A Novel Algorithm for Speaker Change Detection Based on Support Vector Machine

以支援向量機爲基礎之新穎語者切換偵測演算法

王駿發、林博川、王家慶、宋豪靜

wangjf@mail.ncku.edu.tw,tony@icwang.ee.ncku.edu.tw

國立成功大學電機研究所

摘要

對於不同的語者的切換，我們可以利用其不同之語音特徵來加以區別，本論文提出一個以支援向量機(support vector machine, SVM)爲基礎的新穎語者切換偵測演算法；我們定義一個「SVM 訓練分類錯誤率」來判斷語者之料之間的可分離性，藉此判斷是否爲同一個語者的聲音資料。實驗證明我們提出的演算法比貝氏訊息準則(Bayesian information criterion, BIC)演算法有更好的偵測能力並且有更低的誤報率(false-alarm rate)。同時對於兩秒鐘以下的語者變換也可以有效的加以判斷。

關鍵字: 支援向量機(support vector machine, SVM)、貝氏訊息準則(Bayesian information criterion, BIC)、語者切割(speaker segmentation)、語者切換點偵測(speaker change detection)。

I. 簡介

一般而言，對於不同的語者的切換，可利用其不同語音之特徵來加以區別，這種預處理的動作在廣播新聞分類、語音辨識、電話語音分類、自動字幕系統、自動會議記錄、語者識別、語者追蹤(speaker tracking)、語者聚類(speaker clustering)、口述語言資料檢索(spoken document retrieval, SDR)等都有很大的幫助。因此目前有相當多此類的語者分段(speaker segmentation)研究[1-2], [8], [6-12], [16-20], [22], [27], [29-32]。

同時，新聞廣播之聲訊信號源是目前常被研究的信號來源[8-9], [17], [20], [22], [24], [27], [33-37]，因爲其多樣性（包含純音樂、純語音、窄頻語音、具有背景環境或是噪音的語音…等）。而切換點的偵測主要是依照不同的語者、環境、channel等加以標記，最後將相同語者的訊號做群聚(clustering)與合併(merging)。相反地；如果只想取出純音樂的段落可以將非音樂段加以刪除。

對於一個語音串列(speech stream)，事前預蒐集聲學或是語者的模型不是很好的方法且有其困難性，因此不需要事先收集語者資料與任何模型或是訓練的偵測方法(unsupervised manner)是必須的[6], [11]。目前已經有許多研究從事於 unsupervised語者切割之研究 [1], [6], [8], [10-11], [24]。而這些方法主要分爲metric-based、model selection-based與energy-based三類:

A. Energy-based法

一般對話系統的行爲模式中，語者切換點之間通常有靜音段存在；利用能量的大小來判斷切換點[8], [11], [18-19]是很直覺也很簡單的方法。

B. Metric-based法

此方法在語音串列以平移的方式 [3-4], [8-9], [16-17], [27]，使用許多聲學距離：如 Kullback-Leibler distance (KL, KL2) [8], [28], generalized likelihood ratio (GLR) [10], [31], Mahalanobis distance 與 Bhattacharyya distance [9] 來評估兩個相鄰window的相似度；藉此產生

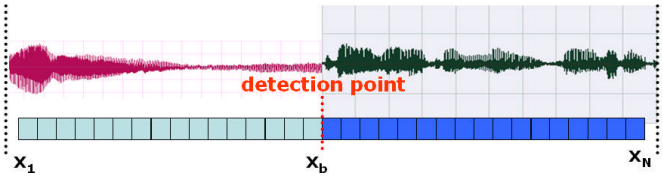
距離曲線(distance curve)；經由一個低頻濾波器濾除因爲雜訊造成的微小波動後[3], [9-11]；取出區域最大值的時間點做爲切換點的輸出。雖然可以有效的加速判斷，但也有許多缺點：(1)需要一個臨界值來選取區域最大值，無法確保所有的語音訊號都適用；(2)只利用鄰近兩個window蒐集的資料來判斷相似度；(3)若使用貝氏訊息準則爲基礎來做相似度判斷[3]；則爲了有充足的資訊量；使得window大小必須大於兩秒鐘；造成對於一秒鐘以下的切換點無法有效偵測[10]。

C. Model selection-based法

由 Schwarz[14]所提出的貝氏訊息準則爲一以模型複雜度加以懲罰(penalize)的可能性準則(likelihood criterion)，被廣泛使用在語者切換點偵測上[1-2], [4] , [6-7], [10], [12-13], [19-23]，對於一個模型 M_i 而言，BIC 定義如下

$$BIC(M_i) = \log P(D_1, D_2, \cdots, D_N \mid M_i) - \frac{1}{2} d_i \log N \text{.....(1)}$$

模型選擇的問題是要從候選模型 M_i， $i = 1,2,\cdots,m$ 中選擇一個來表示一個資料集合 $D = (D_1, D_2, \cdots, D_N)$，$d_i$ 是模型參數集合中獨立的參數個數，$P(D_1, D_2, \cdots, D_t \mid M_i)$ 是模型的最大資料概似(maximized data likelihood)，而 $\frac{1}{2} d_i \log N$ 項的相減動作是模型由 log-likelihood 用來懲罰(penalize)模型複雜度之用。



圖一、使用貝氏訊息準則做語者切換點偵測方法

而使用貝氏訊息準則做語者切換點偵測方法如圖一所示，定義 $X = x_i \in R^d, i = 1,2,\cdots,N$ 是只有一個語者切換點的語音串列；每個 frame 都有一組 cepstral 向量。假設在 $b \in (1, N)$ 這個 frame 有一個切換點，如果可以假設每個聲學 homogeneous 區塊都可以用 multivariate Gaussian process $X \sim N(\mu, \Sigma)$ 加以模型化。則語者切換點的偵測可以視爲一種介於以下巢狀模型(nested model)之模型選擇問題[11]：

$$\begin{aligned} & M_1 : X : x_1, x_2, \cdots, x_N \sim N(\mu, \Sigma) \\ & and \\ & M_2 : x_1, x_2, \cdots, x_b \sim N(\mu_1, \Sigma_1); \\ & x_{b+1}, x_{b+2}, \cdots, x_N \sim N(\mu_2, \Sigma_2) \text{.....(2)} \end{aligned}$$

其中，在 X 中假設所有的取樣都是獨立且類似一個高斯分佈，而 M_1 中假設前 b 個取樣也是一個高斯分佈，而 M_2 中假設最後的 *N-b* 個取樣則是另一個高斯分佈。而切換點 b 的判斷是利用兩個模型差值 $\Delta BIC(b)$ 的正負號來判斷：

$$\Delta BIC(b) = B\bar{I}C(M_2) - B\bar{I}C(M_1) \text{.....(3)}$$

$$= \frac{1}{2}(N \log\left|\hat{\Sigma}\right| - b \log\left|\hat{\Sigma}_1\right| - (N-b)\log\left|\hat{\Sigma}_2\right|) - \frac{1}{2}\lambda(d + \frac{1}{2}d(d+1))\log N .....(4)$$

其中 $\hat{\Sigma}$、$\hat{\Sigma}_1$ 與 $\hat{\Sigma}_2$ 是由相對應的資料所估算出的 ML covariance，λ 是懲罰係數(penalty factor)以補償少量取樣的情況，d 是 cepstral 參數的維度。根據貝氏訊息準則，如果 $\Delta BIC(b) > 0$ 則代表 b 是一個語者切換點，經由 MLE (Maximum Likelihood Estimation)，最終的 BIC 語者切換點判斷式如下:

$$\hat{b} = \arg \max_{1<b<N, \Delta BIC>0} \Delta BIC(b) .....(5)$$

然而Chen的方法[1]在運算量上是屬於二次複雜度(quadratic complexity)，無法實用在real time的系統之中。此外，BIC往往需要足夠的資訊蒐集量才足以判斷出不同語者的切換點，對於較短時間的語者片段無法有效的切割[10]。

在這一篇論文中，我們針對 unsupervised 語者切割提出一個以支援向量機爲基礎的新穎語者切換偵測演算法；定義一個 SVM 訓練分類錯誤率來判斷語者資料之間的可分離性，我們稱其爲「以 SVM 可分離性爲基礎(separability-based)之語者切換偵測演算法」；藉此判斷是否爲同一個語者的聲音資料。本論文的架構如下:在第 **II** 節之中我們對 SVM 演算法做一回顧並提出一個 separability-based 之語者切換偵測演算法，在第 **III** 節之中，SVM 與 BIC 之語者切換點偵測能力將做一系列的評估，而我們所提出的演算法將在第 **IV** 節之中做完整描述，第 **V** 節是實驗環境介紹與實驗結果，最後在第 **VI** 節中做總結。

II. 應用 SVM 訓練錯誤率判斷之語者切換點偵測演算法

A. SVM 演算法簡介

支援向量機(support vector machine, SVM) 是一個新穎的統計學習方法；且近年來引起越來越多研究學者的注意[5][15][25-26][38-39]。SVM 是基於"結構風險最小化" (structural risk minimization, SRM)所構思的歸納原理[40]，主要是以歸納方式求取最小化邊界的錯誤爲目的；而不是以方均誤差最小化爲主。在許多應用之中；SVM 已被證實比傳統 learning machines 有更好的效果，且在分類問題[41]與回歸問題[42]上，被當作是有利的工具。例如在分類議題上；獨立字之數位手寫辨識[25],[43]、語者確認、人臉辨識、知識基礎分類器(knowledge-based classifier)[44]、文件分類(text categorization)[45-46]都是SVM的應用範疇。在迴歸估計(regression estimation)方面，SVM對於benchmark time series prediction tests [47-48]、financial forecasting [49-50]與Boston housing problem [51]都很有競爭力。

本論文使用到的 SVM 演算法部分，是將 l 筆訓練資料(每筆資料 $x_i \in R^N$)利用超平面(hyperplane)分類成爲 $-1,1$ 兩類，而 $y_i \in \{-1,1\}$ 是分類的標記。Hyperplane 的定義爲 $f_H(\mathrm{x}) = \mathrm{w}\cdot\mathrm{x} + b$ ，而最佳的 w 標記爲 $\overline{\mathrm{w}}$ ；求法如下:

$$\overline{\mathrm{w}} = \sum_{i=1}^{l} \overline{\alpha}_i y_i \mathrm{x}_i .....(6)$$

其中 $\overline{\alpha}_i$ 是求解 hyperplane 中導入的 Lagrange multipliers。求出最佳的 hyperplane 後；利用下列公式來加以將資料分類：

$$f_D(\mathrm{x}) = sign(f_H(\mathrm{x})) = sign(\sum_{i=1}^{l} \bar{\alpha}_i y_i \mathrm{x}_i \cdot \mathrm{x} + \bar{b}).....(7)$$

對於無法使用線性方程式將資料分類的情況，我們必須將樣本 x 轉換到高爲度特徵空間 Z 中來處理，即 $x \to \varphi(x) : R^N \to Z$ ，新的 hyperplane 爲：

$$f_H(\mathrm{x}) = \bar{\mathrm{w}} \cdot \mathrm{z} + \bar{b} = \sum_{i=1}^{l} \bar{\alpha}_i y_i K(\mathrm{x}_i, \mathrm{x}) + \bar{b}.....(8)$$

判斷函數爲

$$f_D(\mathrm{x}) = sign(f_H(\mathrm{x})) = sign(\sum_{i=1}^{l} \bar{\alpha}_i y_i K(\mathrm{x}_i \cdot \mathrm{x}) + \bar{b}).....(9)$$

其中 $K(*,*)$ 稱爲 kernel functions；常用的有以下四種：

Linear Kernel: $K(\mathrm{x}, \mathrm{y}) = \mathrm{x} \cdot \mathrm{y}.....(10)$

Polynomial: $K(\mathrm{x}, \mathrm{y}) = (\gamma \cdot \mathrm{x} \cdot \mathrm{y} + c)^d .....(11)$

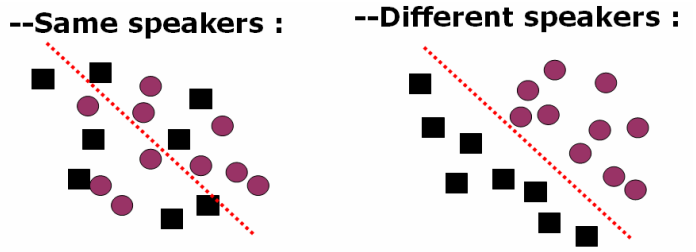
Gaussian Radial Basis Kernel: $K(\mathrm{x}, \mathrm{y}) = \exp(-\gamma \cdot |\mathrm{x} \cdot \mathrm{y}|^2).....(12)$

Sigmoidal Neural Network Kernel: $K(\mathrm{x}, y) = \tanh(\gamma \cdot \mathrm{x} \cdot \mathrm{y} + c)....(13)$

其中 γ 與 c 是常數，而 d 是維度。

B. 以 SVM 之可分離性爲基礎之語者切換偵測演算法

如圖二所示，對於不同的語者而言，可以利用 hyperplane 來將資料分成兩類(+1 與-1)：相反地，如果是同一個語者的語音特徵，hyperplane 將無法有效的將資料分成兩類。



圖二、使用 hyperplane 對相同/不同語者分類示意圖

因此我們可以利用 SVM 在訓練的過程中所找到的 hyperplane 用於計算訓練分類錯誤率(training misclassification rate)來作爲語者切換點的判斷，SVM 的 training misclassification rate 可以分爲兩種，分別是(-1)類誤判爲(+1)的比率與(+1)類誤判爲(-1)的比率；我們分別將其定義爲 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 。

對於同一個語者的語音特徵而言，由於系統所訓練出來的 hyperplane 無法有效的將資料分爲兩類； $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 都會相當高；相反地；對於不同語者的語音特徵而言，由於系統所訓練出來的 hyperplane 可以有效的將資料分爲兩類；因此 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 都會驅近於零，此判斷方式對於語者切換點的偵測能力(detectability)我們將於 III 中測試並評估。

以下詳細說明如何利用 SVM 的 training misclassification rate 來偵測語者切換點，由圖三所示:

Step 0: 將一個語音段 $X = \{x_i : i = 1, \cdots, N\}$ 的每個 frame 取出其語音特徵參數(13 階 MFCCs)。

Step 1:在虛線處(第 *k* 個 frame)假設有一個語者切換點(是不是真的切點還不知道)，立即將虛線左邊 window 的每個 frame 都全部分別標記成 -1 即 $X^- = \{x_i : i = 1, \cdots, k\} \in tag(-1)$；同時將虛線右邊 window 的每個 frame 都全部分別標記成 +1 即 $X^+ = \{x_i : i = k+1, \cdots, N\} \in tag(+1)$。

Step 2:利用 SVM 的訓練過程找出 hyperplane，嘗試將兩個 window 的語音段分開。

Step 3:利用剛才找到的 hyperplane 實際的來將所有 fraeme 分類並標記。假設(-1)類誤判爲(+1)的個數爲 *p*，(+1)類誤判爲(-1)的個數爲 *m*

Step 4:分別計算 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$，其中

$$mis^+(\hat{R}) = m \big/ N - k \quad ....(14)$$

$$mis^-(\hat{R}) = p / k \quad .....(15)$$

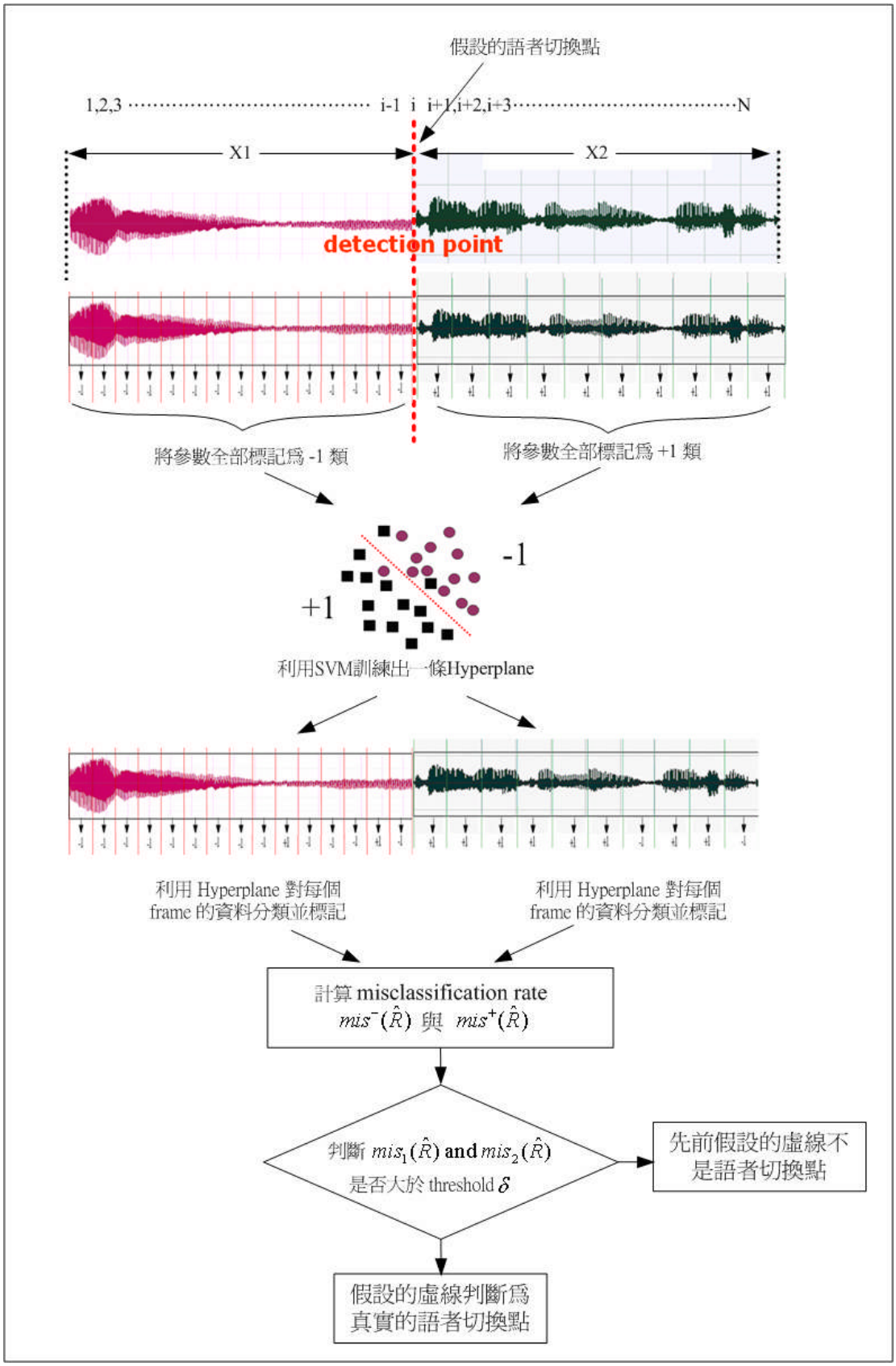
Step 5:判斷 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 使否都小於一個臨界值 δ；成立則判斷 Step 1 所假設的虛線爲真實的語者切換點，不成立則代表 Step 1 的假設錯誤。

III. SVM 與 BIC 之語者切換點偵測能力評估

我們做了以下實驗來評估 SVM 與 BIC 兩個演算法對於語者切換點之偵測能力[1]，實驗中；我們將 BIC 的懲罰係數(penalty factor) λ 設定爲 1[1]。

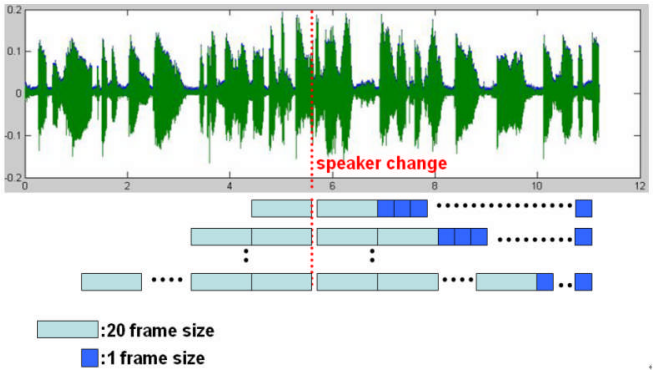
A. 使用左右不同尺寸之 window 來偵測一個已知的語者切換點

本實驗使用了一個 20 秒鐘的語音段(含有兩位語者) 來做爲實驗對象，每位語者分別說 10 秒鐘的語音，由圖四(a)所示，虛線處代表實際的語者切換點。我們從切點處左右挑選出不同的 window 大小來做實驗。實驗中我們先固定切點左邊 window 的大小初值爲 M 爲 20 個 frame；同時定義一個右邊 window 擴增動作來測試，此擴增動作如圖四(a)所示:切點右邊 window 從 M 個 frame 開始，每次增加一個 frame，一直增加到 350 個 frame 爲止。圖(b)與(c)的 X 軸代表切點右邊 window 擴增的大小，而每次挑出的 window 都分別將 ΔBIC 值與 SVM 的兩個 training misclassification rate 即 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 畫在 Y 軸。當切點右邊的 window 每做完一次擴增後，我們將切點左邊 window 的大小 M+10 並重複上述的擴增動作，一直增加到 M=100 爲止，並以不同的顏色來代表。由於 ΔBIC 對於較接近 window 邊界的切點或是 window 收集量太少的情況下判斷力會因爲統計的距離不足而變差[3],[11]，但是如果增加 window 資料收集量將會面臨一個不可避免的問題：那就是 miss detection rate 也將提高[3]。因此這個實驗主要評估兩個驗算法在資料收集量不同下的切點偵測能力。

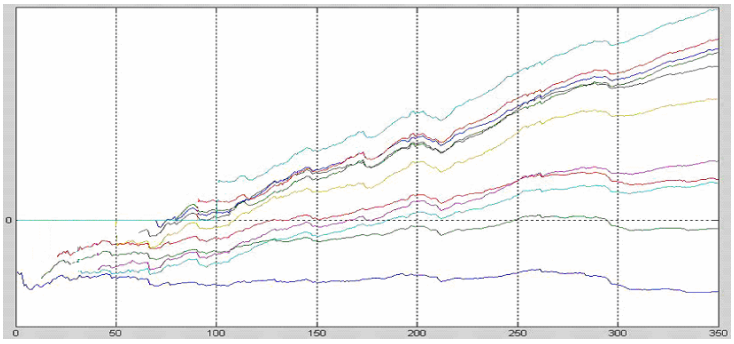


圖三、利用 SVM 的 training misclassification rate 來偵測語者切換點

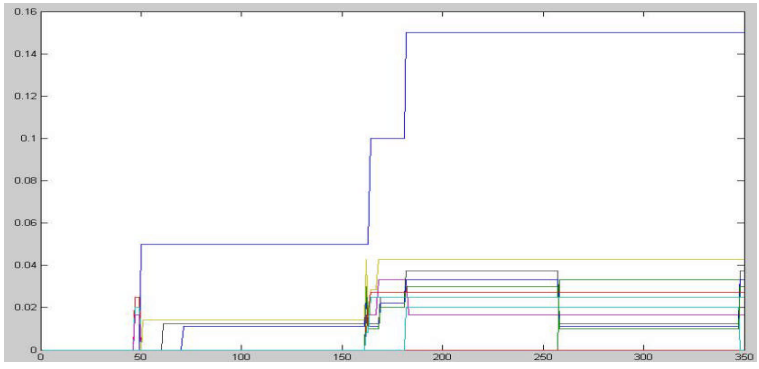
由圖四(b)我們可以發現 BIC 演算法必須要求切點左邊的 window 大小 M 必須大於 90 個 frame 以上(大於一秒鐘)才能使得 ΔBIC 大於零。而 SVM 如圖無論 window 大小如何，都可以得極低的 training misclassification rate；這樣的分佈意味著 SVM 即使在資料收集量極低的情況下仍然可以有良好的判斷力。



(a)



(b)



(c)

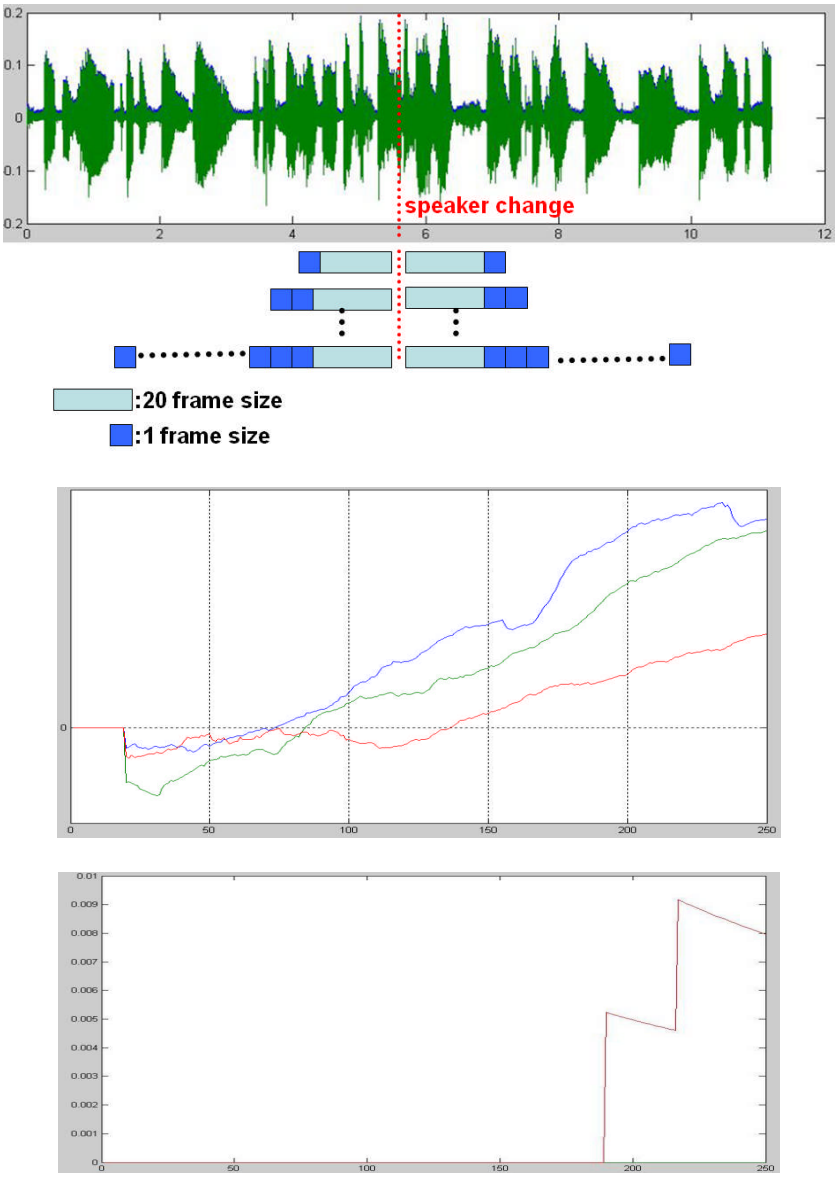
圖四、使用左右不同尺寸之 window 來偵測一個已知的語者切換點

B. 使用左右相同尺寸之 window 來偵測已知的語者切換點

本實驗使用了三個 20 秒鐘的語音串(每個語音串含有兩位語者)來作爲實驗對象，每位語者分別說 10 秒鐘的語音，由圖五所示，虛線處代表實際的語者切換點。我們從切點處左右挑選出相同的 window 大小來做實驗，X 軸代表切點左右兩邊 window 從 20 個 frame 開始，每次切點兩邊的 window 增加一個 frame，一直增加到 250 個 frame 爲止。Y 軸表示每次挑出的 window 所對應的 ΔBIC 值與 SVM 的 training misclassification rate。

由圖五顯示 ΔBIC 值必須要 window 左右兩邊分別收集到大約 137 個 frame(大於一秒鐘)以

上才能保證三個音檔的切點都被偵測出來，這意味著如果不同的語者段落若小於一秒以下將無法被偵測。而 SVM 的效果就相當理想，當 window 收集的資料量在很小的情況下，training misclassification rate 都等於零；亦即 SVM 幾乎都可以百分之百的將資料分成兩類(兩個語者)，也就是 SVM 即使在資料收集量極低的情況下仍然可以有良好的判斷力。唯有其中一個測試語音；當資料量收到大於 190 frames (1.5 秒)時，會有少部分資料呈現誤分類(misclassified)情形，進而造成 training misclassification rate 稍微上升。



圖五、使用左右相同尺寸之 window 來偵測已知的語者切換點

C. 以固定 window 大小掃描之結果分佈來評估 SVM 與 BIC 之易測性

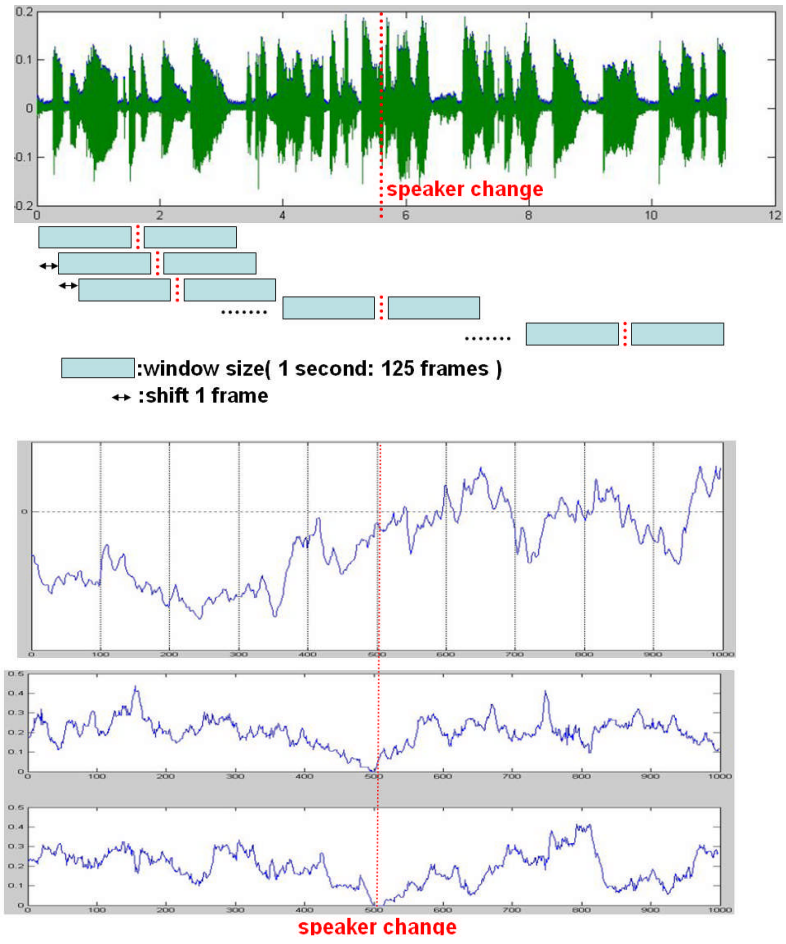
Metric-based 法[3-4], [8-9], [16-17], [27]以假設語者切換點的位置(虛線處)左右相鄰之固定大小 window 來計算 ΔBIC 值，利用 sliding window 的方式掃描出一條 ΔBIC 值曲線，藉由 local peak 來判斷切換點。而這樣的掃描方式，我們恰巧可以套用在 II.***B*** 所提出的判斷方式上；即以 SVM 演算法的 training misclassification rate 曲線來判斷切換點。使用這種方式來偵測切換點的另一個理由是：這種掃描方法可以避免 error broadcasting (先前 windows 的切點判斷錯誤將影響下一個

window 的判斷)並且比 windows 擴張法(expanding scheme)[1-2], [6-7], [11], [21-22]節省運算量[3], [11]。

如圖六所示，同樣以 III.***A*** 的語音段來作爲實驗的對象；我們使用預估切點左右兩邊都是 125 frame (一秒鐘)的 window 去掃描，並且每次右移(shift)一個 frame 直到右邊 window 到達語音終點爲止。

圖六也同時顯示 ΔBIC 值與 SVM 演算法的 training misclassification rate，我們可以看出在實際的語者切換點附近 ΔBIC 的值並不是最高，甚至沒有大於零，因此這是一個 miss detection 的情形。此外也有許多 ΔBIC 值大於零(false-alarm)的情形發生；儘管使用 local peaks 或是一個低頻濾波器來處理這個曲線[9-10]，仍然會有太多的假切點會被判斷成可能的切點(candidate)。

反觀 SVM 演算法的兩個 training misclassification rate；即公式(14)與(15)中的 $mis^{-}(\hat{R})$ 與 $mis^{+}(\hat{R})$ 都在切點附近明顯的下降並趨近於零，我們可以很容易的找出一個有效的 threshold 來判斷，亦即當 $mis^{-}(\hat{R})$ 與 $mis^{+}(\hat{R})$ 同時都低於此 threshold (如 0.05) 時，將其判斷成可能的切點。上述的判斷機制將比[9]更爲簡單有效，並且沒有任何 false-alarm 發生，就語者切點的易測性而言；SVM 演算法有絕對的優勢。



圖六、以固定 window 大小掃描之結果分佈來評估 SVM 與 BIC 之易測性

D. 低於兩秒以下之語者切換點偵測能力評估

由於 BIC 在資料的蒐集上必須有一定長度，因此在 metric-based 的系統上大部分在 window 的設計上都是以預測切換點左右兩秒鐘爲主，而擴增法的系統則都是以兩秒鐘 window 開始擴增。因此在以往的文獻中，對於低於兩秒以下之語者切換點都無法有效的判斷[1-4], [8-10], [16], [32]。爲了測試 SVM 的 training misclassification rate 在低於兩秒鐘語者切換點判斷上的強健性程度，我們設計了兩個實驗來加以測試。

如圖七所示，我們錄製一個由三位語者發聲的語音，每兩秒鐘即切換到另一個語者，由實驗結果發現 ΔBIC 仍然會有 false-alarm，而 SVM 的 training misclassification rate 依然相當準確，易測性也相當高。

如圖八所示，在第二個實驗中；同樣由三位語者發聲；但是第二位語者只有錄製一秒鐘，由實驗結果發現 BIC 無法判斷出任何的語者切換點(所有 ΔBIC 都小於零)，而 SVM 的 training misclassification rate 依然相當準確；易測性也相當高。

藉由上述的實驗評估；我們發現 SVM 具有以下優勢：

1. 可以用更少的資料收集量來判斷切點；對更短的語者切點做偵測。亦即對於 duration 在兩秒鐘以下的短語音段落[11]，有更好的判斷力與偵測能力，因此 miss detection rate 可以降低[3]。
2. 易測性高，不需要額外使用一個低頻濾波器[3][9-11]。
3. 相較於 ΔBIC 值，對於一個不可避免的 heuristic threshold 而言，SVM 的 training misclassification rate 之分佈很明顯，因此 threshold 大小很容易選擇。
4. False-alarm 機率極低。

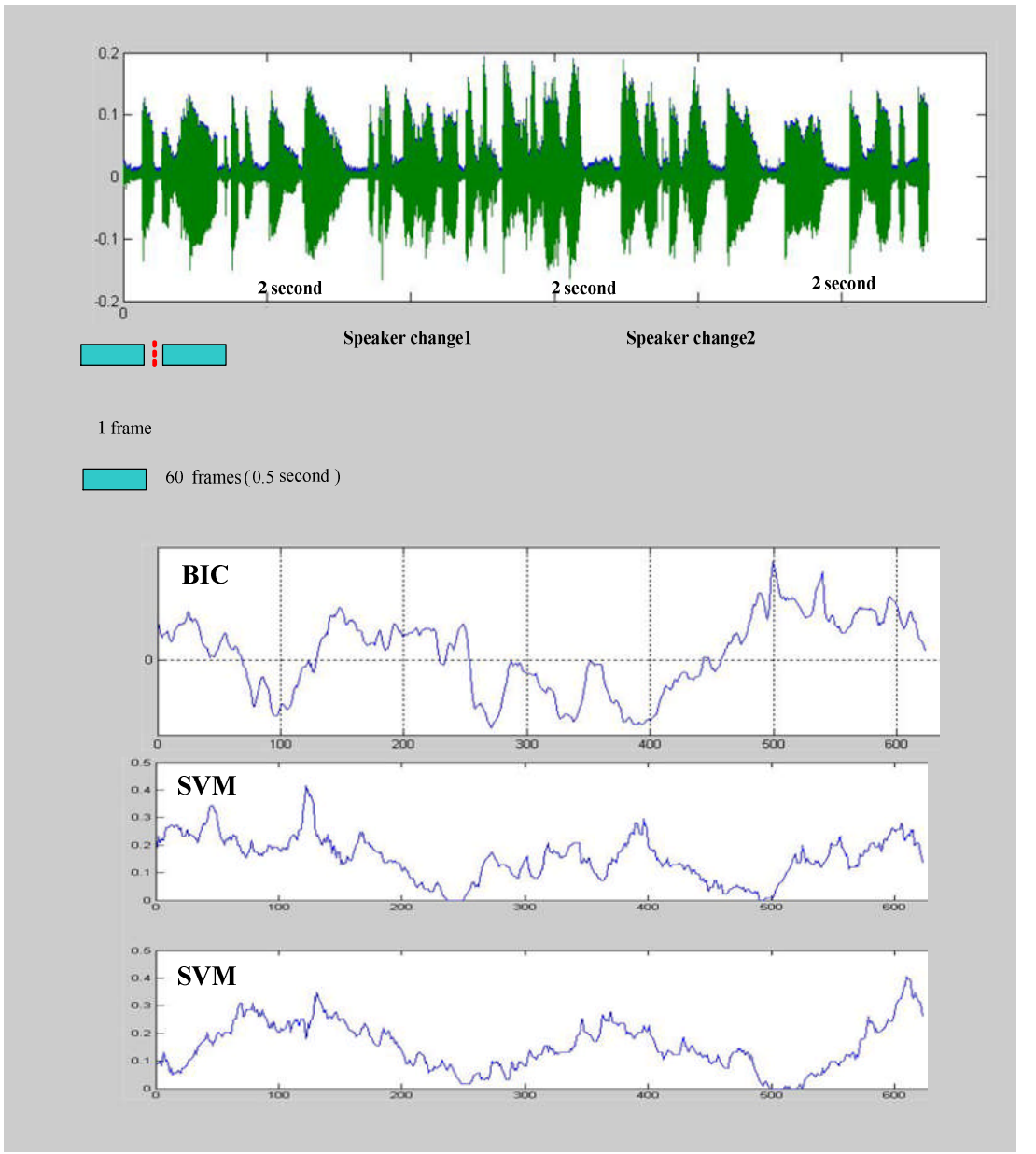
E. 使用相同語者之語音串評估滑動窗與 training misclassification rate 之 threshold 大小

如圖九所示，爲了要找出一個適合於 SVM 的 sliding window 大小與 training misclassification rate 之 threshold，我們使用四個語音串(以不同顏色區分)來做評估，每個語音串都只有一位語者。利用 III.*B* 中；使用切點左右相同尺寸之 window 來掃描。由 SVM 的 training misclassification rate 來觀察； $mis^{-}(\hat{R})$ 與 $mis^{+}(\hat{R})$ 在 0.05 (紫色虛線)以上將會穩定的爬升，意味著當 window 收集資料越多，SVM 對於同一個語者的語音將越無法有效的找出一條 hyperplane 將資料分類，進而導致 training misclassification rate 越來越大。因此我們使用 0.05 來當作 training misclassification rate 之 threshold，當 training misclassification rate 大於 0.05 則表示左右兩邊的 window 是同一個語者的聲音，因此原先假設的切換點無效。相反地，misclassification rate 小於 0.05 表示 SVM 找到一條有效的 hyperplane 將資料分類，因此原先假設的切換點正確。

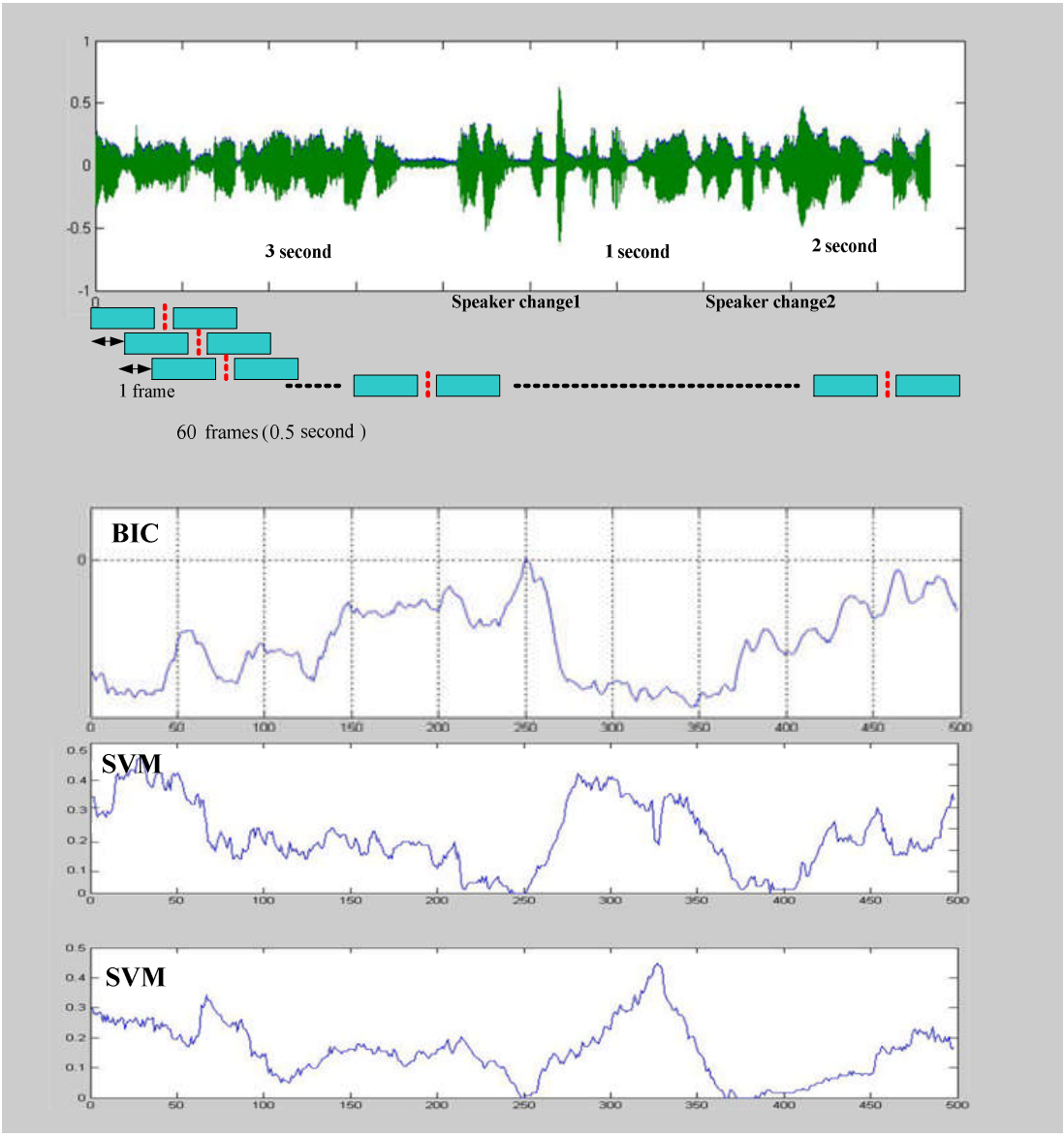
此外，當 sliding window 大於 70 個 frame 時，所有語音串的 training misclassification rate 都會大於 0.05 並且穩定的爬升。我們可以說：當切點兩邊相鄰的 window 收集資料都大於 70 個 frame (左右兩邊共 140 個 frame) 時，SVM 將可以有效地判斷出左右兩邊的 window 是同一個語者的聲音。

IV. 語者切換點偵測驗算法

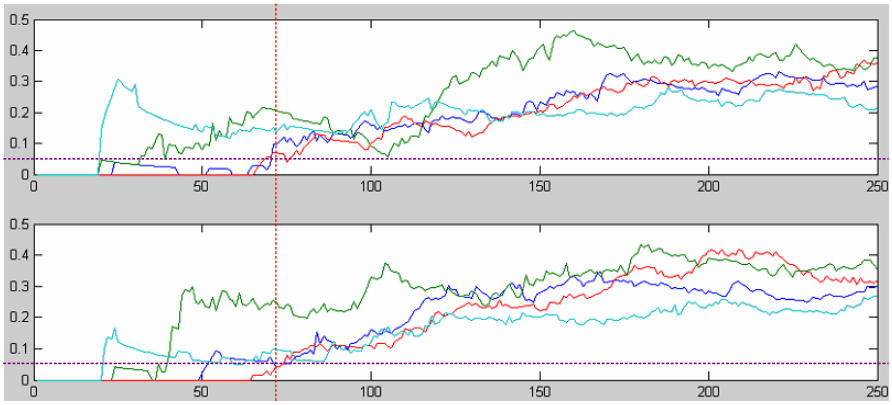
我們所提出的演算法主要分爲三個步驟：首先，第一步驟我們將可能的語音切點(potential speaker change point)偵測出來，這個步驟是整各演算法的關鍵步驟，由於接下來的第二步驟：確認處理(confirmation process)完全仰賴第一步驟的偵測品質[11]。在第三步驟中我們將相鄰且是相同語者的語音段加以合併(merging)。



圖七、低於兩秒以下之語者切換點偵測能力評估一



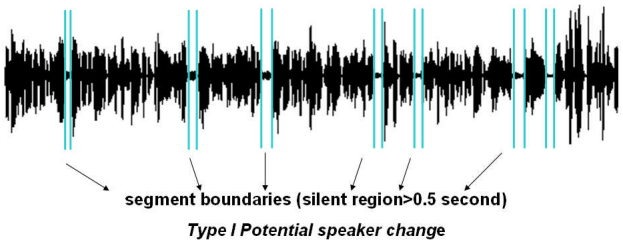
圖八、低於兩秒以下之語者切換點偵測能力評估二



圖九、使用相同語者之語音串評估滑動窗與 training misclassification rate 之 threshold 大小

A. 步驟一之第一類可能語者切換點(位於靜音段邊緣的切換點)之偵測

對於位於靜音段邊緣的切換點，我們使用最直覺的 energy-base 判斷方式來做偵測，而靜音段的 duration 要多久以上才有可能是一個語者切換點? 我們可以使用一個經驗值的 threshold 來判斷。如圖十所示，是一個使用 0.5 秒爲 threshold 的靜音段來判斷可能切換點的切割結果。



圖十、第一類可能語者切換點(位於靜音段邊緣的切換點)之偵測

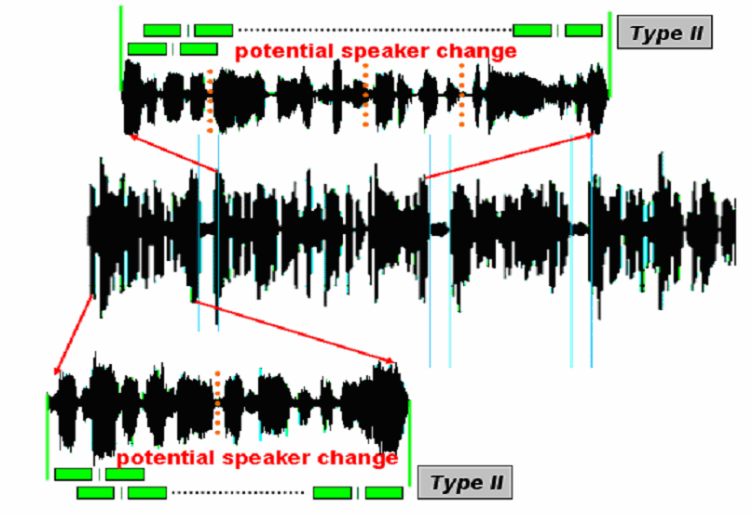
B. 步驟一之第二類可能語者切換點(非靜音語音段的切換點)偵測

如圖十一所示，針對此類非靜音語音段的語者切換點，我們使用 II.***B*** 所提到的應用 SVM 訓練錯誤率判斷之語者切換點偵測演算法；每次右移一個 frame 來做掃描，藉此來判斷出此類可能的切換點。在 III.***E*** 中，我們曾對所提出演算法的 window size 做評估，其結論是 70 個 frame 是一個不錯的資料收集量，但對可能的語者切換點而言；我們使用較小的 60 個 frame 來當作 sliding window 的大小，這樣設定的原因是我們寧願提高 false-alarm 將所有可能的切換點都找出來，藉此降低 miss detection rate。而那些多出來的 false-alarm 可以使用「切換點確認」與「音段合併」等步驟有效的加以排除[3], [12-13]。而 60 個 frame (左右兩邊共 120 個 frame)之 window size 可以更有效的判斷一秒鐘左右(每秒 125 個 frame)的切換點。相較於其他系統[1-4], [8-10], [16], [32]只能判斷兩秒鐘以上的語者切換點而言，有更多的優勢。另一個優點是：對於每次 shift 一個 frame 的 sliding 機制而言，僅用 60 個 frame 的大小可以大幅的降低運算量。

C. 步驟二:切換點確認

從 **IV.*A*** 與 **IV.*B*** 所偵測出來的可能切換點只是初步的推測，在這個步驟中，我們取出那些可能切點左右兩邊各 1.5 秒鐘的 window 重複地使用 **II.*B*** 提出的方法來做判斷，企圖以更多的資料收集量來確認先前在 **IV.*A*** 與 **IV.*B*** 找出的切換點，同時刪除 false-alarm 的切換點，詳細情形如圖十二所示。

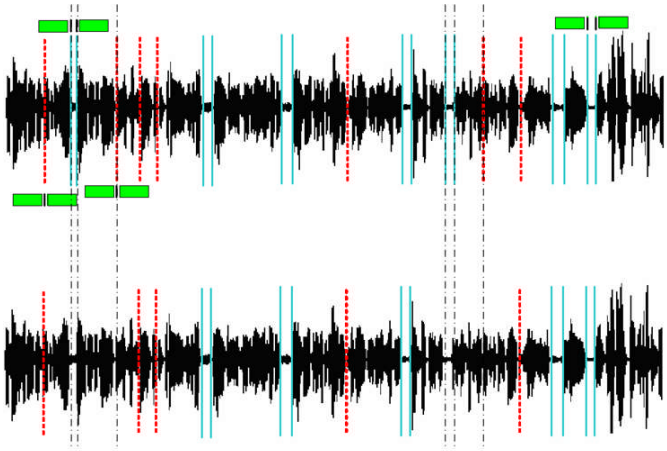
由於靜音段的資料對於判斷切換點沒有幫助甚至會影響 training misclassification rate 的判斷，因此對於第一類靜音的切換點而言(藍色實線)；我們取出切音點左右兩邊非靜音的 1.5 秒來當做 window。對於第二類非靜音語音段的語者切換點(紅色虛線)而言，直接取出可能切點左右兩邊各 1.5 秒鐘的 window 即可。黑色虛線部份即爲 false-alarm 的切換點刪除效果。



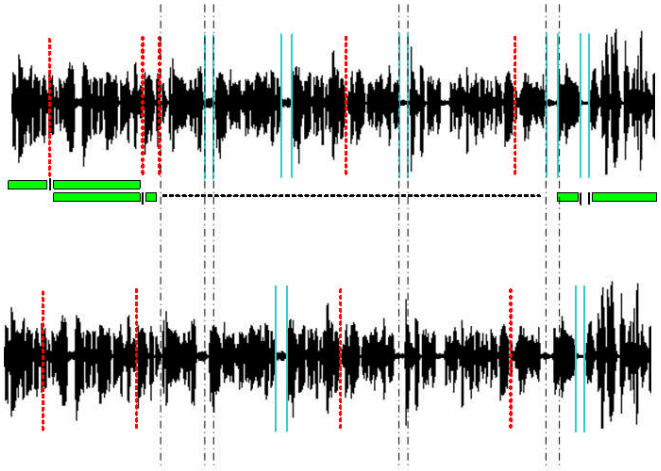
圖十一、第二類可能語者切換點(非靜音語音段的切換點)之偵測

D. 步驟三:相同語者音段合併

最後一個步驟中我們將先前所確認後的語音段落加以合併，採用左右不固定 window 大小的方式；直接將兩個鄰近的語音段落使用 **II.*B*** 提出的方法再次做切換點偵測，藉此將同一個語者的語音段落加以合併，詳細情形如圖十三所示。當語音段太多的時候，這個步驟的動作可以重複兩次或是三次，以求效果。



圖十二、切換點確認



圖十三、相同語者音段合併

V. 實驗結果

A. 實驗環境介紹

我們使用的參數爲 13 階的的 MFCC 參數，分別對兩個語音串做實驗；平均每個語音串有 30 分鐘的語音；並且含有六位語者交互說話。

針對語者切換點偵測系統而言；一般有兩種錯誤，對於系統無法找到實際換點位置的錯誤稱爲 detection error，另一種錯誤指的是系統找到的切換點位置不是或是無法對應到實際的切換點位置，稱之爲 false-alarm 或是 segment insertion。我們也可以用資訊檢索的 precision 與 recall 來表示上述的兩個錯誤，其定義如下:

$$\Pr ecision = \frac{number\ of\ correctly\ found\ boundaries}{total\ numer\ of\ boundaries} .....(16)$$

$$\mathrm{Re} call = \frac{number\ of\ correctly\ found\ boundaries}{numer\ of\ hypothesized\ boundaries} .....(17)$$

另外我們也使用 F-measure [13]來對 precision 及 recall 進行綜合評估，針對對等參數(neutral parameterization)而言；我們給予 precision 與 recall 相同的權重，因此 F-score 的定義如下:

$$Fscore = \frac{2 \times \Pr ecision \times \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call} .....(18)$$

我們定義一個放寬時間(tolerance) Δt 來做判斷，對於一個真實切換點位置在 t_0 的切換點而言，如果偵測出來的切換點發生在 $t_0 - \Delta t < t_0 < t_0 + \Delta t$ 之間；都視爲正確的偵測，在我們的實驗中我們設定 $\Delta t = 0.5$ 。此外；這樣的判斷也突顯演算法的強健性，對於系統偵測出來的切換點；我們僅容忍偏移量在左右 0.5 秒以內，才算是正確的偵測，並非其他系統所採用的：左右偏移 1 秒鐘[3]都算偵測正確的評估方式。

B. 相關參數設定實驗

對於偵測第二類語者切換點所需要的相關參數設定；在 **IV.*B*** 我們已經決定了 sliding window 大小爲 60 個 frame，以及在 **III.*E*** 中，我們決定了 SVM 的 training misclassification rate 爲 0.05。

而針對切換點確認(**c**onfirmation)與合併(merging)；我們提出四種不同的組合來實驗，分別是 CS_MS、CS_MB、CB_MS 與 CB_MB，藉此觀察 SVM 與 BIC 的效能。其縮寫定義如下:

- **CS**: **C**onfirmation by **S**VM.
- **CB**: **C**onfirmation by **B**IC.
- **MS**: **M**erging by **S**VM.
- **MB**: **M**erging by **B**IC.

針對以上 BIC 的懲罰係數(penalty factor)與 SVM 的 training misclassification rat 臨界值之設定；我們採用 10 個 10 秒鐘的語音串來做實驗；每個語音串都只有一個語者切換點。從這 10 個小測試中求得一個最適當的參數設定。最後決定的參數設定如表一所示：

表一、BIC 的懲罰係數(penalty factor)與 SVM 的 training misclassification rat 臨界值之設定

Algorithm	SVM based	BIC based
Parameter	Threshold of training misclassification rate	Penalty factor
Confirmation	0.1	1.4
Merging	0.05	1.4

在 SVM 的 **c**onfirmation 過程中我們採用較高的 training misclassification rate 臨界值來判斷；藉此放寬 threshold。最後的 merging 步驟再使用較低的 training misclassification rate 臨界值來判

斷，這樣的設定主要是希望獲的更高的 recall rate，因爲對於一個切換點偵測系統而言；確實偵測到語者切換點是最主要的目的，因此我們對 recall rate 的要求更勝於 precision rate。

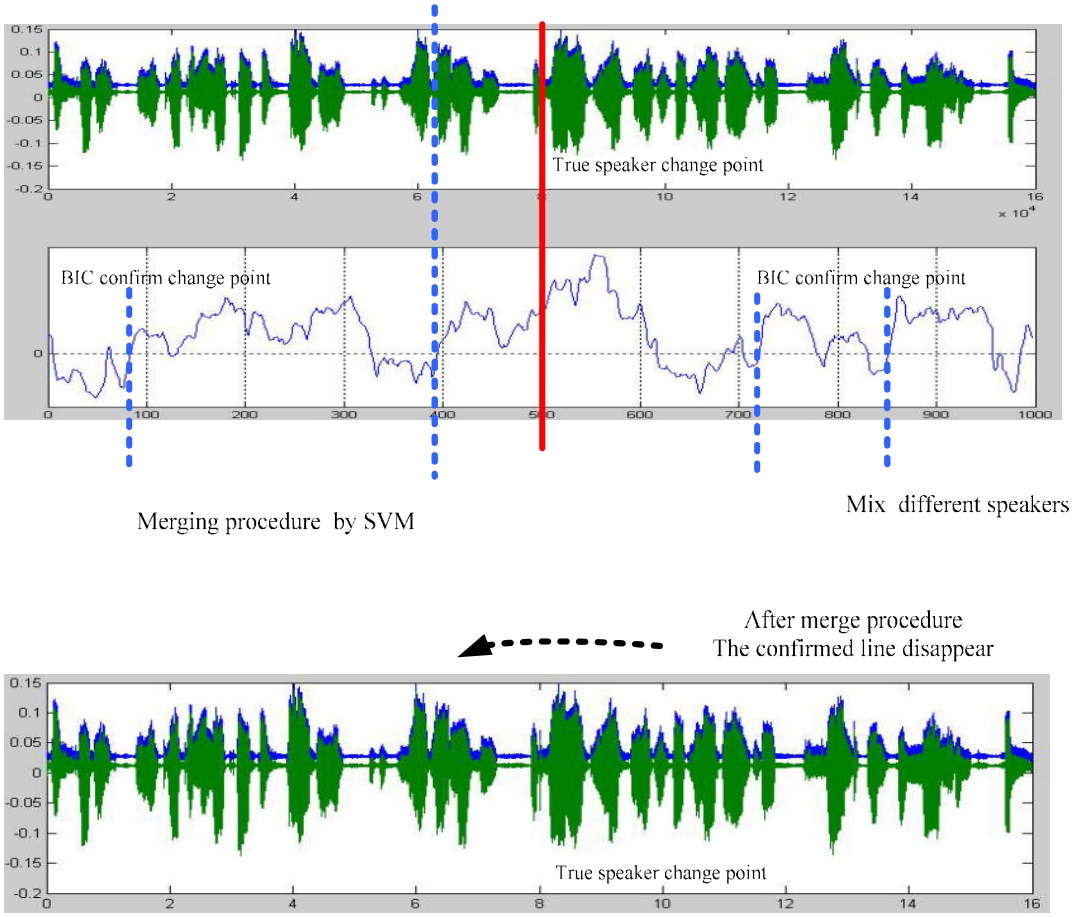
C. 整體實驗結果比較與討論

整個實驗的結果整理如表二所示，其中 CS_MS (Confirmation by SVM & Merging by SVM) 有最好的偵測效果，我們推測是因爲在 **IV.*B*** 之中；針對第二類可能的切換點使用較小的 window 去掃描的原因；再加上 training misclassification rate 的強健性，所有可能的切換點都可以被找到的原因。

表二、實驗結果整理

Data		Precision	Recall	F-score
Chinese	CS_MS	0.76	0.90	0.82
	CS_MB	0.79	0.81	0.79
	CB_MB	0.57	0.77	0.65
	CB_MS	0.53	0.75	0.62
English	CS_MS	0.70	0.92	0.79
	CS_MB	0.72	0.80	0.75
	CB_MB	0.44	0.85	0.57
	CB_MS	0.57	0.62	0.59

在 CB_MS(Confirmation by BIC & Merging by SVM)實驗中；BIC 的確認點與實際的語者切換點有偏移的現象，如圖十四所示：實際的切換點爲紅色實線，而 BIC 的 confirmation 結果爲藍色虛線；這樣子的確認結果在 SVM 的 merging 之步驟中會將紅色橢圓虛線視爲同一語者的聲音，但其實已經混雜兩個語者的語音資料。這樣的結果會導致 training misclassification rate 上升，進而造成 SVM 的誤判；recall rate 因此最差。



圖十四、CB_MS(Confirmation by BIC & Merging by SVM)實驗分析

VI. 結論

在這一篇論文之中我們提出了以SVM可分離性爲基礎(separability-based)之語者切換偵測演算法，這種新穎的語者切換點判斷機制，不同於以往採用的演算法，而是使用 SVM 的 training misclassification rate 與 metric-based 的掃描方式加以結合。在偵測能力的評估上，我們亦做了許多實驗來證實 SVM 演算法的強健性。此外，本方法也同樣擁有像 BIC 不需要事先訓練語者資料或是建立任何模型的優勢。由實驗發現 CS_MS (Confirmation by SVM & Merging by SVM)有最好的偵測效果，證明 SVM separability-based 法應用在語者切換點偵測上之優越性。

參考文獻

[1]. S. Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion, “in Proc. DARPA Broadcast News Transcription Understanding Workshop, Feb. 1998, pp. 127-132.

[2]. M. Cettolo and M. Vescovi, “Efficient audio segmentation algorithms based on the BIC,” Proceedings of ICASSP2003.A Sequential Metric-based Audio Segmentation

[3]. Shih-Sian Cheng and Hsin-min Wang "METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation," Proc. International Conference on Spoken Language Processing (ICSLP2004), Jeju Island, Korea 2004.

[4]. S. S. Cheng and H. M. Wang, “A sequential metric-based audio segmentation method via the

Bayesian Information Criterion," Proceedings of Eurospeech 2003.

[5]. Jhing-Fa Wang Taiwan - Jia-Ching Wang, Tze-Hsuan Huang and Cheng-Shu Hsu, "Home Environmental Sound Recognition Based on MPEG-7 Features," 46th IEEE International Midwest symposium on Circuits and systems, 2003.

[6]. B. W. Zhou, and John H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion," Proceedings of ICSLP 2000.

[7]. A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian Information Criterion," Proceedings of Eurospeech 1999.

[8]. M. Siegler, U. Jain, B.Raj, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio, " in Proc. DARPA Speech Recognition Workshop, Feb, 1997,pp. 97-99

[9]. J.W. Hung, H.M. Wang, and L.S. Lee. Automatic Metricbased speech segmentation for broadcast news via principal component analysis. Proceeding of ICSLP'2000.

[10]. P. Delacourt, C. J. Welkens, DISTBIC: A Speaker-based segmentation for Audio Data Indexing, Speech Communication, v. 32, pp 111-126, 2000.

[11]. Bowen Zhou and John H. L. Hansen, "Efficient Audio Stream Segmentation via the Combined T2 Statistic and Bayesian Information Criterion", IEEE Transactions On Speech And Audio Processing, Vol.13, No.4, July 2005.

[12]. Meinedo, H., Neto, J.A., "Audio Segmentation, Classification and Clustering in a Broadcast News Task", Proc. ICASSP'2003 - Hong Kong, China, Apr. 2003.

[13]. X. Zhong, M. Clements, and S. Lim, "Acoustic change detection and segment clustering of two-way telephone conversation," Proceedings of Eurospeech2003.

[14]. G. Schwarz, "Estimating the dimension of a model," The Annals of Statistics, vol. 6, no. 2, pp.461–464, 1978.

[15]. Cheng-Shu Hsu, "Home Environmental Audio Classifier Based on SVM and MPEG-7 Audio Low-level Descriptors", Master Thesis, NCKU, Taiwan, 2002.

[16]. H. Beigi and S. Maes, ``Speaker, channel and environment change detection", Proceedings of the World Congress on Automation, 1998.

[17]. Luis Perez-Freire and Carmen García-Mateo, "A Multimedia Approach For Audio Segmentation In Tv Broadcast News", ICASSP 2004.

[18]. F. Kubala et al., ``The 1996 BBN Byblos Hub-4 transcription system", Proceedings of the Speech Recognition Workshop, pp 90-93, 1997.

[19]. S. Chen et al., ``IBM's LVCSR System for Transcription of Broadcast News Used in the 1997 Hub4 English Evaluation", Proceedings of the Speech Recognition Workshop, 1998.

[20]. M. Harris, X. Aubert, R. Haeb-Umbach, and P. Beyerlein, "A study of broadcast news audio stream segmentation and segment clustering,"in Proc. EUROSPEECH, Budapest, Hungary, 1999, vol. 3, pp. 1027–1030.

[21]. P. Sivakumaran, J. Fortuna, and A. M. Ariyaeeinia, "On the use of the Bayesian Information Criterion in multiple speaker detection," in Proc. EUROSPEECH, Aalborg, Denmark, 2001, vol.

2, pp. 795–798.

[22]. M. Cettolo, “Segmentation, classification and clustering of an Italian broadcast news corpus,” in Proc. of the 6th RIAO-Content-Based Multimedia Information Access - conference, Paris, France, 2000.

[23]. A. Raftery, "Bayesian Model Selection in Social Research", Tech. Reports, Dept. of Stat., Univ. of Washington, 1994.

[24]. R. Bakis et al., "Transcription of broadcast news shows with the IBM large vocabulary speech recognition system ", Proceedings of the Speech Recognition Workshop, pp 67-72, 1997.

[25]. C. Cortes and V. Vapnik, “Support vector networks,”Machine Learning, vol. 20, pp. 273-297, 1995.

[26]. V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer, 1995.

[27]. Laurent Couvreur and Jean-Marc Boite, "Speaker Tracking in Broadcast Audio Material in the Framework of the THISL Project", Proc. of European Speech Communication Association (ESCA) European Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio, 1999.

[28]. M. Basseville, "Distance Measures for Signal Processing and Pattern Recognition", Signal Processing, Vol. 18, 1989.

[29]. P. Woodland and al., “The development of the 1996 HTK broadcast news transcription system,” in DARPA speech recognition workshop, 1997.

[30]. J. Johnson and P. Woodland, “Speaker clustering using direct maximisation of the MLLR-adapted likelihood,” in ICSLP98, 1998.

[31]. H. Gish and N. Schmidt, “Text-independent speaker identification,” IEEE signal processing magazine, oct. 1994.

[32]. P. Delacourt and C. J. Wellekens, “Audio data indexing: use of second-order statistics for speaker-based segmentation,” in ICMCS, 1999.

[33]. S. Chen, E. Eide, M. Gales, R. Gopinath, D. Kanevsky, and P. Olsen, "Recent improvements to IBM’s speech recognition system for automatic transcription of broadcast news,” in Proc. DARPA Broadcast News Transcription Workshop, 1999.

[34]. J. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system,” Speech Commun., vol. 37, no. 1–2, pp. 89–108, 2002. [33] Recent improvements to IBM’s speech recognition system for automatic transcription of broadcast news

[35]. T. Hain, S. E. Johnson, A. Tuerk, P. C.Woodland, and S. J. Young, “Segment generation and clustering in the HTK broadcast news transcription system,” in Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, 1998, pp. 133–137.

[36]. S. Wegmann, P. Zhan, and L. Gillick, “Progress in broadcast news transcription at Dragon systems,” in Proc. IEEE ICASSP-99: Inter. Conf. Acoust., Speech, Signal Process., May 1999, 1912.

[37]. P. Zhan, S.Wegmann, and L. Gillick, “Dragon systems’ 1998 broadcast news transcription

system for Mandarin," in Proc. DARPA Broadcast News Transcription Workshop, 1998.

[38]. V. Vapnik, Statistical Learning Theory. New York: Wiley, 1998

[39]. B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. Smola, "Input space vs. feature space in kernel-based methods," IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 1000-1017, 1999

[40]. V. Vapnik, Estimation of Dependences Based on Empirical Data. Springer-Verlag, 1982

[41]. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.

[42]. Smola and B. Schölkopf, "A tutorial on support vector regression," Tech. Rep. NC2-TR-1998-030, Neural and Computational Learning II, 1998

[43]. J. C. Burges and B. Schölkopf, "Improving the accuracy and speed of support vector learning machines," in Advances in Neural Information Processing Systems 9 (M. Mozer, M. Jordan, and T. Petsche, eds.), pp. 375-381, Cambridge, MA: MIT Press, 1997.

[44]. G. Fung, O. L. Mangasarian, and J. Shavlik, "Knowledge-based support vector machine classifiers," in Advances in Neural Information Processing, 2002.

[45]. T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in Proceedings of ECML-98, 10th European Conference on Machine Learning (C. Nédellec and C. Rouveirol, eds.), (Chemnitz, DE), pp. 137-142, Springer Verlag, Heidelberg, DE, 1998.

[46]. K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," in Computational Learning Theory, pp. 35-46, 2000

[47]. K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in Articial Neural Networks - ICANN'97 (W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, eds.), pp. 999-1004, 1997.

[48]. S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using support vector machines," in 1997 IEEE Workshop on Neural Networks for Signal Processing, pp. 511-519, 1997.

[49]. F. E. H. Tay and L. Cao, "Application of support vector machines in financial time series forecasting," Omega, vol. 29, pp. 309-317, 2001.

[50]. L. J. Cao, K. S. Chua, and L. K. Guan, "c-ascending support vector machines for financial time series forecasting," in 2003 International Conference on Computational Intelligence for Financial Engineering (CIFEr2003), (Hong Kong), pp. 317-323, 2003.

[51]. H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in Advances in Neural Information Processing Systems, vol. 9, p. 155, The MIT Press, 1997.

錄音資料中的語者切割與分群

Speaker Segmentation and Clustering for the Recorded Speech

蘇峻慶、王小川
Chun-Ching Su and Hsiao-Chuan Wang

國立清華大學電機工程學系
Department of Electrical Engineering, National Tsing Hua University

Email: g923990@oz.nthu.edu.tw hcwang@ee.nthu.edu.tw

摘要

本論文主要在探討錄音資料中語者切割與語者分群的問題，在語者切割方面，採用三個步驟，第一步是利用貝氏資訊準則約略找出語者轉換點大概的位置，第二步利用交叉偵測法作精確化，第三步再確認是否為轉換點，實驗上顯示此方法擁有運算量少及高準確率的優點。在語者分群方面，群集之語者模型採用高斯混合模型，音段與每一個群集模型作最大概似法估測，找出最靠近之群集，然後再利用一個門檻值判斷是要合併或是分離出新的群集。實驗結果顯示音段群中包含語者數愈多，其整體分群效能愈低。

關鍵詞：語者切割、語者分群、語者轉換點偵測、群集模型

一、 緒論

語言是人類溝通及傳達意念最自然的方法，語音訊號不只包含了說話者所要表達的意思，更是隱含了說話者的個人特徵，因此在一段語音信號中，我們不僅要能夠聽出其中所要表達的意思，更要知道這一段話究意是誰所講的。

近年來從有線或無線網路上以語音擷取資訊的應用增加，身份確認或說話人辯識變得更為重要，愈來愈多人投入自動語者辨識的研究領域。在多人說話的環境下，變成需要先對語音做分段，然後再辨認各個音段是誰在說話，因此就需事先作切割與分群。舉例來說，在一個重要會議場合的錄音，其內容包含若干人的談話，若想將這些語者的語音訊號分開，利用人工方法是既費時又不經濟，因此有必要發展出一套正確率高，速度又快的切割與分群方法。

過去已有許多語者切割的方法被提出[1][2]，而這些被提出的方法大致可分類為以解碼為基礎之切割法(Decoder-Guided Segmentation)、以模型為基礎之切割法(Model-based Segmentation)、以及以距離為基礎之切割法(Metric-Based Segmentation)。以上三種方法都有其優缺點，像以解碼為基礎之切割法，只能粗略地分類出語音、音樂、靜音等，並無法用來偵測出語者轉換點的位置。以模型為基礎之切割法，需要事先搜集相關語料建立相對應的模型，這並不符合實際。以距離為基礎之切割法，則需設定門檻值(Threshold Value)來決定語者轉換點的位置，因此缺少穩定性(Stability)和強健性(Robustness)。

語者分群是一個活躍多年的研究領域，大致上在作語者分群時有幾個基本的問題[3]：

1. 聚集(agglomeration)：對一群音段作語者分群時，其形成群集的方式有兩種，一種是凝聚，另一種是分裂。
2. 停止準則(stopping criteria)：在作語者分群時，通常是不曉得音段群裡包含多少個語者，因此需設立一個停止準則，當群集數達到此一停止準則，即停止再分新群。
3. 距離量測(distance measures)：利用一個距離量測的方法，用以決定所偵測的音段是屬於哪一群。

本文在語者切割方面，採用三個步驟，第一步是利用貝氏資訊準則約略找出語者轉換點大概的位置，第二步利用交叉偵測法作精確化，第三步再確認是否為轉換點，實驗上顯示此方法擁有運算量少及高準確率的優點。在語者分群方面，群集之語者模型採用高斯混合模型，音段與每個群集模型作最大概似法估測，找出最靠近之群集，然後再利用一門檻值判斷是要合併或是分離出新的群集。

本文內容安排如下：第二節詳細說明語者切割的基本技術及本論文所使用的方法，第三節說明本論文使用的語者分群方法，第四節是實驗設計及對實驗結果做討論，第五節為結論。

二、 語者切割

2.1 語者轉換點偵測(Speaker Change Detection)

語者轉換點偵測就是偵測說話者改變時的轉換點，最常被使用來偵測的方法，一為貝氏資訊準則(Bayesian Information Criterion, BIC)，另一為廣義概似比(Generalized Likelihood Ratio, GLR)，以下分別介紹這兩種方法。

(A) 貝氏資訊準則(Bayesian Information Criterion, BIC)[4]

假設 $M = M_1, M_2, M_3, ..., M_k$ 是所有的候選模型集合，k_j 是 M_j 這一個模型的參數數目，$X = X_1, X_2, X_3, ..., X_N$ 為一群資料集，根據定義，BIC 可寫成下式：

$$BIC(M_j) = \log L\left\langle X_1, X_2, ..., X_N \middle| M_j \right\rangle - \frac{1}{2}\lambda k_j \log N \qquad (1)$$

其中 $L\left\langle X_1, X_2, ..., X_N \middle| M_j \right\rangle$ 為模型 M_j 和資料集 X 的最大概似值 (Maximum Likelihood)，λ 為損失權重，根據(1)式，就可從眾多模型中找出一個最佳的模型來描述資料集 X。

(B) 貝氏偵測法[1]

假設 $X = \{x_1, x_2, x_3, ..., x_N\}$ 代表一語音段的特徵向量，且只包含一個語者轉換點，如圖 1 所示。假設語者轉換點發生在 i 的時間點上，我們設定二個假說測試(Hypothesis Testing)，其定義如下：

$$H_0 : x_1, x_2, ..., x_N \sim N(\mu, \Sigma) \qquad (2)$$

$$H_1 : x_1, x_2, ..., x_i \sim N(\mu_1, \Sigma_1); x_{i+1}, x_{i+2}, ..., x_N \sim N(\mu_2, \Sigma_2) \qquad (3)$$

(2) 式表示全音段的特徵參數序列，呈高斯分布。(3) 式表示分成兩段音段的特徵參數序列，也是呈高斯分布。

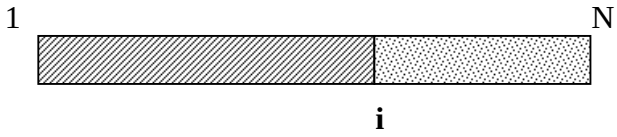


圖 1　長度為 N 並包含一個語者轉換點的語音段

將 H_0 與 H_1 兩模型作比較，比較的式子定義如下：

$$\Delta BIC = BIC(H_1) - BIC(H_0) \tag{4}$$

把(1)式、(2)式及(3)式代入上面的(4)式，可得到下列的結果：

$$\Delta BIC = R(i) - \lambda P \tag{5}$$

其中 λ 為一個加權值，$R(i)$ 為最大概似比(Maximum Likelihood Ratio)：

$$R(i) = N\log|\Sigma| - i\log|\Sigma_1| - (N-i)\log|\Sigma_2| \tag{6}$$

P 為懲罰值(penalty)：

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1))\log N \tag{7}$$

d 為特徵參數維度，N為特徵參數數量。

若該 i 點的 ΔBIC 值最大，而且為正值，我們認為此時間點為一語者轉換點。

$$\arg\max_i \Delta BIC(i) > 0 \tag{8}$$

(C) 廣義概似比(Generalized Likelihood Ratio, GLR)偵測法[5]

圖 2 所示為廣義概似比偵測法的流程圖，其演算和貝氏偵測法一樣，必須先定義兩個假說測試 H_0 與 H_1，不過貝氏偵測法是移動可變時間點 i 作語者轉換偵測，廣義概似比偵測法則以兩個固定長度的語音段作語者轉換點偵測，其測量距離的式子定義如下，

$$R = \frac{L(X, N(\mu,\Sigma))}{L(X_1, N(\mu_1,\Sigma_1))L(X_2, N(\mu_2,\Sigma_2))} \tag{9}$$

X_1 與 X_2 是相鄰的兩段語音參數序列，其連接的語音訊號序列就是 $X = X_1 \cup X_2$，呈高斯分佈，即 $X \sim N(\mu,\Sigma)$。X_1 與 X_2 也是呈高斯分佈，$X_1 \sim N(\mu_1,\Sigma_1)$，$X_2 \sim N(\mu_2,\Sigma_2)$。當 R 值愈小，代表兩個相鄰的語音段愈可能為不同說說者，反之，則愈可能為同一說話者。廣義概似比偵測法最大的缺點，即比較難去定義門檻值來判斷是同一說話者或不同說話者。

2-2 本論文使用之方法

A. 偵測單一語者轉換點

本論文偵測單一語者轉換點的方法，是當語音段進行特徵參數抽取後，會先將貝氏資訊準則應用到以距離為基礎之順序偵測法(Sequential Metric-based segmentation via BIC)[6]，找出語者轉換點大概的位置，然後再透過交叉偵測法[7]，將剛才所找出的語者轉換點作精確化，也就是

讓偵測到的轉換點離真實轉換點更近，最後再確認是否為轉換點，各功能方塊描述如圖 3 所示。

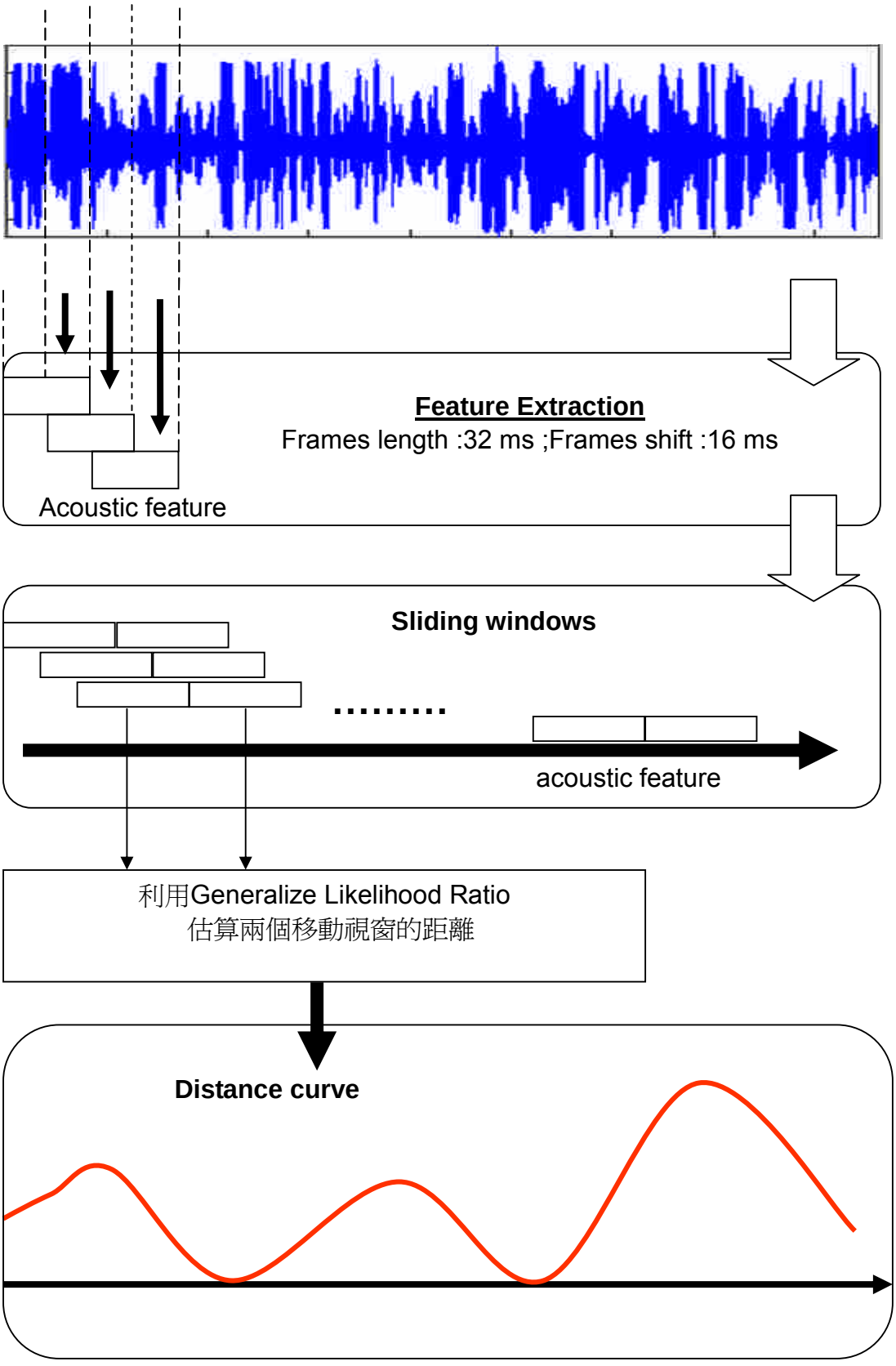


圖 2　廣義概似比偵測法之流程圖

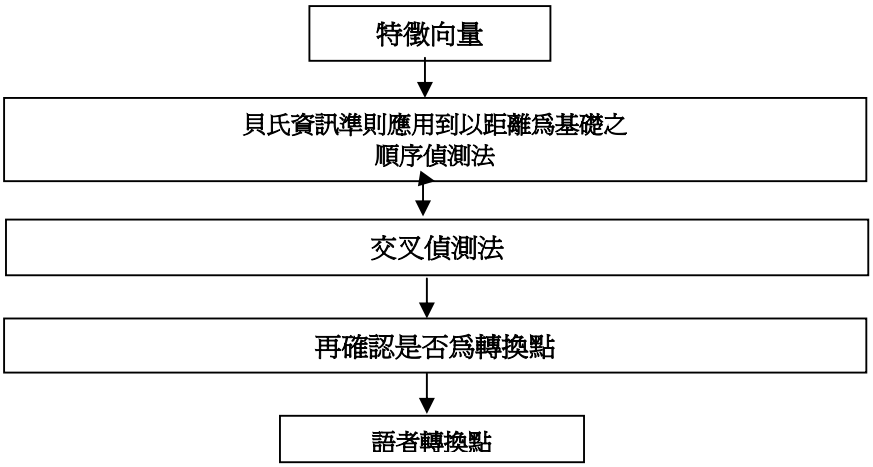


圖 3　偵測單一語者轉換點之架構流程圖

(1)貝氏資訊準則應用到以距離為基礎之順序偵測法(Sequential Metric-based segmentation via BIC)

前述的貝氏偵測法，是根據不同的時間點 i 建立兩個假說測試 H_0 與 H_1，然後計算其每個不同點 i 的 ΔBIC 值，最後由這些 ΔBIC 值決定語者轉換點。若我們將不同時間點 i 估計 ΔBIC 值的方式，改成廣義概似比固定長度的方式，來做 ΔBIC 值的估計，如圖 4 所示。

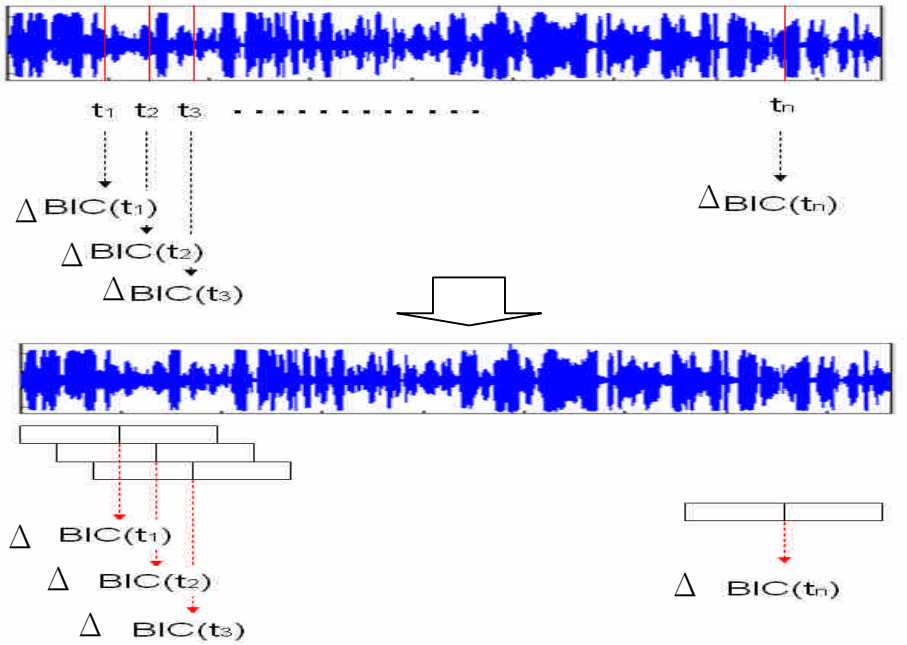


圖 4　計算方式由不同點改為固定長度之示意圖

計算方式改變後可得到下列的結果：

$$
\begin{aligned}
\Delta BIC &= BIC(H_1) - BIC(H_0) \\
&= \log pr\langle X \mid \mu x, \Sigma x\rangle + \log pr\langle Y \mid \mu y, \Sigma y\rangle \\
&\quad - \log pr\langle Z \mid \mu, \Sigma \rangle - P \\
&= \log \frac{pr\langle X \mid \mu x, \Sigma x\rangle pr\langle Y \mid \mu y, \Sigma y\rangle}{pr\langle Z \mid \mu, \Sigma \rangle} - P \\
&= GLR - P.
\end{aligned}
\tag{10}
$$

由(10)式可知，以這樣的方式估算ΔBIC值，其實就好像是計算GLR值，再多加個懲罰項P。這種方式就是貝氏資訊準則應用在以距離為基礎的偵測法(Metric-based segmentation via BIC)。

若再進一步的改成如圖5的方式來計算，則稱貝氏資訊準則應用在以距離為基礎的順序偵測法 (Sequential Metric-based segmentation via BIC)。

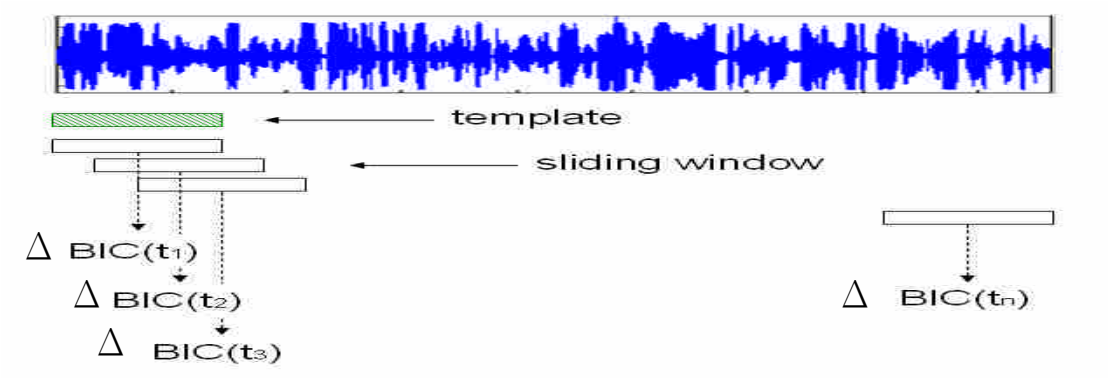


圖5 貝氏資訊準則應用在以距離為基礎的順序偵測法示意圖

首先我們取語音最開始時的短視窗(約2~3秒)作為樣式(template)，之後將此樣式和每個滑動視窗(長度和樣式相同)作ΔBIC的計算，可獲得ΔBIC的曲線，如圖6所示，

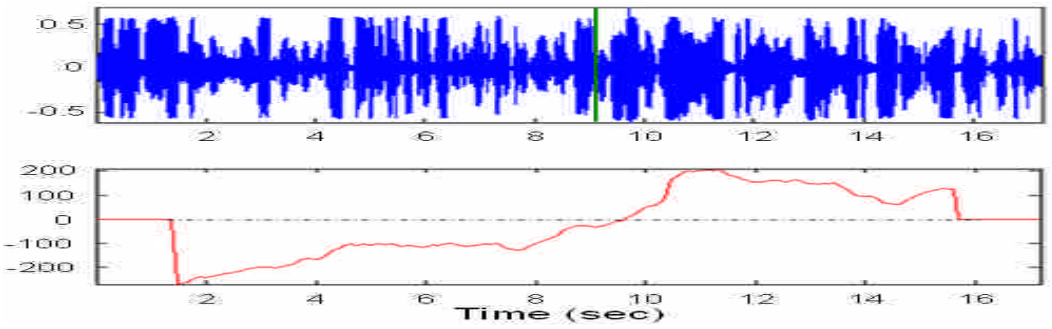


圖6 貝氏資訊準則應用在以距離為基礎的順序偵測法求出的ΔBIC曲線

由圖中觀察可發現，當滑動視窗在語者一的範圍內時，樣式和移動視窗均為語者一的聲音，所以ΔBIC值為負。當滑動視窗到達語者二的範圍內時，滑動視窗變為語者二的聲音，樣式還是語者一的聲音，所以ΔBIC值為正，這正是貝氏資訊準則的特性。在滑動視窗從語者一移到語者二時，ΔBIC值也由負變正，所以我們可以定義，在ΔBIC值為0時，其附近可能有轉換點存在。

(2)交叉偵測法

採用貝氏資訊準則應用在以距離為基礎的順序偵測法來偵測出語者轉換點後，在其轉換點向右延伸 0.5 秒處，往後抓取語者二的樣式，如圖 7 所示，其中向右延伸是為了確保抓取的樣式全包含語者二的語音訊號，而延伸長度選取 0.5 秒，是因為在實驗中發現，貝氏資訊準則應用在以距離為基礎的順序偵測法所偵測出的語者轉換點，大約在真實轉換點的左邊 0.5 秒到右邊 2 秒之間，所以只要向右延伸 0.5 秒就可確保樣式 2 包含語者二的語音訊號。

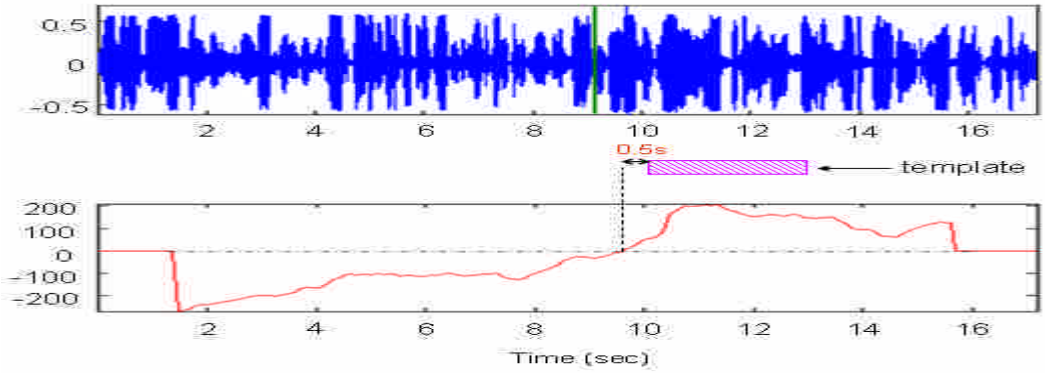


圖 7　尋找語者二的樣式

在找出語者二的樣式後，此樣式和每個滑動視窗作 ΔBIC 估算，可得一條 ΔBIC 曲線，如圖 8 中的藍色曲線，而這條曲線和原先貝氏資訊準則應用在以距離為基礎的順序偵測法所求之曲線的交叉處，即為語者轉換點的地方。會認為在曲線交叉的地方有語者轉換點存在的原因，是因為當滑動視窗移到真實轉換點時，會同時包含語者一和語者二的語音訊號，因此滑動視窗和語者一的樣式作 ΔBIC 計算以及和語者二的樣式作 ΔBIC 計算，其值會差不多。

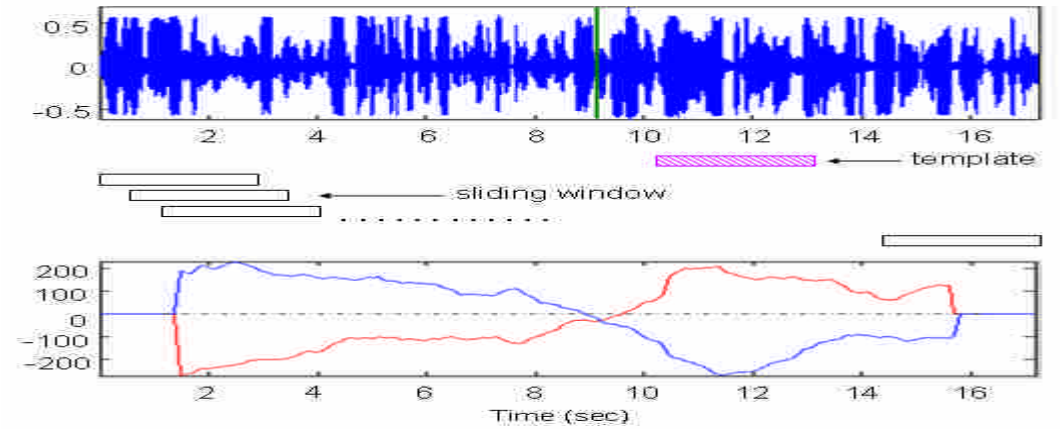


圖 8 交叉偵測法尋找語者轉換點

(3)再確認是否為轉換點

利用第二條曲線起始到交叉的這個區段(如圖 9 綠色框框部份)，將其區段中所有的 ΔBIC 值作符號函數運算後相加，若相加後值為正，則接受此點為語者轉換點，若為負，則拒絕此點為語者轉換點。

$$\begin{cases} \sum_{i=1}^{N} sign(\Delta BIC(i)) > 0 & accept \\ \sum_{i=1}^{N} sign(\Delta BIC(i)) < 0 & reject \end{cases} \tag{11}$$

其中 sign( )為符號函數。

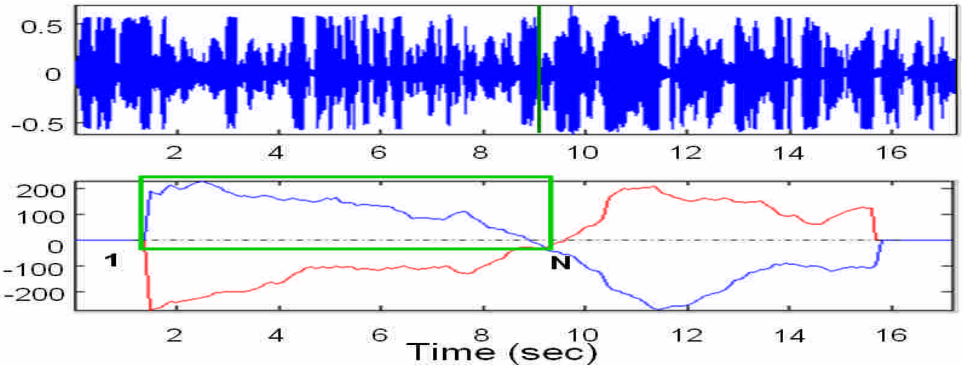


圖 9 作確認之區域圖例

利用(11)式作再確認的原因，是因為在觀察假警報錯誤的情況後，發現一些錯誤是第一條曲線有突起的狀況所造成的，導致本來應該找出語者二的樣本，結果找到語者一的樣本。一般而言，若正確找出語者二的樣本，其所求出的第二條曲線，會如圖 9 中藍色曲線所示一樣，前半段之 ΔBIC 值會大於 0，後半段之 ΔBIC 值會小於 0。若錯誤地找出語者一的樣本，就會如圖 10 中綠色曲線一樣，前半段之 ΔBIC 值小於 0，後半段之 ΔBIC 值大於 0。利用這樣的特性，第二條曲線起點到交叉點的區域，計算 ΔBIC 值是否大於 0，即可進一步確認此點有無偵測錯誤。

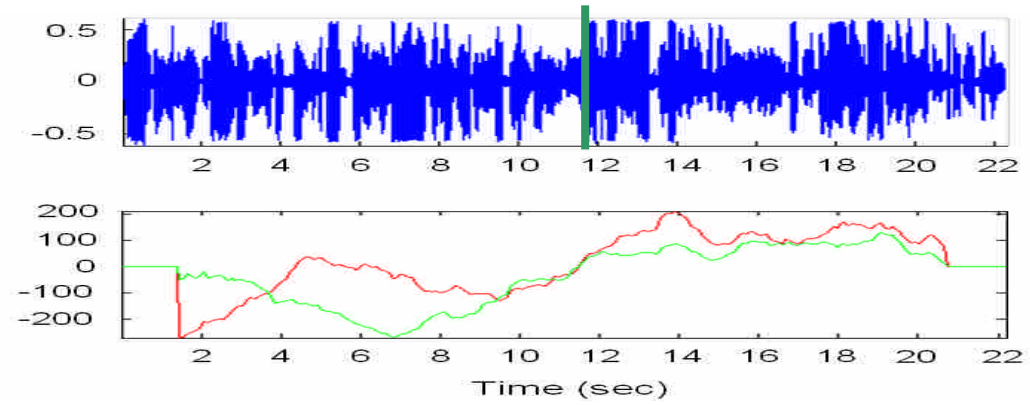


圖 10　錯誤偵測到語者轉換點之範例

B. 偵測多重語者轉換點

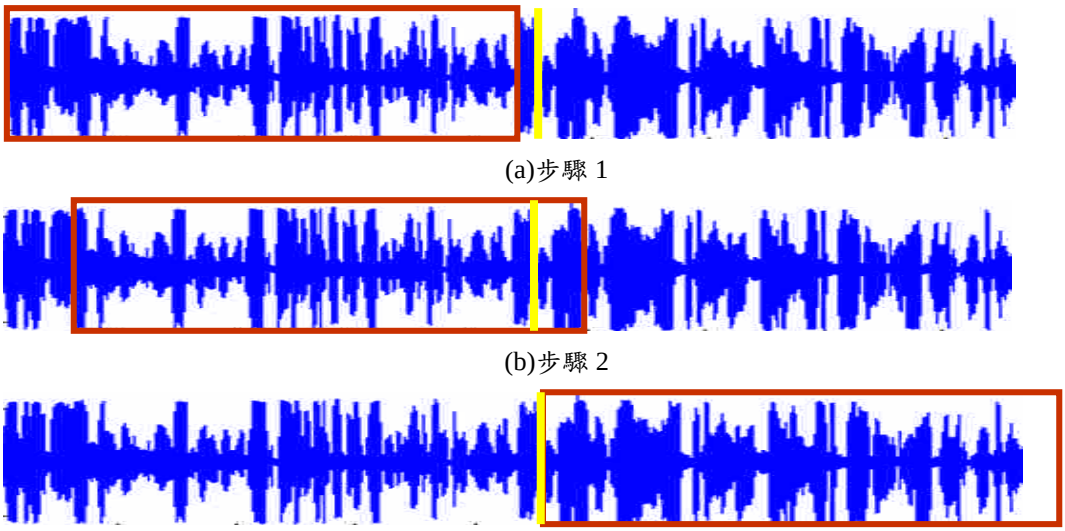
偵測單一語者轉換點的觀念，可以用來偵測多重語者轉換點。其步驟如下，

1.首先設一視窗(長度為 14 秒)，在視窗內作單一語者轉換點偵測。

2.若在上一步驟沒找到語者轉換點，則將視窗向右移動(向右移動 2 秒)，重新在視窗內作單一語者轉換點偵測。若還是沒找到轉換點，再將視窗向右移動，直到找到語者轉換點，或是直到語音結束。

3.若是找到語者轉換點，則記錄此轉換點，並將視窗的起始點設在此語者轉換點上，重新作步驟一及步驟二，直到找到下一個語者轉換點，或是直到語音結束。

圖 11 展示各個步驟之示意圖，圖中黃色線為人工標注之語者轉換點。



(a)步驟 1

(b)步驟 2

(c)步驟 3

圖 11　偵測多重語者轉換點之示意圖

三、 語者分群(speaker clustering)

語者分群系統

圖 12 為語者分群之系統架構，假設有音段 S1…Si…Sm 進入系統，其分群的步驟如下：

1. 開始時(i=1)，將音段 S1 作為第一個群集，訓練出高斯混合模型[8]。
2. 音段 Si+1 和每個群集訓練出的高斯混合模型作最大概似法計算。
3. 找出最靠近的群集。
4. 若最大概似值大於門檻值，則和最靠近的群集合併，重訓練高斯混合模型，若概似值小於門檻值，則創造新的群集，並訓練其高斯混合模型。
5. i=i+1，回到步驟 2，直到 i > m。

本論文所使用的最大概似法，有作了一點修改，就是多除了一個音段本身的長度 T，原因是因為每個音段長度都不同，導致每個音段算出的概似值無法比較，故除以音段長度 T 可以使其基準都一致，如此便可相互比較。

$$\hat{S} = \arg\max_{1\le k\le N} \frac{1}{T}\sum_{t=1}^{T} \log p\left\langle \vec{seg}_t \middle| \lambda_k \right\rangle \tag{12}$$

門檻值之選定，由實驗中得來，如圖 13 所示，藍色曲線為錯誤創新群曲線，紅色曲線為錯誤合併曲線，當門檻值設的越大，則該合併而沒合併的錯誤率越高，當門檻值設的越小，則該創新群而沒創的錯誤率越高。將門檻值選在兩錯誤曲線交叉的地方，實驗中觀察交叉的值，得知其門檻值為-41.3。

四、 實驗結果與討論

4.1 語者切割實驗

A. 實驗語料

此實驗之語料庫(Data Base)為坊間空中英語教室所轉錄出來，其取樣頻率為 44.1kHz，取樣點位元數為 16bits，由雙聲道轉為單聲道，總共有 210 個語者轉換點。以 512 取樣點為音框長度，

音框位移為 256 點，對每一個音框計算其 12 階梅爾刻度倒頻譜係數(MFCC)，做為 12 維語音特徵向量。同時也計算 MFCC 之差分值(Delta MFCC)及二次差分值(Delta Delta MFCC)，連同 MFCC 分別組成 24 維語音特徵向量，或 36 維語音特徵向量。

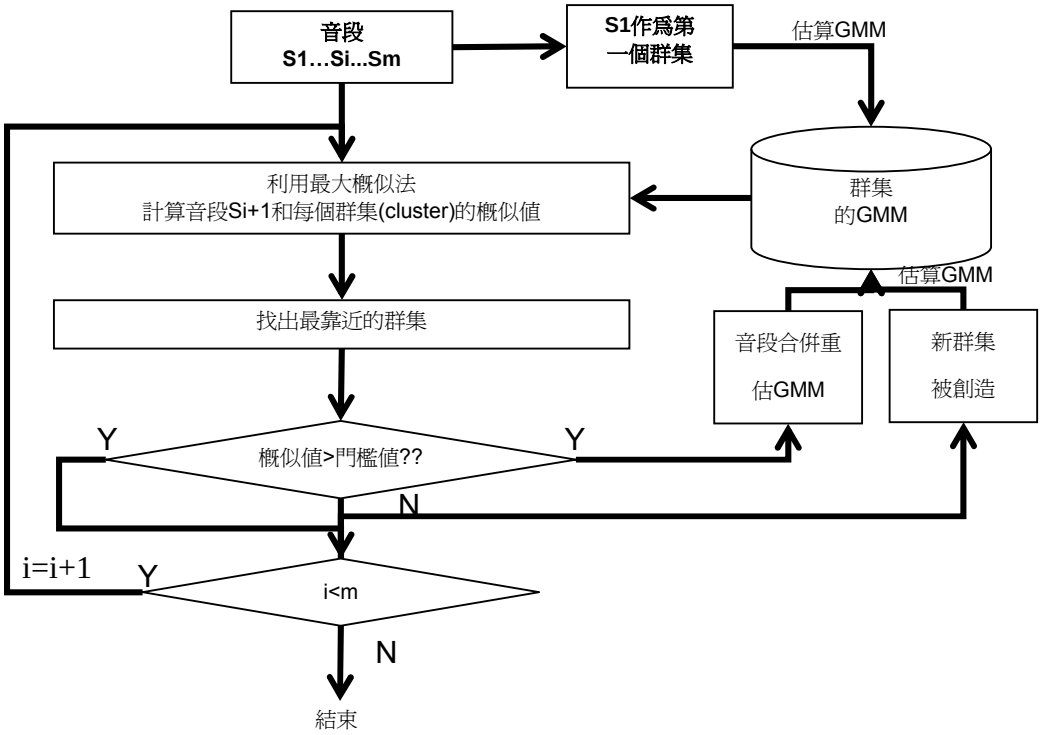


圖 12　語者分群之系統架構圖

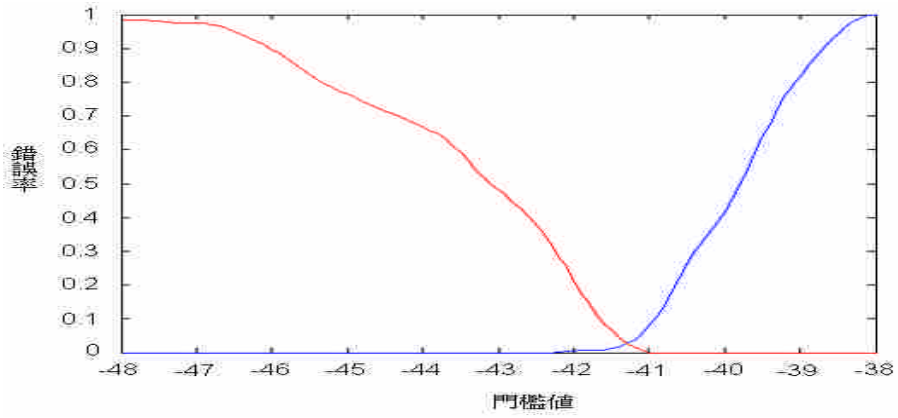


圖 13　合併與分新群之錯誤機率曲線

B. 評估方式[9]

首先定義語者切割會發生的兩種錯誤如下：

1.若在真實轉換點附近(左右 0.6 秒內)，沒偵測到轉換點，則稱這種錯誤為遺失偵測(Miss Detection)。

2.若在偵測出的轉換點附近，沒有真實轉換點存在，則稱此錯誤為假警報(False Alarm)。

利用上述的遺失偵測錯誤和假警報錯誤，可定義出召回率(Recall)和精確度(Precision)：

$$召回率(\mathrm{Re}\,call)=\frac{正確偵測的數目}{正確偵測的數目+遺失偵測} \tag{13}$$

$$精確度(\mathrm{Pr}\,ecision)=\frac{正確偵測的數目}{正確偵測的數目+假警報} \tag{14}$$

將召回率和精確度合併成為一個單一估測值，稱為 F-估測值 (F-measure)：

$$F\text{-評估值}=\frac{2\cdot 召回率\cdot 精確度}{召回率+精確度} \tag{15}$$

F 估測值愈大，代表偵測到的語者轉換點愈準確。

C. 貝氏資訊準則應用到以距離為基礎之順序偵測法與交叉偵測法所偵測到之語者轉換點落點比較

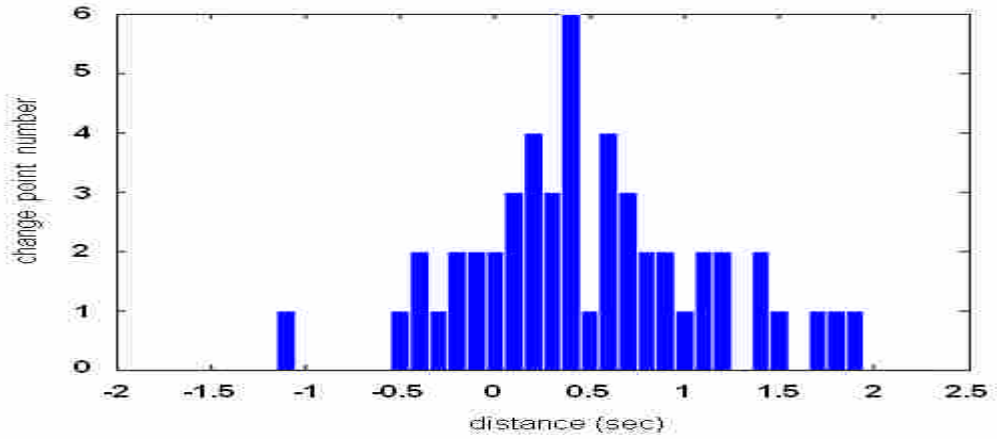


圖 14　貝氏資訊準則應用在以距離為基礎之順序偵測法所偵測出的語者轉換點落點分布之直方圖

圖 14 為貝氏資訊準則應用在以距離為基礎之順序偵測法所偵測出的轉換點落點分布情形，縱軸為轉換點數目，橫軸為偵測出之轉換點與真實轉換點間的距離，以秒為單位，由圖中發現偵測出的轉換點分布大約在真實轉換點左邊 0.5 秒至右邊 2 秒之間。

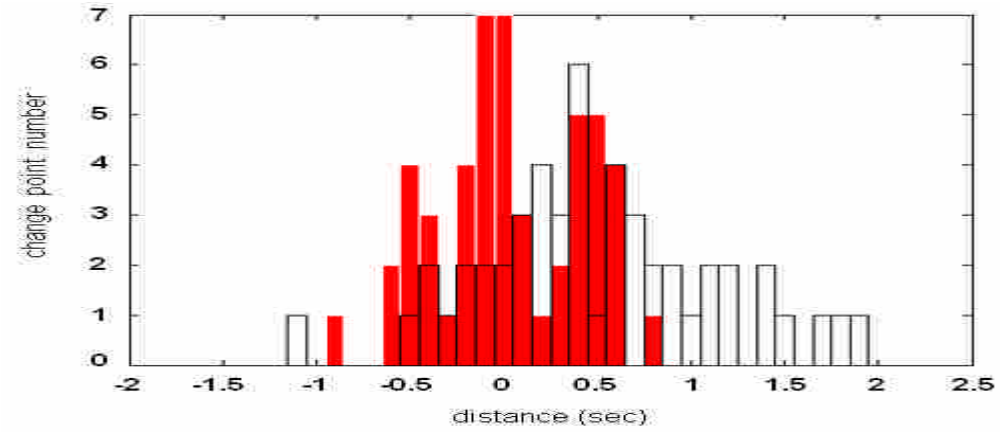


圖 15 貝氏資訊準則應用在以距離為基礎之順序偵測法與交叉偵測法所偵測出的語者轉換點分布之直方圖

圖 15 中紅色的部份為交叉偵測法所偵測出的語者轉換點落點分布直方圖，白色部份為貝氏資訊

準則應用在以距離為基礎之順序偵測法所偵測出的語者轉換點落點分布直方圖，比較兩者，可以發現交叉偵測法所偵測之語者轉換點落點範圍明顯較小，大約是在真實轉換點左邊 0.6 秒至右邊 0.6 秒之間，因此交叉偵測法可以有效將貝氏資訊準則應用在以距離為基礎之順序偵測法所偵測出的語者轉換點作更精確化的處理。

D. 判斷式之影響

在偵測單一語者轉換點時，所用的判斷式對系統效能有何影響，在此針對這個問題作實驗。

	遺失偵測數	假警報數	F-估測
沒加判斷式	18	33	88.26%
加判斷式	22	16	90.81%

表 1　有無判斷式對切割效能的影響

從表 1 中觀察得知，加判斷式後遺失偵測數會有少許的增加，但假警報數卻大量的減少，所以利用此判斷式會犧牲少數對的偵測點來減少多數的假警報數。在整體效能方面，F-估測值從 88.26% 增加到 90.81%，因此加入判斷式對系統偵測是有幫助的。

E. 與貝氏偵測法及廣義概似比法之比較

演算法	花費時間(s)	遺失偵測數	假警報數	F-估測
GLR	66.96	25	48	84.47%
BIC	6714.8	17	29	88.9%
本方法	532.57	22	16	90.81%

表 2　與貝氏偵測法及廣義概似比法所需時間、假警報數、
遺失偵測數與 F-估測之比較

從表 2 中的實驗數據可以發現，廣義概似比法偵測所花費的時間相當短，但對於整體的偵測效能並不是很好。貝氏偵測法剛好和廣義概似比法相反，整體的偵測效能不錯，但偵測所花費的時間相當長。本論文方法，在整體效能方面均優於廣義概似比法和貝氏偵測法，在花費時間方面，雖然比廣義概似比法慢，但卻比貝氏偵測法快的多。

F. 不同特徵維度之影響

在這個實驗中採用不同的特徵維度，觀察本方法在不同特徵維度時，對於語者轉換點偵測是否有不同的影響。

	12 維	24 維	36 維
假警報數	16	18	13
遺失偵測數	22	28	35
F-估量	90.81%	88.78%	87.93%

表 3　不同的特徵維度對偵測效能之影響

由表 3 可看出，增加特徵參數維度並無法增加偵測的成效，尤其在遺失偵測數上，特徵維度越高，遺失偵測數也越多，因此增加特徵維度不但使運算量變大，且對於偵測的效果也沒幫助。

G. 不同取樣頻率之影響

這個實驗是對本論文所使用的方法，觀察其在不同取樣頻率下的效果。

	44.1kHz	32kHz	16kHz	8kHz
假警報數	16	14	39	23
遺失偵測數	22	25	23	50
F-估測	90.81%	90.45%	85.77%	81.42%

表 4　不同取樣頻率對偵測效能之影響

由表 4 可知，本方法在取樣頻率高時可得到較好的成效，在取樣頻率為 32kHz 與 44.1kHz 的情況下，其 F-估測值並不會差太多，而在取樣頻率為 16kHz 時，其 F-估測值已有明顯的下降，在取樣頻率為 8kHz 時，F-估測值降到 81.42%，非常不理想。

4.2 語者分群實驗

A. 實驗語料

本實驗採用三個測試檔案，如表 5 所示，取樣頻率為 44.1kHz，取樣點位元數為 16bits，由雙聲道轉為單聲道。

檔案	音段數	語者數
檔案一	63	3
檔案二	74	5
檔案三	88	7

表 5　三個測試檔案的音段數及語者數一覽表

B. 評估方式[10]

我們計算以下兩個數據，作為評量指標。

1.平均群集純度(Average Cluster Purity,　ACP)

$$p_m = \sum_{j=1}^{R} \frac{n_{mj}^{\ 2}}{n_m^{\ 2}} \qquad acp = \frac{1}{N}\sum_{m=1}^{M} p_m \cdot n_m \tag{16}$$

2.平均語者純度(Average Speaker Purity,　ASP)

$$p_j = \sum_{m=1}^{M} \frac{n_{mj}^{\ 2}}{n_j^{\ 2}} \qquad asp = \frac{1}{N}\sum_{j=1}^{R} p_j \cdot n_j \tag{17}$$

其中 M 為群集的數目，R 為語者的數目， N 為音段的數目。n_m 為第 m 個群集裡的音段數目。n_{mj} 為第 m 個群集裡由第 j 個語者所講的音段數目。n_j 為第 j 個語者所講的音段數目。

進一步的將上述的 acp 與 asp 作幾何平均數，可得到一個單一參數 K：

$$K = \sqrt{acp \cdot asp} \tag{18}$$

以上幾個評估參數的特性如下：

(1)平均群集純度愈高代表該群集包含人數愈接近一。

(2)平均語者純度愈高代表該語者被分配到的群數愈接近一。

(3)K 的值愈大，整體分群效能愈好。

C. 高斯混合數對分群之影響

分別對高斯混合數 2、4、8、16 及 32 作實驗，觀察高斯混合數對 K 值有什麼影響。由圖

16 中可觀察到，隨著高斯混合數的增加，其分群的效能也愈好，而當高斯混合數等於 16 時，*K* 值已達最好結果，繼續增加高斯混合數，*K* 值並不會再增加。

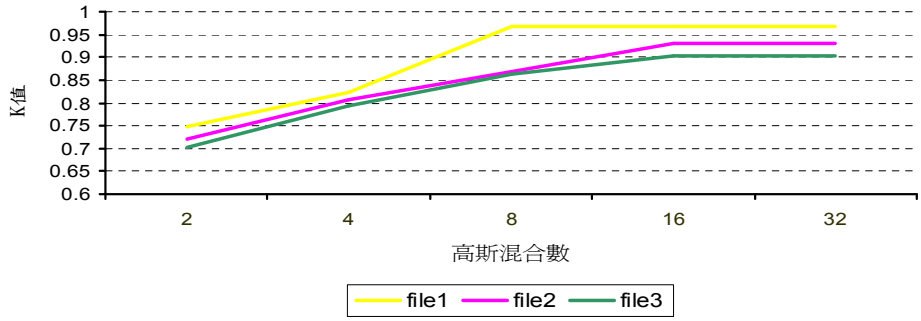


圖 16　高斯混合數對分群之影響

D. 各檔案分群實驗結果

實驗中得知在高斯混合數為 16 時，整體的分群效果已達最好，因此在這個實驗中，將高斯混合數設在 16，來觀察檔案一、檔案二、與檔案三的分群結果。

	實際語者數	群集數	平均群集純度	平均語者純度	K 值
檔案一	3	3	0.970	0.969	0.969
檔案二	5	7	0.974	0.887	0.929
檔案三	7	10	0.960	0.852	0.903

表 6　各檔案分群實驗結果

從表 6 中得知，在包含三個語者的檔案一中，作分群後正確分出三個群集，在包含五個語者的檔案二中，作分群後分出七個群集(多兩個群集)，在包含七個語者的檔案三中，作分群後分出十個群集(多三個群集)。由此可見，當檔案包含的語者數愈多，會錯誤多分出的群集也愈多，而這樣的原因也導致平均語者純度隨語者數的增加而下降。在整體分群效能方面，雖然平均群集純度不太受語者數多寡的影響，但由於平均語者純度的關係，使得 *K* 值也是隨語者數的增加而下降。

五、 結論

本論文探討錄音資料中之語者切割與分群，在語者切割方面，交叉偵測法的確是修正了貝氏資訊準則應用在以距離為基礎之順序偵測法所偵測到的語者轉換點，也証明了利用判斷式作再確認的動作，能有效使假警報數下降。本方法與廣義概似比偵測法及貝氏資訊準則偵測法的比較，從實驗數據中發現，廣義概似比偵測法偵測轉換點花費的時間少，但偵測效能比較差，而貝氏資訊準則偵測法是偵測效能好，但偵測轉換點花費的時間相當長，本方法花費的時間雖比廣義概似比偵測法稍長，但比貝氏資訊準則偵測法卻短很多，且偵測效能為三者之冠，可說是同時擁有廣義概似比偵測法運算量少的優點及貝氏資訊準則偵測法高準確率的優點。在實驗中也發現，增加特徵維度對於整體切割效果並沒有幫助，反而使偵測的效能往下掉，而且偵測轉換點花費的時間

也增多。另外，在取樣率為 32kHz 及 44.1kHz 時，其結果差不多，都有不錯的偵測效能。

在語者分群部份，主要針對 3 個測試檔案做實驗，在實驗中可發現，增加高斯混合數對分群的結果是有幫助的，高斯混合數等於 16 時，其結果已達最好。當要分群的音段群中包含語者數愈多，其整體分群效能愈低。

致謝

本研究受國科會專題研究計畫補助，計畫編號 NSC-93-2213-E-007-019。

參考文獻

[1] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," *DARPA Speech Recognition Workshop*, 1998.

[2] 詹順凱,"在多語者環境下之語者分割與語言辨認研究", 電機工程研究所，國立清華大學，中華民國九十一年六月。

[3] Y. Moh, P. Nguyen, and J.-C. Junqua, "Towards domain independent Speaker clustering," *Proc. ICASSP 2003*, pp. I-85-88.

[4] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461-464, 1978.

[5] J.F. Bonastre, P. Delacourt, C. Fredouille, "A Speaker Tracking System Based On Speaker Turn Detection For NIST Evaluation," *Proc. ICASSP 2000*, paper no. 1628.

[6] S. S. Cheng and H. M. Wang, "A sequential metric-based audio Segmentation method via the Bayesian Information Criterion," *Proc. Eurospeech 2003*, pp. 945-948.

[7] A. Adami, S. Kajarekar and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," *Proc. ICASSP 2002*, pp. IV-3908-3911.

[8] D. Reynolds and R. Rose, "Robust test-independent speaker identification using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, Vol.3, No.1, 1995.

[9] J. Ajmera, I. McCowan, and H. Bourlard, "Robust Speaker Change Detection," *IEEE Signal Processing Letters*, pp. 649-651, Vol. 11, No. 8, pp. 649-651, August.2004

[10] I. Lapidot, "SOM as Likelihood Estimator for Speaker Clustering," *Proc. Eurospeech 2003*, pp. 3001-3004.

結合聲學與韻律訊息之強健性語者辨認方法

Combination of Acoustic and Prosodic Information for Robust Speaker Identification

[1]廖元甫，[1]莊智顯，[2]陳子和，[2]莊堯棠

Yuan-Fu Liao, Zhi-Xian Zhuang, Zi-He Chen and Yau-Tarng Juang

[1]Department of Electronic Engineering & Institute of Computer, Communication and Control, National Taipei University of Technology

[2]Department of Electrical Engineering, National Central University, Chung-Li, Taoyuan, 32054, Taiwan, [1] yfliao@ntut.edu.tw

摘要

語者辨認系統在公共電話網路中，通常會遇到未知不匹配話筒和辨認語料不足的問題。為增進語者辨認系統對未知話筒之強健性，與有效利用有限語料，我們提出一融合下層聲學與上層韻律訊息之架構，首先利用(1)最大相似先驗知識內插法（maximum likelihood-*a priori* knowledge interpolation，ML-AKI）方法估計與補償話筒聲學特性，並以(2)最小錯誤鑑別式法則(Minimum Classification Error, MCE)訓練語者模型，以拉大不同語者間聲學模型的距離，與利用(3)韻律訊息特徵分析(eigen-prosody analysis, EPA)為輔助，量測不同語者間的韻律模型距離，最後利用(4)線性迴歸的方式融合聲學與韻律模型分數得到最後的辨識結果。

實驗使用 Handset TIMIT（HTIMIT）語料庫，以 leave-one-out 方式輪流使用九種不同的話筒當作未知話筒，驗證所提出之方法。實驗結果顯示，在有限的訓練與辨認語料情形下，若以傳統 maximum *a priori* probability adapted Gaussian mixture model/cepstral mean subtraction（MAP-GMM/CMS）的方法當作 baseline，其平均語者辨認率可達 60.2%。但若結合 ML-AKI，MCE，EPA 與 MAP-GMM/CMS 方法，則平均辨認率可提升到 79.3%。而若只觀察未知話筒部份，則平均語者辨識率亦可由 58.3%提升到 74.6%，因此可知所提出之方法無論對已知話筒和未知話筒皆能有效改善系統之強健性。

1. 緒論

語者辨認系統在公共電話網路中，通常會遇到話筒不匹配和訓練／辨認語料不足的問題，尤其是當遇到不匹配的話筒，且其特性在事前無法得知時(未知話筒)，系統效能通常會

劇烈下降。為抵抗未知且話筒特性不匹配的問題，近年來的相關研究【1】，常嘗試結合聲學與韻律兩層次的訊息，包括在聲學層次作話筒特性補償，與使用較不受話筒特性影響的韻律訊息來幫助系統辨認語者。

在聲學訊息層次上，傳統上常使用 Cepstral Mean Subtraction (CMS)【2】、Signal Bias Removal (SBR)【3】及 handset detector【4】等方法補償話筒不匹配效應。而在韻律訊息層次上，常使用 Gaussian mixture models (GMMs)【5】，描述音高軌跡（pitch contour）的短程(short-term)變化，或是使用 N-gram 和 discrete hidden Markov model(DHMM)【5】去表現韻律訊息隨時間的長程（long-term）變化。

然而 CMS 和 SBR 不單只是移除話筒的特性，常也會把語者的特性移除。而基於話筒偵測的方法，遇到測試語音來自未知話筒時，通常只能從已知話筒集合中選擇出一個最相似的話筒，或是直接把它拒絕掉。而使用 GMMs 統計韻律訊息時，一般只能補捉到音高與能量變化等短程的韻律訊息，DHMM 和 N-gram 的方法，雖可以補捉到較長程的韻律訊息變化，但通常得使用大量的訓練/測試語料。

因此在本論文中將針對不匹配未知話筒和訓練／辨認語料不足的問題，在聲學層次以最佳先驗知識內差（Maximum likelihood *a prior* knowledge interpolation，ML-AKI）方法，事先收集先驗話筒知識，再以內差方式估計補償未知話筒的特性，在韻律訊息層次則以韻律特徵值分析（Eigen-prosodic analysis，EPA）方式利用韻律訊息，降低所需估計之參數數目，以減少所需的訓練／辨認語料。最後並融合聲學層次和韻律層次的語者訊息，以加強語者辨認系統對未知話筒不匹配效應的強健性。

其中 ML-AKI 主要是利用 Maximum likelihood linear regression (MLLR)【6】事先估算多組已知話筒之轉換函數，當作先驗知識，測試時以 Maximum likelihood (ML)方法，找出最佳的內插權重組合，估計出測試話筒的特性轉換函數，再以 MLLR 調適語者模型。EPA 主要做法是將語者辨認問題轉化成為文件擷取（document retrieval）問題。首先把語者的韻律特徵參數變化，自動標記成韻律狀態序列，當作一虛擬文件，再運用 latent semantic analysis (LSA)【7】作特徵分析，建立一個特徵韻律訊息空間，以表現不同語者的分佈 (constellation)，最後利用韻律訊息關鍵詞作詢問(query)，以擷取最相似的註冊語者。

本論文其餘內容的安排如下:第二節介紹在聲學層次上提出的 ML-AKI 方法，第三節在韻律層次上提出的 EPA 方法，第四節融合所提出 EPA，ML-AKI，MCE 和傳統 CMS 方法，第五節介紹未知話筒不匹配效應補償實驗，第六節則作一簡單總結。

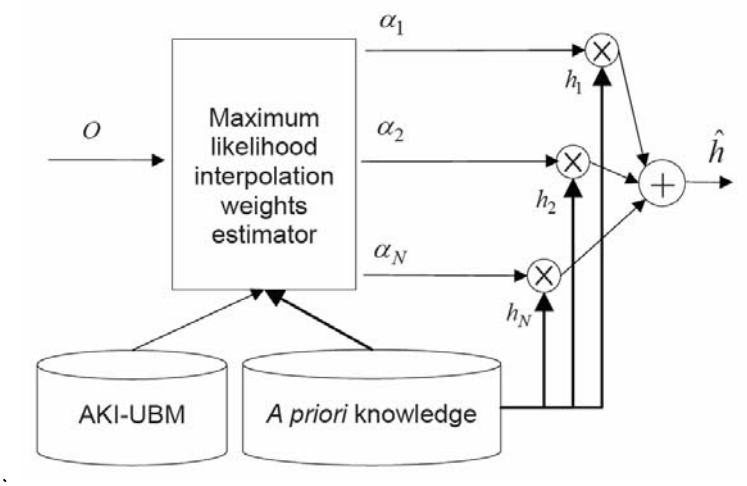
2. 最大相似度先驗知識內差法（ML-AKI）

為補償未知話筒不匹配特性效應，ML-AKI 先收集已知話筒特性集合當作先驗知識，在測試時，則以此先驗知識做線性組合，如式(1)所示，以估計補償未知測試話筒特性，其中先知識內插的最佳權重值的求取，則利用期望值最大化演算法，如圖一所示：

$$\tilde{h}_n = \sum_{n=1}^{N} \alpha_n h_n \qquad (1)$$

其中 α_n 為內插的權重，h_n 為模型領域上的先驗知識。

以下在 2.1 節中將先介紹以 MLLR 方式求取先驗知識，在 2.2 節中則利用 expectation-maximization (EM)【8】演算法求取最佳的先驗知識內插權重，以補償未知測試話筒的特性。



圖一、ML-AKI 的架構圖

2.1. 基於 MLLR 之話筒特性先驗知識

MLLR 是一種模型調適的方法，主要目的是求得一組模型參數轉換函數以調適聲學模型，使其適合辨認測試語料。但在本論文中則先建立 N 個已知話筒與註冊話筒之 GMM 模型，並使用 MLLR 量測此 N 個已知話筒 GMMs 與註冊話筒 GMM 之間的轉換函數，當作話筒特性的先驗知識，以建構話筒特性空間。

MLLR 轉換函數有幾種不同的型式，比較常用的方法是調適平均值及變異數，其轉換函數如式(2.a)與(2.b)所示：

$$\hat{u}_m = \hat{A}_m \cdot u_m + \hat{b}_m \tag{2.a}$$

$$\hat{\Sigma}_m = B_m^T \hat{H}_m B_m \tag{2.b}$$

其中 m 為模型中高斯混合分佈的索引，$\hat{u}_m$ 為調適過的平均值，u_m 為原本模型的平均值，$\hat{A}_m$ 為平均值的轉換函數矩陣，$\hat{b}_m$ 為偏移量；$\hat{\Sigma}_m$ 為調適過的變異數，$\hat{H}_m$ 為變異數的轉換函數，B_m 為 $\hat{\Sigma}_m^{-1}$ 的 Choleski factor 的逆函數，所以

$$\Sigma_m^{-1} = C_m C_m^T \tag{3.a}$$

$$B_m = C_m^{-1} \tag{3.b}$$

在本篇論文中，將使用 MLLR 量測 N 個已知話筒 GMMs 與註冊話筒 GMM 之間的轉換函數，再以此 N 組轉換函數集合，當作話筒特性的先驗知識，即 $\left\{W_{n,m}, \hat{H}_{n,m}, n=1\sim N\right\}$，其中 n 為已知話筒的索引，$W_{n,m} = \left[\hat{b}_{n,m} \;\; \hat{A}_{n,m}\right]$。

2.2. 最佳化內插權重值求取

為補償未知測試話筒的不匹配特性，我們在測試時利用事先求取的話筒先驗知識 $\left\{W_{n,m}, \hat{H}_{n,m}, n=1\sim N\right\}$，以內插方式估計未知測試話筒特性的轉換函數，以調適語者辨認模型，其中內差轉換函數的方式如式(4.a)與(4.b)所示：

$$\vec{W} = \sum_{n=1}^{N} \alpha_n W_n \tag{4.a}$$

$$\vec{H} = \sum_{n=1}^{N} \alpha_n \hat{H}_n \tag{4.b}$$

以下將依據 ML 準則，以 EM 演算法求取最佳內插權重值，調適語者辨認模型，以補償未知測試話筒的不匹配特性。若使用 GMM 語者辨認模型，且只考慮調適 GMM 模型的平均值，則定義 likelihood function 如下：

$$P(o_t \mid \Phi, \Lambda) = \sum_{m=1}^{M} c_m \mathbb{N}(o_t \mid \sum_{n=1}^{N} \alpha_n W_n \mu_m, \Sigma_m) \tag{5}$$

其中 $O = \{o_1 ... o_T\}$ 為測試語者的觀測值序列，M 為 GMM 的混合高斯數目，c_m 為第 m 個混合高斯所佔的權重。

接著利用期望值最大化演算法，求取最佳先驗知識內差權重 $\hat{\alpha}_n$，定義輔助方程式 $Q(\Phi, \hat{\Phi})$ 如下：

$$Q(\Phi, \hat{\Phi}) = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_m(t) \log \mathbb{N}(o_t \mid \sum_{n=1}^{N} \hat{\alpha}_n W_n \mu_m, \Sigma_m) \tag{6}$$

其中 $O = \{o_1 ... o_T\}$ 為語音特徵參數，Φ 和 $\hat{\Phi}$ 分別為舊和新的內差權重值，$\gamma_m(t)$ 為第 m 個混合高斯的 occupation 機率，其公式如下：

$$\gamma_m(t) = \frac{c_m P_m(o_t \mid \Phi, \Lambda)}{\sum_{m=1}^{M} c_m P_m(o_t \mid \Phi, \Lambda)} \tag{7}$$

若忽略和 $\hat{\alpha}_n$ 無關的項，則可將式子(6)簡化表示如下：

$$M(\hat{\alpha}_1,\hat{\alpha}_2,...,\hat{\alpha}_N)=-\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\left(o_t-\sum_{n=1}^{N}\hat{\alpha}_n W_n\mu_m\right)^T\Sigma_m^{-1}\left(o_t-\sum_{n=1}^{N}\hat{\alpha}_n W_n\mu_m\right) \tag{8}$$

由於內插的權重受限於$\sum_{n=1}^{N}\alpha_n=1,\alpha_n\geq 0,n=1\sim N$，式（8）為一具限制條件之非線性最佳化問題（constrained nonlinear programming, constrained NLP），不好求解，所以先定義了一組新的變數做轉換將限制暫時移除，其轉換公式如下：

$$\hat{\beta}_n=\log\hat{\alpha}_n,\ n=1\sim N \tag{9}$$

接著可將式子(8)表示如下

$$M(\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_N)=-\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\left(o_t-\sum_{n=1}^{N}e^{\hat{\beta}_n}W_n\mu_m\right)^T\Sigma_m^{-1}\left(o_t-\sum_{n=1}^{N}e^{\hat{\beta}_n}W_n\mu_m\right) \tag{10}$$

則藉由使得$\frac{\partial M(\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_N)}{\partial\hat{\beta}_n}=0,n=1\sim N$，可以得到一組聯立方程式如下：

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\left[(e^{\hat{\beta}_n}W_n\mu_m)^T\Sigma_m^{-1}(o(t)-\sum_{j=1}^{N}e^{\hat{\beta}_j}W_n\mu_m)\right]=0,n=1\sim N \tag{11}$$

若解出聯立方程式，可得到一組新的內插權重$\hat{\beta}_n^*$，最後再將$\hat{\beta}_n^*$轉回$\hat{\alpha}_n^*$，則可以求出新的內插權重值，其轉換式如下：

$$\hat{\alpha}_n^*=\frac{e^{\hat{\beta}_n^*}}{\sum e^{\hat{\beta}_n^*}},\ n=1\sim N \tag{12}$$

最後反覆執行 EM 演算法，直到所求的內差權重值收斂為止，即求到最佳的內插權重值，代入式子(4.a)與(4.b)，可得到一組調適後的 GMM 語者辨認模型。

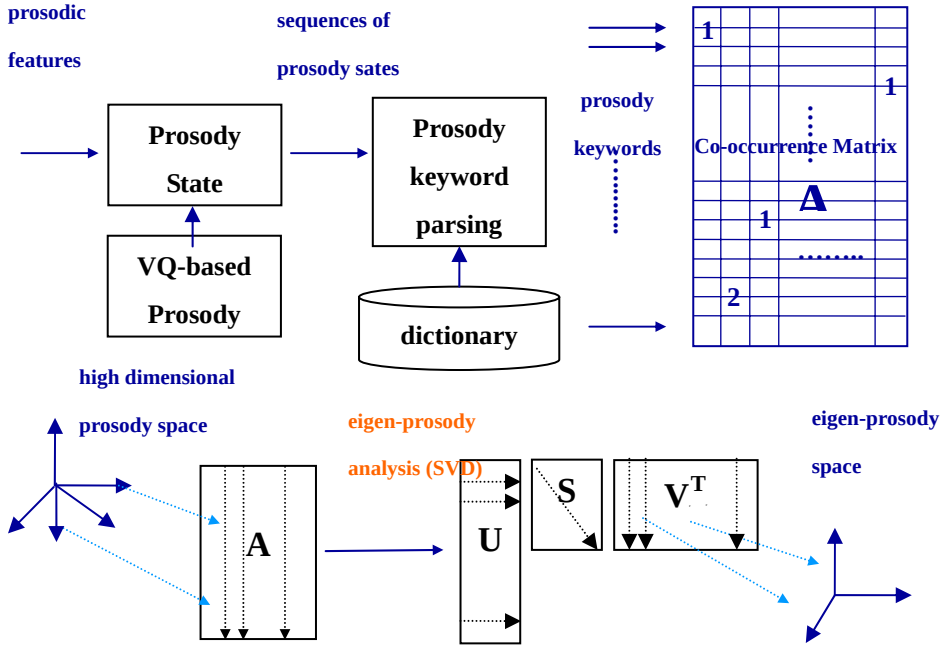
3. 韻律特徵分析（EPA）

為進一步補償未知話筒不匹配特性效應，我們使用較不受話筒影響的韻律訊息，而為了減輕一般韻律訊息模型，如 bi-gram 或是 DHMM，需要大量訓練與測試語料的問題，我們提出 EPA 方式，在有限的訓練／辨認語料限制下，利用韻律訊息，其方塊圖如圖二所示。

EPA 的作法主要是將語者辨認問題轉換為類似文件擷取（document retrieval）的問題。首先把語者的韻律特徵參數，利用一以 vector quantization（VQ）建立的韻律模型，自動標記成韻律狀態序列，當作一虛擬文件。再利用出現頻率較高的韻律序列組合，當成韻律關鍵詞。接著利用所得到的韻律關鍵詞建立韻律關鍵詞詞典，並利用建立好的韻律關鍵詞詞典，剖析進來的虛擬文件，建立語者—韻律關鍵詞關係矩陣。最後運用 latent semantic analysis（LSA）作分析，建立一個特徵韻律訊息空間，以表現不同韻律行為特徵語者的分佈（constellation），最後利用

韻律訊息關鍵詞作詢問(query)，以擷取最相似的註冊語者。

以下在 3.1 節中介紹 VQ 韻律模型與自動韻律狀態標記，在 3.2 節中介紹特徵韻律分析的詳細步驟，在每個章節中並將使用 HTIMIT 語料庫【9】，以從 ESPS 軟體演變而來的 snack 軟體【10】，求取其音高與能量軌跡，並使用 TIMIT 語料庫所提供的切割位置，做初步的實驗（HTIMIT 語料庫與實驗詳細資料請見第五節），說明所有步驟的物理意義。



圖二、EPA 的架構圖

3.1. VQ 韻律模型與自動韻律狀態標記

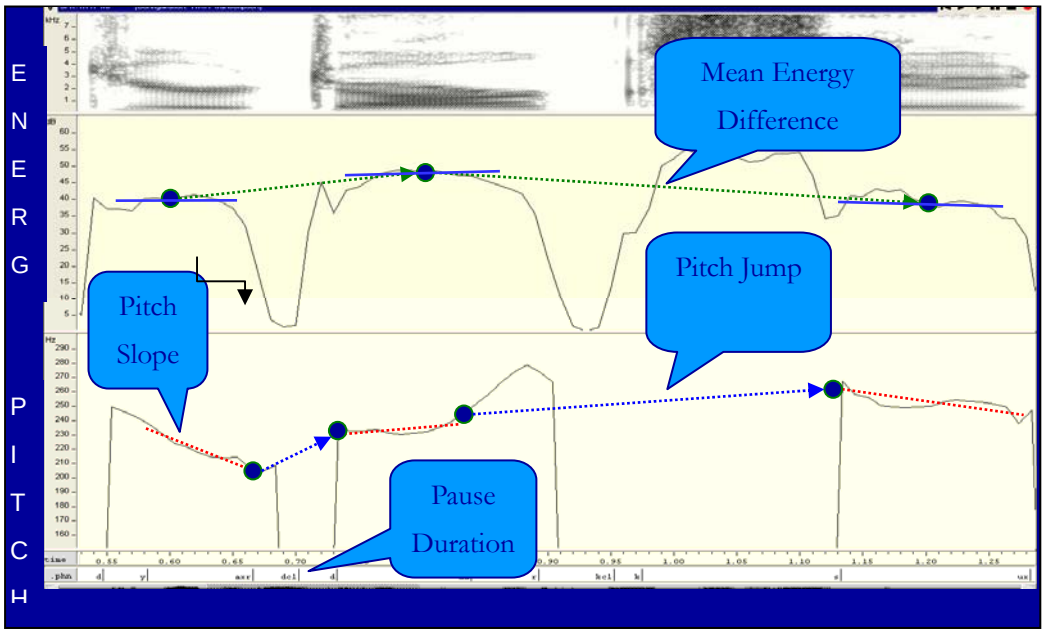
3.1.1. 韻律特徵參數求取與正規化

因為音節為最小的韻律單位，我們採用五種音節層次的韻律特徵參數，包括（1）一個母音區段的音高斜率(pitch slope)和長度的延長變化(lengthening factor)，(2) 兩個母音間的對數能量(log-energy)差和音高跳躍(pitch jump)值與（3）兩個音節間的語音暫停長度(pause duration)。其求取方式如圖三所示。

此外為移除語句發音內容（context-information）對韻律變化的影響，必須根據所處音節的母音類型，對這些韻律特徵參數做正規化的動作，以移去任何非韻律特性的影響，其正規化公式如下：

$$\hat{x} = \frac{x - u_{vowel}}{\sigma_{vowel}} \tag{13}$$

其中 x 為韻律特徵參數，u_{vowel} 和 σ_{vowel} 對整個語料庫根據所處音節的母音類型所求的平均值和變異數，而 $\hat{x}$ 為韻律特徵參數經過正規化所得到的結果。



圖三、音節層次之韻律參數求取

3.1.2. 以 VQ 為基礎之韻律訊息模型

接下來，我們將做過正規化處理的韻律特徵參數利用 VQ 分群方式，以 EM 演算法分成 M 個 codewords，則每個 codeword 可視為一特定韻律狀態，據此建立一韻律模型。

為說明以此方式建立之韻律模型的物理意義，以下初步利用 HTIMIT 語料庫中註冊話筒（senh）的語音資料，建立一個 8-codewords 韻律模型，其每個 codewords 的質心值如表一所示，其轉移矩陣的分佈(如表二)。經檢查每個 codewords 的質心值，統計每個 codeword 在句子中出現的位置，和交叉驗證轉移矩陣之後，可以大概知道這些 cordwords(在這之後我們把他統稱韻律狀態(prosodic state))的物理含意。舉例來說，狀態 6 最可能出現在句首，因此狀態 6 可以表示韻律片語起頭（phrase-start）狀態，狀態 3 和 4 的 pause duration 與 lengthening 最長，且 pitch jump 與 energy difference 最大，所以狀態 3 和 4 可以表示主要與次要中斷(major or minor break)狀態，由此可證明由 VQ 的方法，所得到的每個狀態是具有物理意義的。

表 一：利用 HTIMIT 語料庫以註冊話筒(senh)之語音資料訓練出之 8-state VQ 韻律模型。

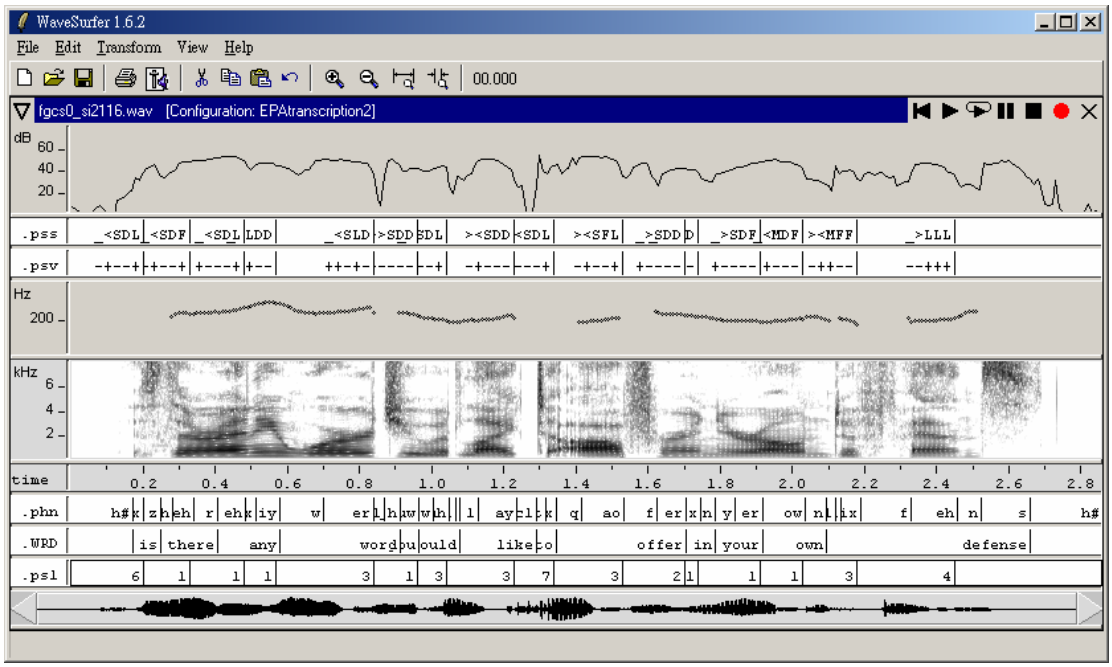
Feature/State	1	2	3	4	5	6	7	8
Pitch slop	-0.1	0.7	-0.1	-0.2	0.1	0.3	-0.2	-2.5
Energy diff.	-0.4	-0.5	-0.8	-1.9	-0.1	0.2	1.3	0.1
Pitch jump	-0.2	-0.2	1.3	1.4	-0.1	-0.9	0.3	-0.6
Lengthening	0.3	-0.5	0.3	1.4	0.1	-0.1	0.1	0.1
Pause	0.4	-0.5	0.5	2.6	0.2	-0.3	0.3	0.1

表 二：利用 HTIMIT 語料庫註冊話筒(senh)語音資料訓練出 8-state 之 VQ 韻律模型狀，其狀態轉移矩陣統計。

Previous / Next	1	2	3	4	5	6	7	8
1	3424	1256	854	429	1059	2783	919	304
2	1304	599	255	209	451	1282	344	192
3	347	122	77	55	109	405	109	43
4	20	18	5	3	18	50	10	3
5	1074	510	237	167	348	894	286	91
6	3218	1544	621	364	1005	2804	891	351
7	882	392	255	102	330	829	416	162
8	331	180	100	63	95	349	129	98

3.1.3. 自動韻律訊息標記

利用建立好的 VQ 韻律模型，即可以自動將輸入之韻律特徵參數軌跡標記成韻律狀態索引序列。以一輸入測試句子來說，其利用韻律模型做自動韻律訊息狀態標記的結果如圖四所示，可看出標示結果符合預期，即 state 6 確實出現在句首(韻律片語起頭)，state 4 則出現在句尾(major break)。



圖四、典型的自動韻律狀態標記範例。

3.2. EPA 分析步驟

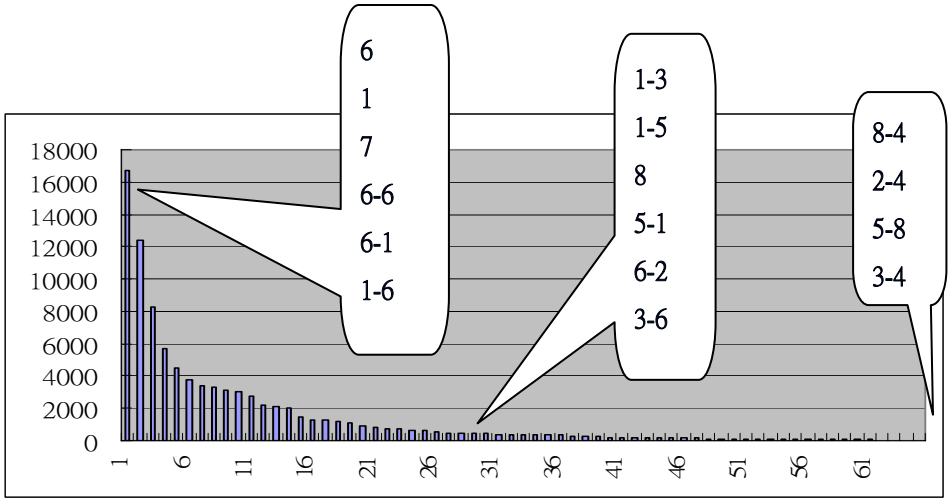
利用 EPA 輔助語者辨認包括四個步驟，包括(1)從所有語者的韻律狀態序列擷取韻律關鍵

詞建立韻律關鍵詞詞典，(2)利用韻律關鍵詞詞典，剖析所有語者的韻律狀態序列做斷詞，統計語者—韻律關鍵詞關係矩陣。(3)利用 SVD 分解此語者—韻律關鍵詞關係矩陣，取前 K 個較大的奇異值來近似，找出一低維度的語者韻律特徵空間，與（4）利用測試語者的測試語句子轉成韻律關鍵詞，投影至語者韻律特徵空間做語者辨認分數的測量。以下詳細說明各步驟的作法。

3.2.1. 韻律關鍵詞詞典

首先將標記好的語者的韻律狀態標記序列，當作一文字文件（韻律文件），並對所有可能發生的韻律標記序列組合，包含單字詞(single words)和雙字詞(word pairs)，統計其發生次數，得到所有可能韻率標記序列組合發生頻率的長條統計圖。接著設定一發生頻率臨界值，擷取所有超過頻率臨界值的韻率標記序列組合作為韻律關鍵詞，藉此建立韻律關鍵詞詞典，藉此表示一般語者常發生的韻律行為。

若同樣以 HTIMIT 語料庫中註冊話筒（senh）的語音資料作初步實驗，經過自動標記統計後產生的長條統計圖如圖五所示，其中可見較常發生的多為對應到韻律片語起頭或韻律片語中段的單字詞，如 state 6，1，7 與 3 等，而較少發生的則多為對應到較少發生的韻律行為，如“8-4”，“2-4”等韻律變化較激烈的雙字詞。



圖五、韻律關鍵詞詞典詞頻統計。

3.2.2. 語者—韻律關鍵詞關係矩陣

接著利用韻律關鍵詞詞典，將每一個語者的韻律文件，以長詞優先方式做斷詞（parsing）處理，然後統計每位語者出現每個韻律關鍵詞的次數，產生每位語者的關鍵詞出現頻率向量，則此向量可以表現出此語者的長程韻律行為特性。若進一步集合所有語者的關鍵詞出現頻率向量，則可建構一語者—韻律關鍵詞關係矩陣 A（見圖二），代表每位語者的韻律行為特性。

而為了減低太常出現的關鍵詞的影響(若大部份語者都出現則不具鑑別性)，與強調較少出現的韻律關鍵詞。此語者—韻律關鍵詞關係矩陣 A，將進一步使用 term frequency-inverse document frequency (TF-IDF)方法【11】作加權。

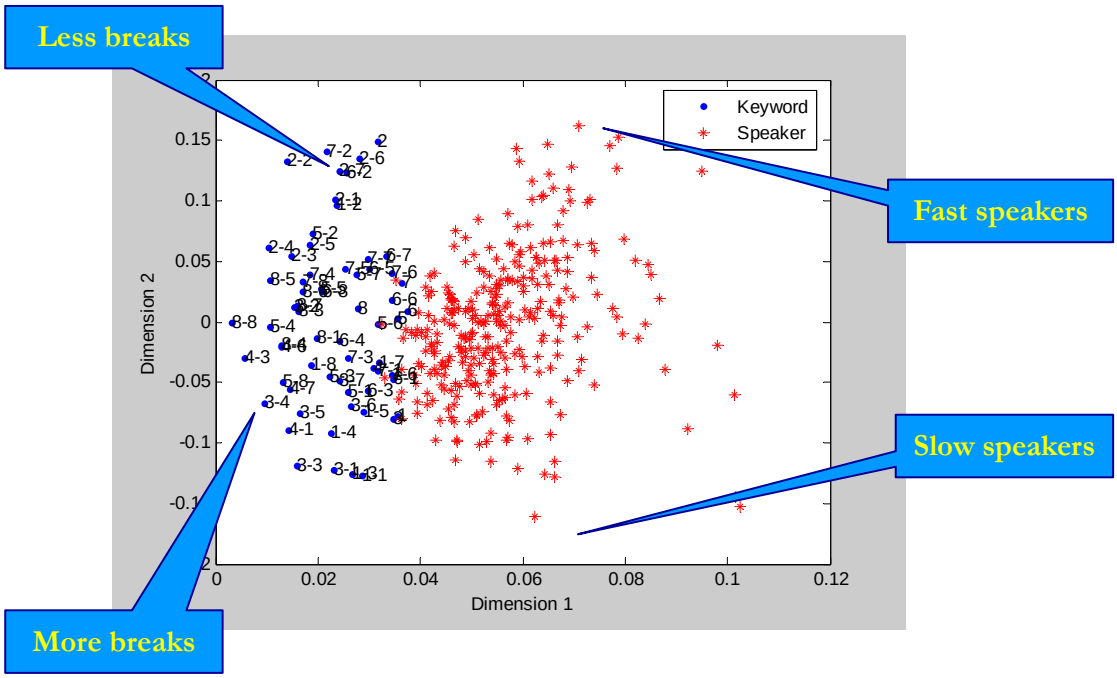
3.2.3. 特徵韻律訊息分析

經過 TF-IDF 加權過後的語者—韻律關鍵詞關係矩陣 A，實際上是一稀疏矩陣，可以利用奇異值分析將其分解，求取其特徵向量，並選取其前 K 個奇異值較大的特增值向量來近似，以產生一語者韻律特徵空間，此 SVD 分解公式表示如下：

$$A = U\Sigma V^T \simeq A_K = U_K \Sigma_K V_K^T \tag{14}$$

其中 A_K, U_K, Σ_K 和 V_K^T 分別為 A, U, Σ 和 V^T 各自矩陣的降秩(rank-reduced)矩陣。

若同樣以 HTIMIT 語料庫作初步實驗，將所有語者，投影到此低維度語者韻律特徵空間，如圖六之二維空間之範例，在此例中，state 2 是 pause duration 最短的，而 state 4 與 3 是 major 與 minor break，可見說話速度較慢的人，聚集在圖的右下角，而說話速度較快的人，多被分配到圖的右上角，由此可以看出 SVD 確實可以將不同韻律特性的語者分離開來。



圖六、語者韻律特徵空間。

3.2.4. 語者辨認分數的測量

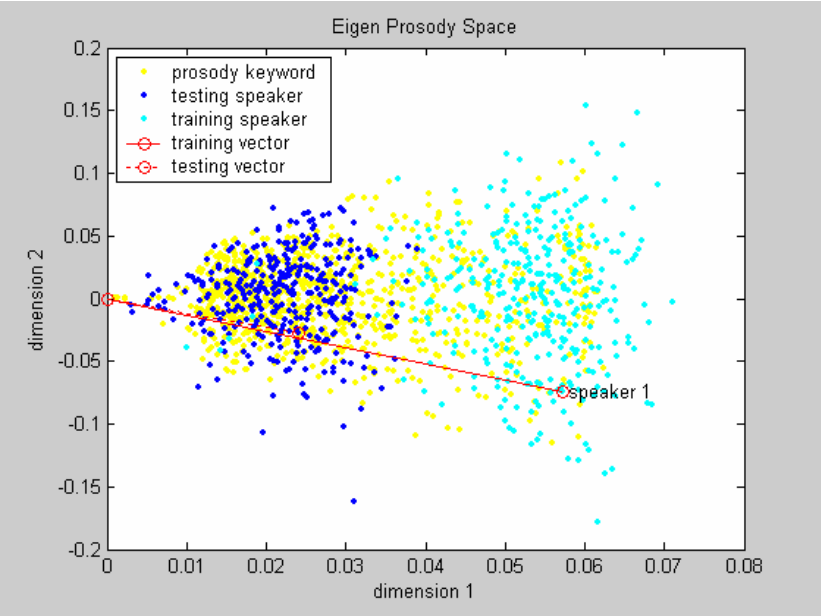
最後可以把語者辨認的問題，看成相當於文件擷取的問題。將輸入之測試語者的測試句子，同樣轉成韻律狀態索引序列，利用韻律關鍵詞詞典剖析後，當做一虛擬詢問向量y_Q，再利用式子(15)，將其投影至特徵-韻律語者空間，得到詢問向量(query vector) v_Q

$$V_Q = y_Q^T U_K \Sigma_K^{-1} \tag{15}$$

接下來則可利用詢問向量v_Q和第i個註冊語者向量$v_{K,i}$間夾角的餘弦值(cosine of angle)，來計算測試語者和第i個註冊語者之間的距離，其中夾角最小的即為韻律行為最相似的註冊語

者，如圖七所示。

經由此 EPA 方法，在作語者辨認時，只需求取每位語者投影到一低維度語者韻律特徵空間的少量座標值，即可利用韻律訊息進行語者辨認，因此所需估計的參數數目很少，可以有效利用韻律訊息，並部分解決使用韻律訊息時，常發生訓練和測試語料不足的問題。



圖七、語者在特徵韻律空間的分佈與辨認分數量測。

4. EPA，ML-AKI＋MCE 和 MAP-GMM/CMS 融合方法

利用聲學與韻律訊息具有互補性之性質，使用聲學訊息的 ML-AKI＋MCE，與使用韻律訊息的 EPA，可以進一步整合，以加強語者辨認系統的強健性。在本論文中使用如圖八的架構，以線性回歸方式，融合 ML-AKI＋MCE，EPA 與傳統 MAP-GMM/CMS 等方法的辨認分數，得到最後的語者辨認結果。

我們首先融合 ML-AKI＋MCE 與傳統 MAP-GMM/CMS 方法，主要是考量到所收集的話筒先驗知識並不可能含蓋所有話筒的特性，總是會有一些例外的未知話筒，因此將 ML-AKI＋MCE 的辨認分數和傳統以 CMS 為基礎之辨認器之分數作融合，其融合的如式子(16)所示：

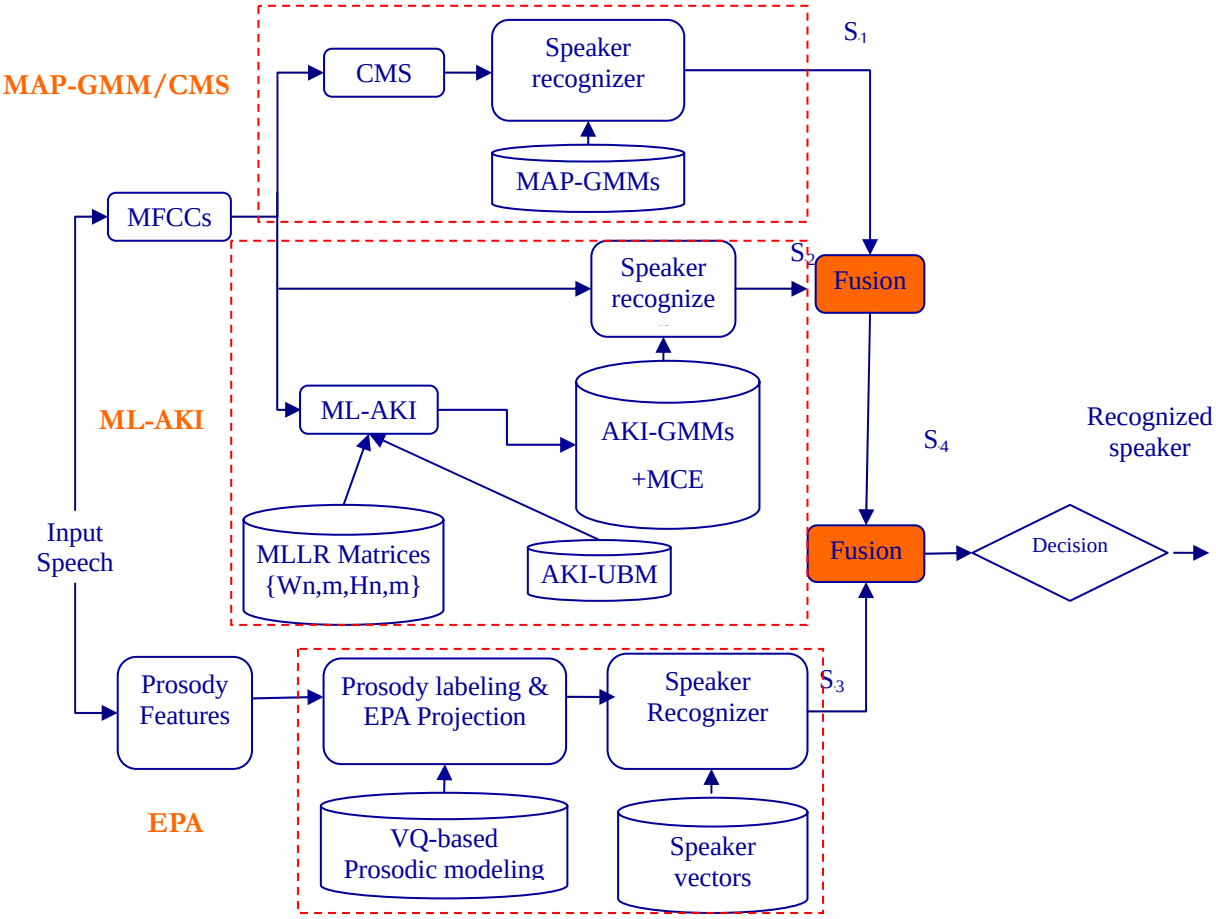
$$S_f = \lambda \cdot \frac{(s_1 - \bar{s}_1)}{\sigma_{s_1}} + (1-\lambda) \cdot \frac{(s_2 - \bar{s}_2)}{\sigma_{s_2}} \qquad (16)$$

其中 λ 為權重常數，s_1 和 s_2 為兩個系統的鑑別分數，$\bar{s}_1$，$\bar{s}_2$，σ_{s_1} 和 σ_{s_2} 分別為 s_1 和 s_2 的期望值和標準差。

然後再把融合後的分數與 EPA 辨認器的分數再做一次分數融合，融合方法使用 sigmod 函數，如式子(17)所示：

$$S_g = \alpha * \left(\frac{1}{1+\exp^{-r\left(\frac{s_3-\bar{s}_3}{std(s_3)}\right)}} \right) + (1-\alpha) * \left(\frac{1}{1+e^{-r\left(\frac{s_4-\bar{s}_4}{std(s_4)}\right)}} \right) \quad (17)$$

其中 α 為權重常數，s_f 和 s_2 為兩個系統的鑑別分數，γ 控制融合非線性程度，$\bar{s}_f$ ，$\bar{s}_2$，σ_{s_f} 和 σ_{s_2} 分別為 s_f 和 s_2 的期望值和標準差。



圖八、融合聲學層次 MAP-GMM/CMS、ML-AKI+MCE 與韻律層次 EPA 分數之架構。

5. 語者辨認實驗

為驗證本論文所提方法之效果，在以下的實驗中，採用含有十種不同話筒的 HTIMIT 語料庫，並以 leave-one-out 方式作九輪實驗，以共 90 次實驗的結果作分析討論。

5.1. HTIMIT 語料庫

HTIMIT是美國Linguistic Data Consortium (LDC) 所發行的語料庫，由Massachusetts Institute of Technology (MIT) 所設計，專門用來探討電話話筒不匹配效應對語者辨認系統的影響。HTIMIT將TIMIT語料庫經過十種不同的電話話筒重新錄製而成，其中包含約 400 位語者，每人錄製 10 個句子 (使用Sennheizer高品質麥克風，senh)，此外並將這些句子，再經過與九種不同的話筒重新錄製，因此每個人皆有十種不同話筒的語料各十句。其中九種話筒包含四種碳墨式(cb1～4)、四種電子式(el1～4)，和一個無線電話話筒 (pt1) 所組成。話筒的選擇條件為不同話質、不同transducer，…，等，其中cb3 和cb4 話筒被選擇是因為它們的聲音特性特別差。

5.2. 實驗條件

本實驗從 HTIMIT 語料庫取出 302 位語者，包含 151 位男性與 151 女性語者。特徵參數使用 38 維 MFCCs，但在求取 MFCCs 時，將 filterbank 的頻帶限制為 300~3200 Hz，以初步減輕 handset 特性的影響。音高與能量軌跡的求取則使用 snack 軟體，並從 TIMIT 中擷取正確音節切割位置。實驗方式採 leave-one-out 方式作九輪實驗，每一輪實驗，皆使用每位語者之 senh 話筒部分之語料為語者註冊語料，並輪流排除某一種話筒當未知話筒 (senh 除外)，使用其餘九種話筒當先驗知識。在訓練與測試語料長度方面則依據下列規則：(1) 在聲學層次訓練時以 senh 話筒中的所有語者的前十六秒語料訓練語者模型；測試時，每一個人使用十種話筒輪流測試，測試語料用所有語者的各種話筒的後四秒語料。(2) 而在韻律層次，訓練時以 senh 話筒中的所有語者的前七句語料訓練語者模型；測試時，則每一個人亦使用十種話筒輪流測試，測試語料用所有語者的各種話筒的後三句語料。

此外 GMM 語者模型使用 256 高斯混合數的 MAP-GMM【12】，並使用語者層次之最小錯誤鑑別式再訓練 MAP-GMM 語者模型【13，14】，VQ 韻律模型則使用 32 mixtures (先前之 8-state VQ 純為方便解說使用)，並找出 432 的韻律關鍵詞，因此韻律關鍵詞-語者矩陣的維度為 432＊302，語者韻律特徵空間則經初步實驗訂為 5 維。

5.3. 實驗結果

首先，我們使用傳統 MAP-GMM 當作基礎，並以 CMS 方式去除話筒的通道效應，其結果如表三和表四所示，MAP-GMM/CMS 方法的十種話筒的平均語者辨認率可達到 60.2%，若不計註冊話筒 (senh)，則為 58.5%。但若進一步利用語者層次之最小錯誤鑑別法則 (MCE)【13】，再次訓練語者模型，則十種話筒的平均語者辨認率可提升到 61.9%，若不計註冊話筒 (senh)，則為 60.3%。

接下來以本論文所提出的 ML-AKI 方法，進行 leave-one-out 實驗，則九輪實驗 (共 90 次實驗) 的平均語者辨認率可提升到 73.7% (如表三所示)，若將九輪實驗中的未知話筒實驗部分獨立出來，則九次未知話筒實驗的平均語者辨認率可提升到 65.0% (如表四所示)。顯示 ML-AKI 不管對已知或未知話筒的不匹配效應皆有不錯的補償效果。

而若將 MCE 與 ML-AKI 疊加起來，使用 MCE 調適後的 GMMs 作為 ML-AKI 的聲學模型 (ML-AKI+MCE)，並和 MAP-GMM/CMS 作辨認分數融合，進行 leave-one-out 實驗，並以如圖九所示的方式找出最佳融合權重，則九輪實驗的平均語者辨認率可再提升到 74.9% (如表三所示)，

而若將九輪實驗中的未知話筒實驗部分獨立出來，則九次未知話筒實驗的平均語者辨認率可提升到 69.7%(如表四所示)。顯示辨認分數融合方式可以截長補短，以互補方式提升辨認率，且 ML-AKI＋MCE 與 MAP-GMM/CMS 的權重比重約為九比一，權重偏重 ML-AKI＋MCE。

最後我們再疊加上 EPA 方法，並作辨認分數融合，一樣進行 leave-one-out 實驗，並以如圖十所示的方法找出的最佳融合權重，則九輪實驗的平均語者辨認率可提升到 79.3%（如表三所示），若將九輪實驗中的未知話筒實驗部分獨立出來，則九次未知話筒實驗的平均語者辨認率可提升到 74.6%（如表四所示）。顯示韻律訊息與聲學訊息確實具有非常不錯的互補效果，且聲學與韻律訊息的權重比重約為七比三。

表三：在 HTIMIT 語料庫上使用 leave-one-out 實驗方式，嘗試各方法所得到之不同話筒的語者辨認率，與所有話筒的平均語者辨認率。

方法	senh	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Average
(1) MAP-GMM/CMS	75.1	70.9	73.8	30.5	35.8	73.8	63.2	58.9	65.2	54.3	60.2
(2) MCE	75.8	70.2	75.5	32.2	38.7	75.2	64.6	62.3	67.2	57.0	61.9
(3) ML-AKI	84.1	78.5	82.2	50.8	63.3	83.7	76.1	70.5	78.4	69.4	73.7
(1)+(2)+(3)	86.1	81.7	84.3	49.9	62.5	85.4	76.9	74.8	77.9	70.7	74.9
(1)+(2)+(3)+EPA	89.1	83.0	87.1	59.7	67.1	88.5	80.1	79.6	82.4	76.6	**79.3**

表四：在 HTIMIT 語料庫上使用 leave-one-out 實驗方式，只觀察在每一輪實驗中未知話筒所得到之語者辨認率，與所有未知話筒的平均語者辨認率。

方法	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Average
(1) MAP-GMM/CMS	70.9	73.8	30.5	35.8	73.8	63.2	58.9	65.2	54.3	58.5
(2) MCE	70.2	75.5	32.2	38.7	75.2	64.6	62.3	67.2	57.0	60.3
(3) ML-AKI	77.5	79.1	32.5	50.7	80.5	59.9	71.2	73.5	60.3	65.0
(1)+(2)+(3)	80.4	82.8	38.1	57.0	85.4	67.2	74.8	76.5	65.2	69.7
(1)+(2)+(3)+EPA	83.4	84.3	45.7	62.6	87.1	74.8	79.5	80.8	72.8	**74.6**

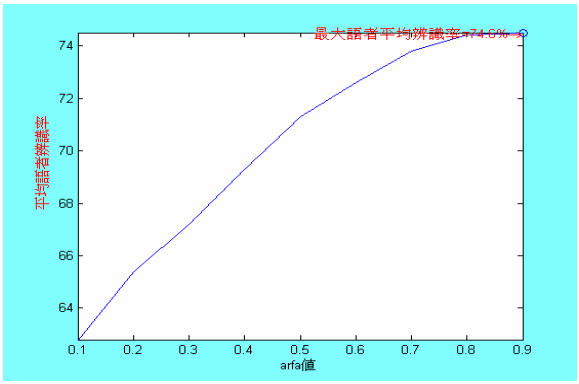
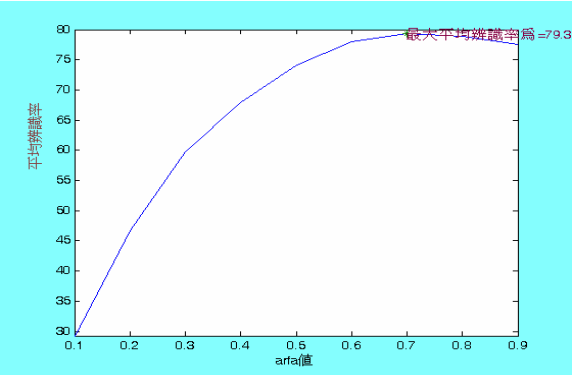


圖 九、融合 ML-AKI＋MCE 和 MAP-GMM/CMS 分數時權重 λ 值與辨認率之關係。

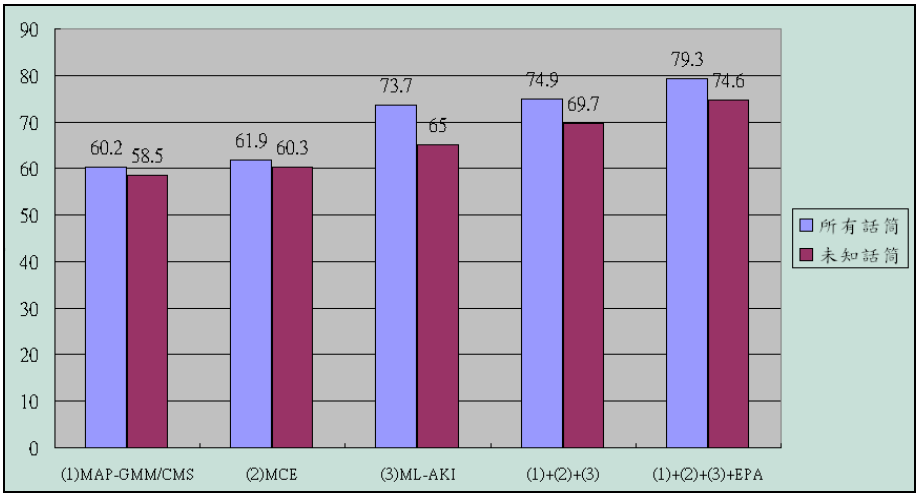


圖十、融合聲學訊息與韻律訊息時 α 值與辨認率之關係。

5.4. 實驗總結與討論

所有實驗結果的總結如圖十一所示，利用傳統 CMS 方法使用 MAP-GMM 語者模型的辨認結果為基礎，疊加上語者層次最小錯誤鑑別法則再訓練語者模型，再疊加上本論文所提出的 ML-AKI 和 EPA 方法，在含有未知話筒情形下，平均語者辨認率由 60.2%提升至 79.3%，若只觀察未知話筒，則平均語者辨認率亦可由 58.5%提升至 74.6%。因此無論對已知話筒或未知話筒而言，都可以得到不錯的提升，足以證明我們所提出的 ML-AKI＋MCE 和 EPA 對於話筒不匹配的問題，的確得到相當程度的改善。

另外由圖九和圖十所示，可看出當 MAP-GMM/CMS 和 ML-AKI+MCE 融合時，最佳融合的係數為 0.9，可看出大部份是依靠 ML-AKI+MCE 為基礎之辨認器之分數。而和 EPA 融合後，所求最佳融合的係數為 0.7，雖然大部份還是依賴 ML-AKI＋MCE 的分數，但融合之後，平均語者辨識率提升到 79.3%，得到很不錯的提升，所以由此得知 EPA 和 ML-AKI＋MCE 是非常具有互補性的。



圖十一、融合 MAP-GMM/CMS，MCE，ML-AKI 與 EPA 等方法，在 HTIMIT 語料庫上使用 leave-one-out 實驗方式所得到之平均語者辨認率，與只觀察每一輪實驗中之未知話筒所得之平均語者辨認率。

6. 結論

本論文嘗試融合聲學與韻律層次訊息，以建立強健式語者辨認系統，包括融合聲學層次的最小錯誤鑑別式法則訓練之 MAP-GMM 語者模型，ML-AKI 和韻律訊息層次的 EPA 分析。由 HTIMIT 實驗結果來看，平均語者辨認率可從傳統 MAP-GMM/CMS 的 60.2%，提升到 79.3%，而若只單看未知話筒部分，平均語者辨認率亦可從 58.3%，提升到 74.6%。因此聲學與韻律層次訊息的融合，的確可對於話筒不匹配問題得到一定程度的解決，尤其在未知話筒方面，也得到不錯的改善。此外若從所使用的語料長度來看，聲學層次系統都只用了 16 秒與 4 秒的訓練與辨認語料，韻律層次系統都只用了七句與三句的訓練與辨認語料，因此確實可善用有限的訓練與辨認語料。綜合以上說明可驗證本論文所題所提方法的有效性。

7. 參考文獻

【1】 http://www.clsp.jhu.edu/ws2002/groups/supersid/SuperSID_Closing_Talk_files/frame.htm

【2】 S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254-272, Apr. 1981.

【3】 M. G. Rahim and B. H. Juang: 'Signal bias removal by maximum likelihood est-imation for robust telephone speech recognition', IEEE Trans. On Speech andAudio Processing, vol. 4, no. 1, pp. 19-30, Jan 1996.

【4】 D. A. Reynolds: 'HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects', in Proc. ICASSP'97, vol. II, pp. 1535-1538, 1997.

【5】 D. A. Reynolds et. Al., "The SuperSID project; exploiting highlevel information for high-accuracy speaker recognition, "Proc. ICASSP'03, vol, IV, pp. 784-787, 2003.

【6】 D. Reyolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaus-sian Mixture Models," Digital Signal Processing, vol. 10, pp. 19-41, January 2000.

【7】 S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, "Indexing by Latent Semantic Analysis," Journal of the Society for Information Science, vol. 41(6), pp. 391-407. 1990.

【8】 Fu Shu-qun, Cao Bing-yuan, Ma Jin-wen" Research on correct convergence of the EM algorithm for Gaussian mixtures," Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference. vol. 5, pp. 18-22 Nov. 2002.

【9】 D. A. Reynolds: 'HTIMIT and LLHDB: Speech corpora for the study of handset

transducer effects', in Proc. ICASSP'97, vol. II, pp. 1535-1538, 1997.

【10】 http://www.speech.kth.se/snack/

【11】 Li-Ping Jing, Hou-Kuan Huang, Hong-Bo Shi" Improved feature selection approach TFIDF in text mining," Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on Vol. 2, 4-5 Nov. 2002 Page(s):944 - 946 vol.2

【12】 Gauvain, J. L. and Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, " IEEE Trans. Speech Audio Process. 2(1994), 291-298.

【13】 W. Chou, B.H. Juang and C.H Lee, "Segmental GPD Training of HMM based Speech Recognizer," In proceedings of ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, page(s): 473 -476, 1992.

【14】 Biing-Hwang Juang, Wu Chou, and Chin-Hui Lee, "Minimum Classification Error Rate Methods for Speech Recognition," IEEE Trans. on Speech and Audio Processing. Vol. 5, NO. 3, May 1997.

以高斯混合模型表徵器與語言模型為基礎之語言辨認

Language Identification based on Gaussian Mixture Model Tokenizer and Language Model

張智傑、王小川
Zhi-Jie Chang and Hsiao-Chuan Wang

國立清華大學電機工程學系
Department of Electrical Engineering, National Tsing Hua University
E-mail : piscesboy@micro.ee.nthu.edu.tw hcwang@ee.nthu.edu.tw

摘要

本論文探討不需要標注資料的自動化語言辨認方法，基本觀念是建立高斯混合模型之表徵器，以表徵器輸出建立語言模型，加上切割處理與後端處理，提升語音資料的語言辨認正確率。所建議的系統架構，分別是串聯高斯混合模型表徵器和語言模型的“高斯混合模型表徵器-語言模型法”，以及將語言模型融合在表徵器裡面的“連結聲學-語言模型法”兩種型式。由實驗結果觀察，加入切割處理的幫助，的確能夠提升系統的辨認正確率。

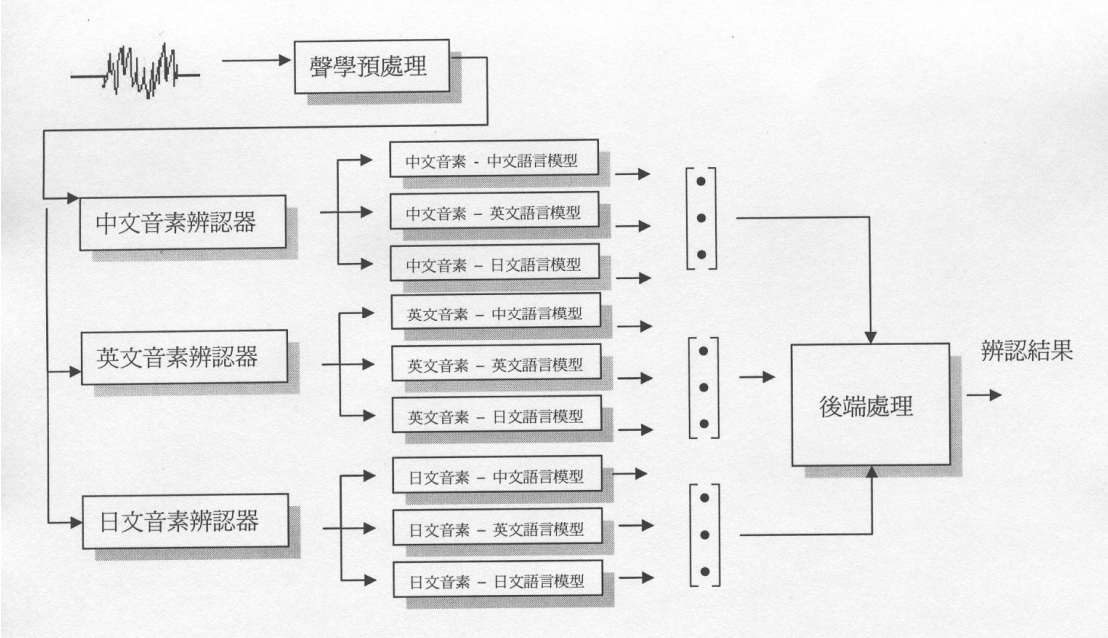
關鍵詞：語言辨認、高斯混合模型、表徵器、語言模型

一、緒論

近代語言辨認的方式，主要是對訓練語音資料，轉換成類音素（phone-like）序列，以類音素序列建立 N-連文模型作為語言模型。在做語言辨認時，計算測試語音之類音素序列與語言模型之間的相似度，經過後端處理做出語言辨認的判斷。所建議的系統有連結語言模型的音素辨認法（PRLM, Phone Recognition Language Model)[1][2]、連結語言模型的平行音素辨認法(PPRLM, Parallel-language PRLM)[1][2]、高斯混合模型表徵器-語言模型法（GMM-tokenizer-language model) [2][3] 、以及連結聲學-語言模型法（Joint-Acoustic Language Model) [4] [5] 等方式。

連結語言模型的音素辨認法[1]是將輸入語料經過預先訓練好的音素辨認器（phone recognizer），得出輸入語料的音素序列（phone sequence），再由音素序列統計產生語言模型(language model)。在辨認過程中，則是計算測試語音的音素序列與 N-連文法（N-gram)語言模型的相似度（likelihood），對應相似度最高的語言模型，就是辨認結果。圖一是以中英日三個語言的辨認為例，展示語言辨認系統之示意圖。輸入的測試語音，分別經由中英日三個語言的音素辨認器，產生三個不同的音素序列，將這三個不同的音素序列分別輸入到三個語言所建立的語

言模型，得出九個相似度值，後端處理器對這九個相似度值做運算，產生最後的辨認結果。



圖一、 連結語言模型的音素辨認法

表徵器-語言模型法的系統，需要有標註好的訓練語料做為音素辨認器訓練之用，要人工的介入才能完成系統建構；因此有研究者提出基本概念相似，但不需人工幫助的高斯混合模型表徵器-語言模型系統。其作法是將高斯混合模型的各個高斯機率密度函式(Gaussian probability density function)視為一個量化單位，給予模型中的每個高斯分布固定的表徵（token）值，將一個音框在各個高斯分布的機率值計算出來後，選擇機率最大的高斯分布作為表徵，視為此音框的代表值。對於輸入的測試語料，以高斯混合模型的表徵值序列（token sequence）取代連結語言模型的音素辨認法的音素序列（phone sequence），再做語言模型的處理。高斯混合模型表徵器-語言模型法在訓練語料充足時，實驗結果顯示有不錯的表現。

連結聲學-語言模型法是將語言模型合併到次字元的辨認器中，在決定音框的次字元時，加入語言模型的考量，以語言模型的機率值做為次字元間的轉移機率。使用連結聲學-語言模型法，可以降低次字元模型的數量 [4]，在次字元模型數更少的情況下，達到接近平行音素辨認法的效果。

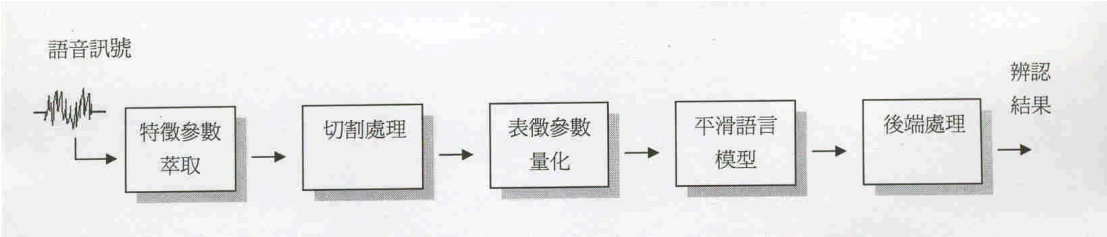
本篇論文的內容如下：第二節介紹本論文所使用的系統架構、流程，以及原理，第三節介紹所使用的程式工具，以及實驗的語料庫，第四節將實驗結果以圖表的方式呈現，並對其所顯示的現象加以討論。

二、 語言辨認之模型與方法

2.1 基本流程

語言辨認的基本流程如圖二所示，從語音訊號萃取出特徵參數後，經過切割成為類音素(phone-like)的聲音單位，經過表徵參數量化器（tokenizer)，將每個聲音單位轉換成一個音素代碼。對於一段輸入語音，即轉換成一序列的音素代碼，用以計算該段輸入語音與每一個語言模型

的對數相似度 (log-likelihood)。最後由後端處理器，對多個表徵器-語言模型(tokenizer –language model) 的相似度做處理後，得出最後的辨認結果。



圖二、 語言辨認的基本流程

2.2 位移差分化倒頻譜參數 (Shifted Delta Cepstrum)

倒頻譜參數由單一音框計算而得，並不包含發音腔道模型隨時間的變化特性。差分化 (delta) 便是將時間特性考慮進來，透過計算連續幾個相連音框的參數差值，表現出特徵參數的變動情況。本篇論文使用 HTK 所採用的回歸方程式 (regression formula) 來處理參數的差分化 ：

$$d_t = \frac{\sum_{k=1}^{K} k \cdot (c_{t+k} - c_{t-k})}{2\sum_{k=1}^{K} k^2} \qquad (1)$$

d_t 為差分化的特徵參數，k 為周圍音框和目前處理音框的距離，c_t 為音框 t 的特徵參數向量，K 為差分化音框 (delta window) 的大小。

位移差分化倒頻譜參數是將距離相同的多個音框的差分化倒頻譜參數結合起來，成為一個維度更大的向量，當做新的特徵參數來使用。一般使用四個關鍵值 (N, d, p, k) 來加以描述；

N : 單一音框計算出的倒頻譜參數的維度。

d : 差分化音框的大小

p : 串接差分向量的音框距離

k: 串接差分向量的個數

本論文中，關鍵值選用 (N, d, p, k) = (10, 1, 3, 3) ; 表示單一音框計算出 10 維倒頻譜參數，以其前後相連的兩音框計算出 10 維的差分化倒頻譜參數後，將間隔 3 音框的 3 組差分化倒頻譜參數串接成維度 30 的新向量，做為該音框的特徵參數向量。

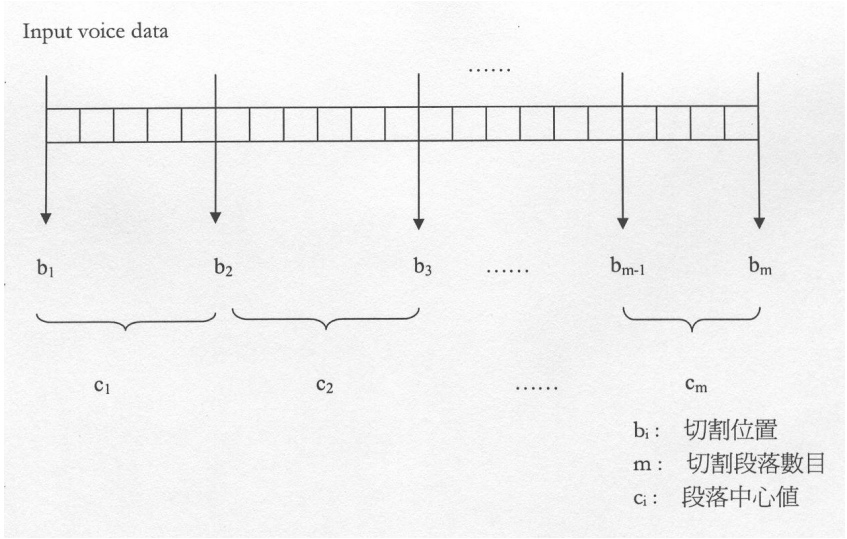
2.3 切割處理 (Segmentation)

將輸入的語音資料特徵參數序列，分割為預先指定的音段數目，稱做切割處理。音段的數目可以事先指定，也可以採用設定限制(constrained) 的方式讓系統自行決定。加入切割處理的目的在於取得類音素 (phone-like) 的單位作為語音段 (speech segment)。根據統計，各國語言音素出現的平均頻率為每秒 10 個音素。因此在本論文中，指定每秒切割出 10 個音段的方式。分割位置的決定，係以最小音段內差異(intra-segment distortion)的準則，讓每個分割出來的音段能夠有最大的聲學一致性 (acoustic homogeneity)。如圖三所示，輸入的語音特徵參數序列為 $X_1^T = (X_1, X_2, ...X_T)$ ， 若得出的分割點為 $B_0^m = (b_0, b_1, ..., b_m)$ ，計算每個音段的中心值 (centroid) $C_1^m = (c_0, c_1, ..., c_m)$ ，以及音段內各音框特徵參數和其中心值的距離，總和起來就

是該語音特徵參數序列的距離總和，

$$D(m,T)=\sum_{i=1}^{m}\sum_{n=b_{i-1}+1}^{b_i} d(X_n,\mu_i) \qquad (2)$$

設定分割點的目標，就是要使得距離總和為最小。



圖三、 特徵參數序列的分割

根據特徵參數選用的不同，各音段中心點以及最後距離總合的計算方式也應該跟著改變[6]。選用梅爾刻度倒頻譜參數時，中心點即為各音段中所有音框特徵參數的算數平均值，距離的計算則是採用歐氏幾何距離。分割位置採用動態程式規劃（Dynamic Programming）的方式，

$$D(i,b_i)=\min_{b_{i-1}}\{D(i-1,b_{i-1})+\Delta(b_{i-1},b_i)\} \qquad (3)$$

將所有可能的分割組合都加以嘗試。

2.4 語言模型（Language Model）

將量化單位給予一個類音素的代碼，於是一段語音就被轉換成一序列的類音素代碼。從各個語言的訓練語料，可以計算出每個類音素代碼量出現的機率。在做語言辨認測試時，即依據訓練語料統計出的機率，計算出測試語料的對數相似度。若一次選取連續 N 個語音段當作統計標準，則稱作 N-連文語言模型（N-gram language model）。以二連文法為例，根據各個類音素代碼的出現次數計算出狀態機率（conditional probability）：

$$P(w_t \mid w_{t-1})=\frac{C(w_{t-1},w_t)}{C(w_{t-1})} \qquad (4)$$

其中，w_{t-1},w_t 表示時間上相連的兩個類音素代碼，$C(w_{t-1},w_t)$ 為兩者共同出現的次數，$C(w_{t-1})$ 則為 w_{t-1} 出現的總次數。測試語料經由前端處理器轉換成類音素代碼序列

$W=\{w_0,w_1,...,w_T\}$ 後，輸入到各語言的 N-連文法語言模型 λ_l^{NG}，計算出其對數相似度

$$L(W \mid \lambda_l^{NG}) = \sum_{t=1}^{T} \log P(w_t \mid w_{t-1}, \ldots, w_{t-(N-1)}, \lambda_l^{NG}) \qquad (5)$$

做為語言模型的輸出結果。

如果其中某一個類音素代碼在訓練語料中沒有出現，可能會發生 $P(w_t \mid w_{t-1}) = 0$ 的狀況，使得測試時出現 log 0 這樣無意義的成份。解決的方法是以平滑化的 *N* 連文語言模型（smoothed N-gram language model）來取代原本的 *N* 連文語言模型。以二連文法為例，平滑化的二連模型為二連及一連模型的線性組合

$$\tilde{P}(w_t \mid w_{t-1}) = \alpha_2 P(w_t \mid w_{t-1}) + \alpha_1 P(w_t) + \alpha_0 \qquad (6)$$

其中，$P(w_t \mid w_{t-1})$ 為二連文法語言模型，$P(w_t)$ 為一連文法語言模型，α_2, α_1 為其對應的權重值，α_0 則為防止兩者皆為 0 的偏差值。α 的大小是由評估最大值演算法迭代求得。根據前人研究的結果 [1]，$0.3 < \alpha_1, \alpha_2 < 0.7$ 時系統會有最佳的表現。

在本論文中，選取的語言模型為平滑化的二連語言模型，對應的參數則選取 $\alpha_2 = 0.666, \alpha_1 = 0.333, \alpha_0 = 0.001$。另外為了避免 $P(w_t \mid w_{t-1})$ 分母出現 0 的情況，每個語音段量化單位出現次數至少為 1。

2.5 常用的系統組合與方法

A. 高斯混合模型表徵法 (GMM Tokenization)

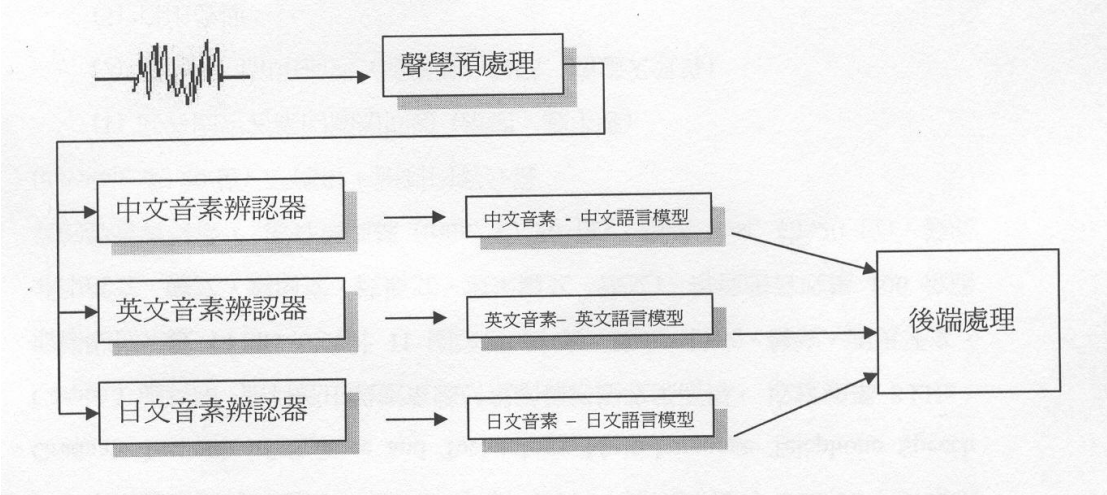
高斯混合模型表徵法是在高斯混合模型中每個高斯分布有一個固定的表徵（token）值，計算出單一音框在高斯混合模型中各個高斯分布發生的機率值，選擇機率最大的高斯分布的表徵做為此音框的代表值。其概念和向量量化（Vector Quantization, VQ）近似，不同的是，在向量量化中，我們取用和輸入音框距離最接近的中心值（centroid）的代號當做輸入資料的量化值。在高斯混合模型表徵法中，則是採用音框機率值成份最大的高斯模型的代號當作量化結果。

經過高斯混合模型表徵器 **(GMM Tokenizer)** 的處理後，輸入的特徵參數序列轉變成表徵序列 **(token sequence)**，於是就可以進行下個階段的語言模型處理。語言模型的大小決定於前端高斯混合模型表徵器混合數 **(mixture)**。混合數越大，表示原資料在量化時的單位越細，語言模型在統計時就必須記錄下越多種可能發生的語音段排列。以二連文法為例，混合數為 N 時，對應的二連文法語言模型大小為 N^2。

當訓練語料不夠多時，會使得許多成份出現機率太小，在測試端使用時就會讓相似度的變動範圍變小而難以做比較。在使用高斯混合模型表徵法時，混合數和語言模型的大小必須跟著訓練語料的大小、分布做調整。

B. 平行模式法 (Parallel Model)

如圖四所示，平行模式是每個音素辨認器只連接和其語言相同的語言模型，目的在於彌補音素辨認器未考慮到的時間性質及順序問題。如果是採用高斯混合模型表徵法，則是將前端的音素辨認器改為高斯混合模型表徵器。



圖四、 平行模式法

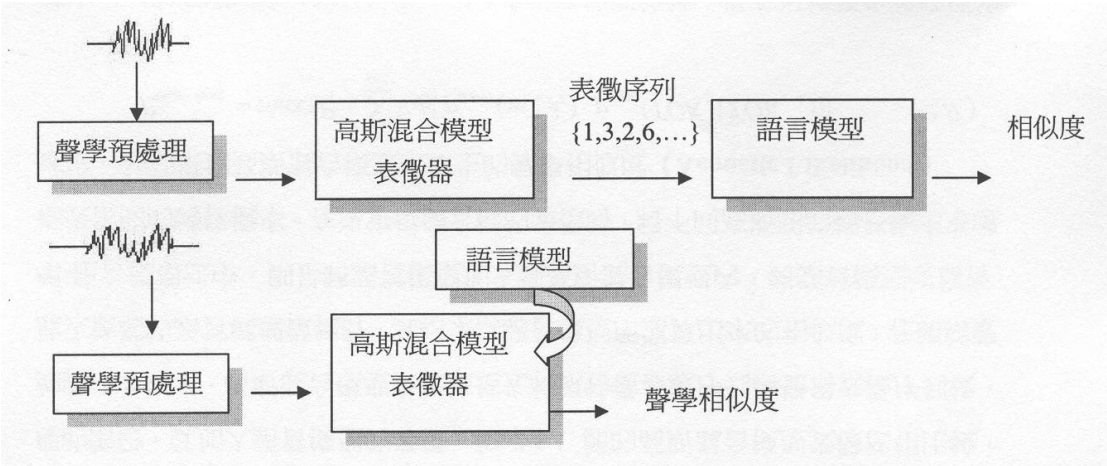
C. 連結聲學-語言模型法 (Joint-Acoustic-Language Model)

連結聲學-語言模型法的概念是，在做表徵序列 **(tokenization sequence)** 處理的階段，就加入語言模型的考量。高斯混合模型表徵法是先將特徵參數序列轉變為表徵序列後，作為語言模型輸入，得出對於該語言模型的相似度。在連結聲學-語言模型法中，則是將語言模型加入到高斯混合模型內，將語言模型視為音框表徵間的轉移機率。在決定每個音框的表徵時，將轉移機率考慮進去。其結果為高斯混合模型所產生的聲學相似度 **(Acoustic Likelihood)**

$$P_{token-lang} = \max_{\lambda}\{P_1 + \sum_{t=2}^{T} \log[\, p_{token}(w_t \mid \lambda_t) \cdot p_{lang}(TOK_t \mid TOK_{t-1})]\} \qquad (7)$$

其中，T 為音框總數，$\lambda = (\lambda_1, \lambda_2, ..., \lambda_T)$ 表示最有可能產生該音框的高斯成份序列，

$p_{token}(w_t \mid \lambda_t)$ 為音框 w_t 在 λ_t 的機率值， $p_{lang}(TOK_t \mid TOK_{t-1})$ 為語言模型，表示第 t-1 個音框表徵連接第 t 個音框表徵的發生機率，P_1 為第一個音框的最大高斯成份機率。



圖五、連結聲學-語言模型法

2.6 後端處理 (Back-end Process)

A. 專家投票法 (Voting)

專家投票法就是將所有要辨認的語言視為候選者, 每個語言模型的輸出視對候選者投下的選票。統計每個語言模型的輸出，以得票數最高的語言模型當做辨識結果。

B. 幾何平均法

使用平行音素辨認器時，某一個語言音素辨認器的輸出，只經過該語言音素所訓練的語言辨認器，產生一序列的相似度值。將各個語言所產生的相似度值做相乘後開次方，得出的結果即為測試語料和該國語言的相似度。

C. 算術平均法

和幾何平均法相同，得出一序列的相似度值，而後端處理改為對相似度值做相加後平均。由於語言模型的輸出結果為機率值乘積，取對數的結果

$$L_k = \log(\prod_{i=1}^{T} p_k(x_i)) \qquad (8)$$

將不同音素辨認器中同語言模型的部份相加，

$$L_1 + L_2 + L_3 = \log(\prod_{i=1}^{T} p_1(x_i)\, p_2(x_i)\, p_3(x_i)) \qquad (9)$$

在計算輸入音素的相似度時, 將不同的音素-語言模型組合視為彼此獨立的事件。

三、 實驗語料庫及使用工具

3.1 OGI-TS 語料庫

本篇論文使用的語料庫是 1992 年完成錄製的語料庫 OGI-TS (Oregon Graduate Institute of Science and Technology Multi-language Telephone Speech Corpus)[7]。取樣頻率為 8 kHz，取樣點位元數為 14 bits，共計有 11 國語言(中文、英文、日文、德文、西班牙文、北印度文、韓文、越南文、波斯文、坦米爾文、法文)。每種語言收集大約 100 位語者的聲音資料，這些資料分為訓練 (train, 約 50 位)、測試 (test, 約 20 位)、發展 (develop, 約 20 位) 三部份。語料內容包含

(1) 回答固定答案的簡短問題 (星期、數字等)

(2) 對問題之簡單描述 (最喜歡的餐點、描述天氣等)

(3) 自由發揮

3.2 語料庫之修改

在後半段的實驗中，我們從發展語料 (developing data) 取出部份檔案，將各語言的訓練語料填補至 85 分鐘左右的長度，如表一所示。

訓練語料只選取 45 秒的長句，每種語言 20 筆資料，測試語料長句數未滿 20 者，從發展語料填補。在 11 國語言辨認涉及偏差值 (bias) 的實驗中， 使用訓練語料內約 50 筆 45 秒的長句來做偏差值的計算。

表一、填補後的訓練語料

語言	訓練語料大小 (Byte)	語者數
中文	80,301,372	62
英文	81,481,458	56
日文	81,089,452	54

德文	80,817,724	52
西班牙文	80,724,618	51
印度文	81,031,972	107
韓文	80,559,128	59
越南文	80,481,152	61
波斯文	80,937,144	57
坦米爾文	81,040,500	61
法文	80,479,746	52

3.3 使用工具

特徵參數之萃取，採用劍橋大學提供的 Hidden Markov Model Toolkit (HTK)，版本為 3.2.1。輸入語料的取樣頻率是 8 kHz，經過 $(1- 0.97Z^{-1})$ 的高通濾波器做預處理。音框大小為 32 ms，音框間距為 16 ms，每個音框乘上漢明窗 (Hamming Window)。倒頻譜參數的臨界頻帶 (critical band) 數設定為 20, 並作倒頻譜參數均值刪除法 (Cepstrum Mean Subtraction) 處理。

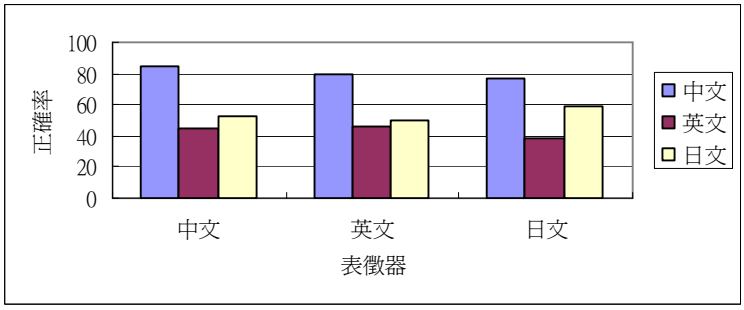
採用程式是建立在 C 語言上的 mat 2D 函式庫，版本為 1.8.1，作者為 Mike Schuster。該函式庫主要功能在處理向量及矩陣的基本運算、分類處理、基本統計模型建立、以及傅立葉轉換等語音處理經常使用到的演算。

四、 實驗結果及討論

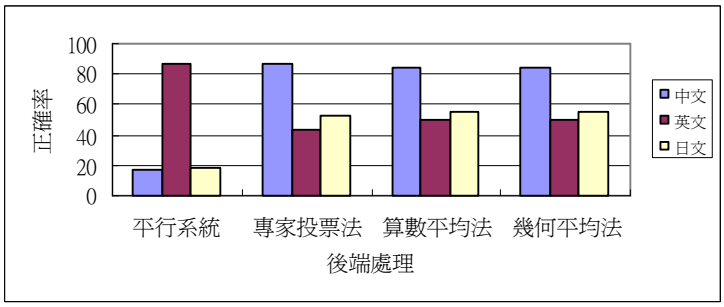
實驗的安排如下; 4.1 節與 4.2 節針對 OGI-TS 的三國語言 (中、英、日) 原本設定的訓練及測試語料做模擬實驗。對兩種不同的特徵參數 (30 階位移差分化倒頻譜參數、38 階梅爾刻度式倒頻譜參數)，作高斯混合模型表徵器-語言模型 (GMM-tokenizer language model)的語言辨認。在前端採用了單一表徵器及多表徵器兩種形式，也試驗了多種後端處理 (專家投票法、幾何平均法、算數平均法) 的方法。 4.3 節以連結聲學語言模型做為辨認系統，4.4 節則是將語料庫重新整理分類後，讓訓練語料大小相似，測試語料統一為 20 句 45 秒的長句。

4.1 中英日三國語言辨認 - 30 階位移式差分化倒頻譜參數

本節使用 30 階位移式差分化倒頻譜參數做為特徵參數，使用在單一表徵器及多表徵器兩種系統上。



(a) 單一表徵器



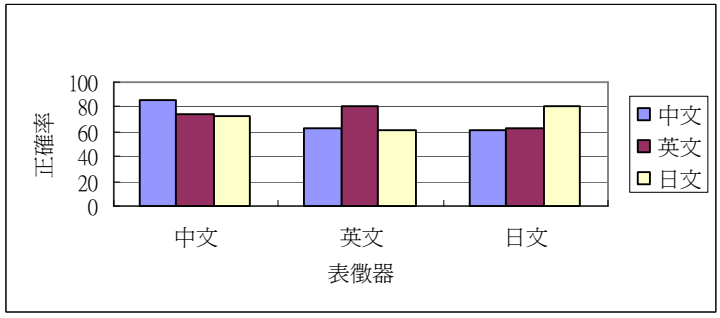
(b) 多表徵器

圖六、 使用 30 階位移式差分化倒頻譜參數之辨認結果

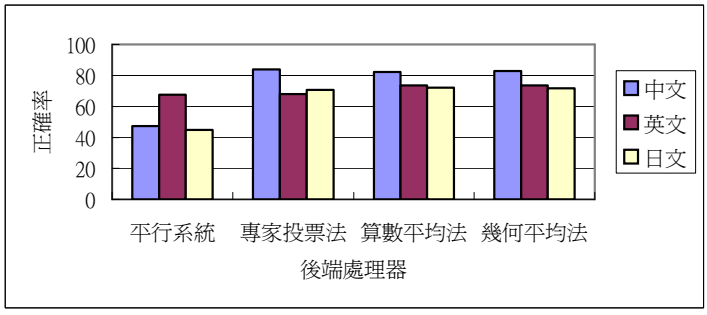
由圖六中發現， 使用位移式差分化倒頻譜參數時， 用單一表徵器及多表徵器的效果都不太理想，正確率大多不到五成。可能造成的原因是，位移式差分化倒頻譜參數將更多隨時間變化的因素加入參數萃取的步驟內，因此和語言模型相結合後，訓練語料必須要夠充足才能提供變化特性。參考文獻 [8] 使用的是比 OGI-TS 語料庫 (1.5 hr) 大 6 倍的 CallFriend 語料庫 (10 hr)，所以得到較高的辨認率。

4.2 中英日三國語言辨認 – 38 階梅爾刻度式倒頻譜參數

圖七顯示，使用梅爾刻度式倒頻譜參數，用單一表徵器時相同語言的語言模型辨認率較高，各種後端處理器的表現則相差不大，效能最好的是算術平均法，平均正確率為 75.92 % 。



(a) 單一表徵器



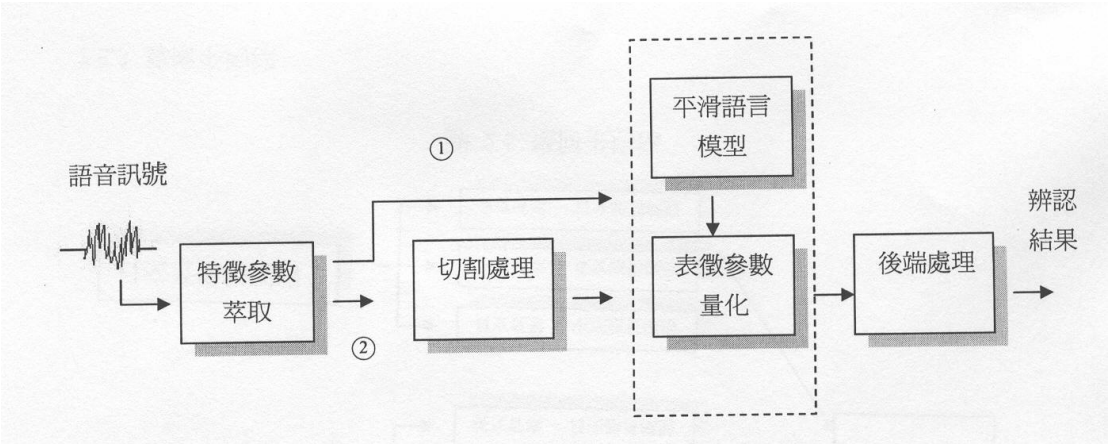
(b) 多表徵器

圖七、 使用 38 階梅爾刻度式倒頻譜參數之辨認結果

將圖六與圖七的實驗結果加以比較，38 階梅爾刻度式倒頻譜參數在後端處理為算術平均時，表現最好也最平均，平均正確率為 75.92 % 。因此我們選用該參數及算術平均之後端處理，應用在後面的實驗中。

4.3 中英日三國語言辨認 – 連結聲學語言模型

本實驗使用連結聲學語言模型做為辨認方法，由於使用這個方法可能會有偏差值的現象產生，即某些語言的相似度值會偏高，使得辨認結果偏向該國語言。使用的高斯混合模型之混合數為 32、64、128 和 256，參數採用的是 38 階梅爾刻度式倒頻譜參數。系統如圖八所示，實驗 A 未使用切割法，直接將特徵參數序列傳進表徵器內，實驗 B 則加入切割法，將特徵參數序列切割為事先指定的每秒 10 個單位，再傳入表徵器，表徵器的訓練語料也同樣經過切割處理。

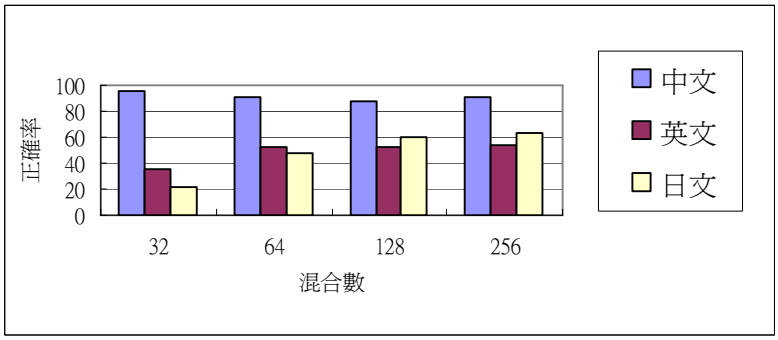


(1) 實驗 A　　(2) 實驗 B

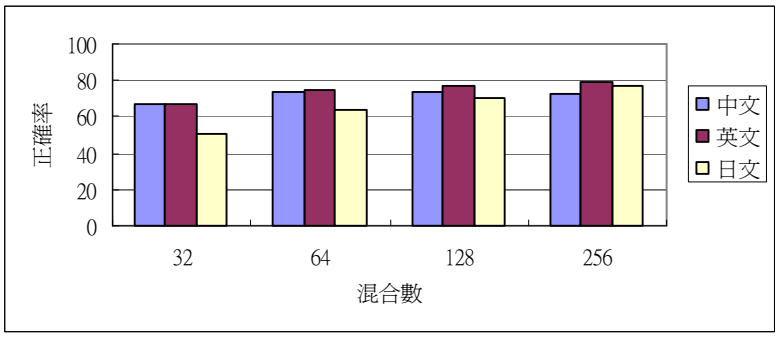
圖八、 切割處理與未切割處理之連結聲學語言模型法

(1) 實驗 A

圖九為針對測試語料做開放測試 (open test) 的實驗結果，以未經偏差值刪除的相似度做為辨認依據。



(a) 含偏差值



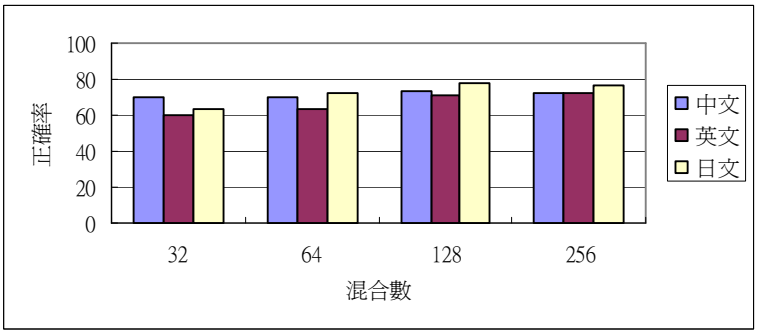
(b) 去除偏差值

圖九、 連結聲學語言模型之實驗結果(未使用切割法)

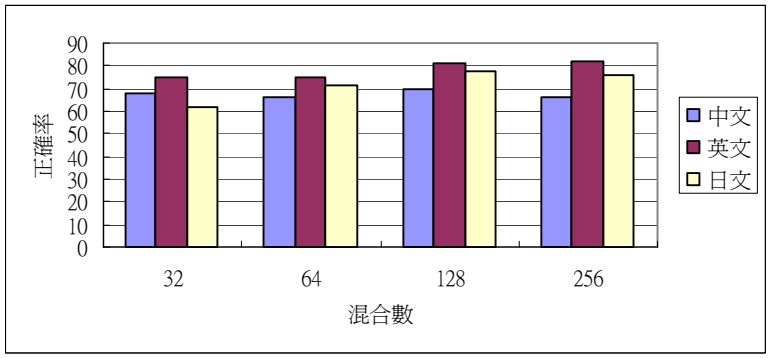
由圖中可看出，未除去偏差值時，中文語料的相似度有偏高的狀況，使得其它兩國語言的辨認效能很難提升，即使提高表徵器的混合數，也沒有顯著的效果。當除去偏差值的影響後，隨著表徵器混合數的增加，三國語言的辨認率都會提升，且會逐漸拉近，混合數 256 時有 76.27 % 的平均辨認率，和圖六的實驗結果 75.92 % 相比上升了一點。

(2) 實驗 B

本實驗對於做切割處理 **(segmentation)** 後的語料做語音辨認。



(a) 含偏差值



(b) 去除偏差值

圖十、 連結聲學語言模型之實驗結果(使用切割法)

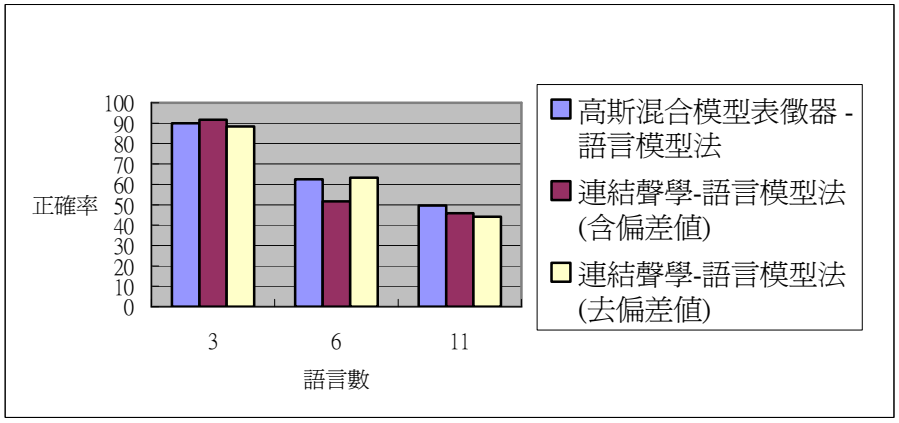
圖十顯示，混合數 128 時的表現和混合數 256 差不多，去除偏差值的情形下，在混合數 128 時平均正確率為 76.10 % 。對於切割語料，偏差值的考慮與否效果並不明顯。根據上述的觀察，

我們在接下來的實驗中保留偏差值的變因，並對切割過的資料採用混合數 128 的高斯混合模型。

4.4 使用修改語料庫之實驗

在 OGI-TS 語料庫中，日文比起中文和英文短少了約 15 分鐘的語料，這可能是日文辨認效能不高的原因。而測試語料也有長短不一的問題，較短的關鍵字詞語料大約 3 秒就結束，較長的自由發揮問題則有 45 秒的長度。為了做出較客觀的比較，訓練語料的大小和測試語料的長度就必須加以調整。

對於重新分類的語料庫，我們試驗了“高斯混合模型表徵器-語言模型法”、“連結聲學-語言模型法 (含偏差值)”、 以及 “連結聲學-語言模型法 (去除偏差值)” 三種方法。語料未經切割處理，傳入辨認系統的是 38 階梅爾刻度式倒頻譜的特徵參數序列。每種實驗分別測試三國語言 (中、英、日) 、六國語言 (中、英、日、德、西、印) 、 與 11 國語言 (中、英、日、德、西、印、韓、越、波斯、坦米爾、法)的辨認。

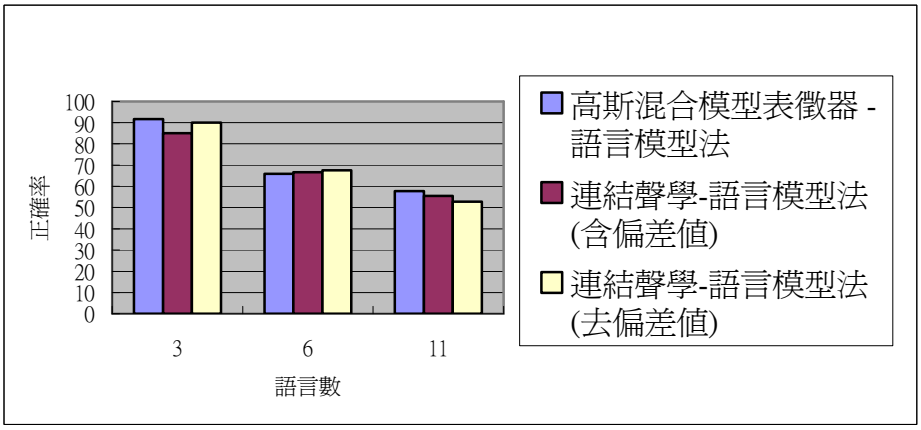


圖十一、 修改語料庫之實驗結果 (未使用切割法)

圖十一顯示，在未經切割處理的語料中，對於三國語言的辨認，效果最好的是連結聲學-語言模型法，平均正確率可以達到 91.67 % ，顯示出當訓練語料長度均衡，且測試語料長度夠的時候，對於這三國語言，可以有不錯的辨認效果。我們也發現，在三國語言中，將偏差值去掉後表現反而下降一點，表示加入偏差值的考量並不一定能提升系統效能。

對於六國語言的辨認，則有不同的現象發生，表現最好的是去掉偏差值的連結聲學語言模型法，正確率為 63.33 %，不考慮偏差值的系統則為 51.67 %， 相差 11 % 左右，顯示在六國語言辨認中，確實有幾個語言有相似度偏高的狀況。

在十一國語言辨認的實驗中顯示，單純採用特徵參數序列時，即使採用不同的系統處理，也很難提高正確率。三種方法的正確率都沒辦法超過五成，可能的原因是，所採用的混合數 128 高斯混合表徵器，將特徵參數間較細微的變化模糊掉。因此，接下來我們嘗試對特徵參數做切割處理，將相近的特徵參數先聚集起來，以其中心值為代表參數，再傳入表徵器做辨認。



圖十二、 修改語料庫之實驗結果 (使用切割法)

圖十二顯示，對於三國語言的辨認，經過切割處理的語料一樣有很高的辨認率，以高斯混合模型表徵器-語言模型法的 91.67 % 為最高。對於六國語言的辨認， 和未切割的實驗結果同樣是以去偏差值的連結聲學-語言模型法最高，有 67.5 % 的辨認效果，比起未切割的語料進步 4 % 左右。在十一國語言辨認的實驗中，加入切割處理能讓三種方法稍微提升正確率，有超過五成的正確率。效果最好的是高斯混合模型表徵器-語言模型法的 57.73 %。多數錯誤發生在英、德、西、法等四國同屬印歐語系的語言，以德文和法文之間的混淆最嚴重；中文和韓文也有蠻高的比例被辨認成越南文，顯示來源語系相近的語言會互相影響。

五、結論

本論文的主要目標在於尋找不需要標註資料的自動化語言辨認方法，所採用的系統是以高斯混合模型表徵器和語言模型為基礎，加上切割處理以及後端處理的輔助，處理語音資料的語言辨認工作。我們對串聯高斯混合模型表徵器和語言模型的“高斯混合模型表徵器-語言模型法”，以及將語言模型融合在表徵器裡面的“連結聲學-語言模型法”分別進行實驗。實驗語料採用 OGI-TS 語料庫，實驗之語言數分別是三國語言 (中、英、日) 、六國語言 (中、英、日、德、西、印) 、 與 11 國語言 (中、英、日、德、西、印、韓、越、波斯、坦米爾、法)。三國語言辨認實驗結果顯示，以 38 階梅爾刻度式倒頻譜參數，使用“高斯混合模型表徵器-語言模型法”，並加入切割處理，效果最好，有 91.67 % 的辨認率。在六國語言的辨認中，則是去偏差值的“連結聲學-語言模型法”表現最好，平均辨認率為 67.50 %。在 11 國語言的辨認上，則以 ”高斯混合模型表徵器-語言模型法”的 57.73 % 辨認率為最高。由實驗結果觀察，加入切割處理能夠使辨認效能稍微提升，表示在語言辨認的問題中，以音框為單位的特徵參數序列可能太過細微，如果能夠尋找聲學一致性更大、更為粗糙的單位來取代特徵參數，有機會能夠提升辨認的效能。

致謝

本研究受國科會專題研究計畫補助，計畫編號 NSC-93-2213-E-007-019。

參考文獻

[1] Marc A. Zissman, “Comparison of Four Approaches to Automatic Language Identification of Telephone Speech”, *IEEE Transactions on Speech and Audio Processing*, pp. 31-44, 1996

[2] Pedro A. Torres-Carrasquillo, Elliot Singer, T. P. Gleason, W. M. Campbell, D. A. Reynolds, "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification", *Proc. Eurospeech 2003*, pp. 1345-1348.

[3] Pedro A. Torres-Carrasquillo, Douglas A. Reynolds, J. R. Deller, Jr. , "Language Identification Using Gaussian Mixture Model Tokenization", *Proc. ICASSP 2002*, pp. I-757-760.

[4] A. K. V. Sai Jayram, V. Ramasubramanian, T. V. Sreenivas, "Language Identification Using Parallel Sub-word Recognition", *Proc. ICASSP 2003*, pp-81-84.

[5] Wuei-He Tsai, Wen-Whei Chang, "Discriminative training of Gaussian Mixture Bigram Models with Application to Chinese Dialect Identification", *Speech Communication*, vol. 36, pp. 317-326, 2002

[6] A. K. V. Sai Jayram, V. Ramasubramanian, T. V. Sreenivas, "Robust Parameters For Automatic Segmentation of Speech", *Proc. ICASSP 2002*, pp. I-513-516.

[7] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, " The OGI multi-language telephone speech corpus," *Proc. ICSLP 1992*, pp. II-895-898.

[8] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds, J. R. Deller, Jr. , "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Cepstral Features", *Proc. ICSLP 2002*, pp. 89-92.

A Practical Passage-based Approach for Chinese Document Retrieval

Szu-Yuan Chi[1], Chung-Li Hsiao[1], Lee-Feng Chien[1,2]

1. Department of Information Management, National Taiwan University
2. Institute of Information Science, Academia Sinica, Taipei, Taiwan

{r93036, r93042}@im.ntu.edu.tw and *lfchien@iis.sinica.edu.tw*

Abstract. TF*IDF-based methods, as they are easily to implement, are widely accepted in information retrieval industry. It is interesting to investigate a feasible and practical technique to improve the retrieval performance of these conventional IR methods. In this paper, we are going to introduce a good alternative approach that uses passage-based ranking as the second stage of the retrieval process in them.

1. Introduction

Previous research has shown that passage-level evidences can bring additional benefits to document retrieval when documents are long or span different subject areas. Callan (1994) addressed that the types of passages explored by researchers can be grouped into three classes: discourse, semantic, and window. Discourse passages are based upon textual discourse units (e.g. sentences, paragraphs and sections). Semantic passages are based upon the subject or content of the text. Window passages are based upon a number of words. Kaszkiel & Zobel (1997, 2001) also investigated the effectiveness on using passages for information retrieval. Their research showed that passages can be used in different ways. One is to provide a good basis for a question-and-answer style of information retrieval. Another approach is to use passages as proxies for documents, and documents are ranked according to similarities computed for their passages. It also addressed that fixed-length arbitrary passages of 150 words or more and starting at small interval so that the passages heavily overlap can give substantial empowerments in effectiveness, particularly for collections of long documents.

Recently Liu (2002) further demonstrated that passages can be used effectively in a language modeling framework. They found passage retrieval based on language models can provide more reliable performance than retrieval based on full documents. The previous works have proven that passage-based retrieval can get better performance than document-based ones in different applications. It's unfortunately that conventional passage-based retrieval most calculates documents ranks with the aids of passage-level indexing and single-stage processing. Although this can reduce the computing cost of passage-based retrieval, it is not flexible to consider the contextual effects of matched query terms in a passage and determine an appropriate weighting scheme through the access of indices. In some cases, it needs to analyze the content of document, for example, to determine if the occurrences of the matched query terms are appearing in a critical passage. Our research on passage retrieval is just at the beginning. Our long term research goal is to adopt sophisticated text analysis in combination with index-based ranking schemes to reach a balance between retrieval speed and accuracy.

The purpose of this paper aims at developing a practical approach to improving conventional TF*IDF IR methods, without the involvement of using some sophisticated techniques such as query expansion and ontology-based ranking. The goal is not trivial. As users' queries are often short, only a few conventional IR systems provide query expansion.

It's getting popular to improve performance by using a 2-stage strategy in retrieval task (Kwok et la., 1998). Nevertheless, most of the researches used pseudo-relevance feedback as the 2nd stage. Unlike them, the proposed approach is a two-stage retrieval process that uses passage-level analysis as the second stage in the retrieval process of conventional TF*IDF-based methods. The first stage utilizes an Okapi-based ranking algorithm to retrieve top-n relevant documents as an initial set. The passage features are then used in the 2nd stage to try to filter out irrelevant ones from the set. The proposed approach has been tested with the Chinese monolingual IR task of NTCIR-4 (Kando, 2004). The obtained preliminary result shows that the Okapi-based approach can be improved using the two-stage

process and passage-based ranking. Though the archived performance is close to NTCIR4 participants' average, the proposed approach is believed easier to be applied in commercial applications.

2. Related Work on NCTIR-4 Experiments

In this section, we will review the performance of Chinese-to-Chinese (C-C) monolingual runs achieved by NTCIR-4 participants. Table 1 shows the obtained average, median, maximum and minimum values of MAP by type of run based on rigid relevance. We use following notations:

C-C-T: all C-C <TITLE>-only runs (T-runs)
C-C-D: all C-C <DESC>-only runs (D-runs)

Table 2 shows the top 5 groups tanked according to MAP values of D-runs based on rigid relevance. I2R-C-C-D-01 which was based on ontological query expansion achieved the best performance. The research work of the top three groups is briefly summarized.

Table 1 MAP of overall C-C runs

	Average	Median	Min	Max
C-C-T	0.1943	0.1881	0.1327	0.3146
C-C-D	0.1826	0.1741	0.1251	0.3255

Table 2 Top-ranked 5 groups (C-C, Rigid, D-runs)

Run-ID	Mean Average Precision
I2R-C-C-D-01	0.3255
OKI-C-C-D-04	0.2274
Pircs-C-C-D-02	0.2150
RCUNA-C-C-D-01	0.2087
UniNE-C-C-D-03	0.2011

I2R: Using knowledge ontology. (Yang et al., 2004)

The I2R group has built knowledge ontology for query terms by using a search engine on the Internet with manual verification. Firstly, they automatically extract terms from documents and use them to build indexes; secondly, they use short terms in the query and documents to do initial retrieval; thirdly, they build ontology for the query to do query expansion and implement second retrieval. Finally, they use long terms to reorder the top N retrieved documents. The knowledge ontology appears to include narrower terms, related terms and so on. They combine information from the ontology with that from pseudo-relevance-feedback to expand query terms.

OKI: (Nakagawa and Kitamur, 2004)

As widely known, pseudo-relevance-feedback (PRF) of blind feedback brings us improvement or retrieval performance. Some research groups, however, challenge to use **non-standard PRF** methods. For example, the OKI group adopts Ponte's ration method (Ponte, 1998).

PIRCS: (Kwok et al, 2004)

For PIRCS group, Chinese monolingual retrieval was performed as before (Kwok, 2002): based on combination of retrieval lists using bigram + unigram, and short word + character indexing.

3. Two-stage Document Retrieval

In our research, we pursue a simpler approach that can achieve acceptable performance. We propose a two-stage passage-based document retrieval approach. The retrieval process performed at the first stage is similar to that in conventional n-gram-based Chinese IR systems. That is, all unique character unigrams and bigrams in a document except some stop characters will be extracted to form the term vector of the document and a TF*IDF-based weight value is assigned to each term as its significance value. In addition, an Okapi-based similarity estimation function (Robertson et al., 1994)

is adopted to estimate the relevance score between the input query and each indexed document. As any two feature dimensions of a vector-space model are assumed independent, at the first stage it doesn't consider the contextual effects between the query terms and the document of concern.

The second-stage process is proposed as an additional retrieval process performing detailed passage analysis. As shown in previous research (Callan, 1994; Kaszkiel & Zobel, 1997), passage-level evidences can bring additional benefits to document retrieval when documents are long or span different subject areas. The additional process re-examines the occurrences of the query terms and analyzes their presences at the passage level of a document. The addition of the second-stage process is tried to investigate if there is a simple approach to improve conventional document retrieval methods, without the involvement of sophisticated techniques, such as query expansion and ontology-based ranking.

An overview of the two-stage document retrieval approach is shown in Figure 1, in which some techniques and strategies which are planned to be tested are listed.

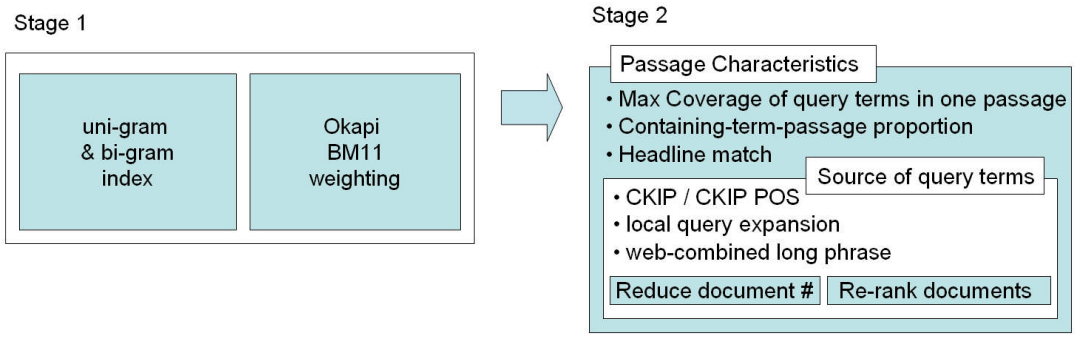


Figure 1: An overview of the two-stage document retrieval approach.

3.1 The Stage One Process

As described previously the purpose of the first-stage processing is to form uni-gram and bi-gram feature vectors for input query and all documents, and an Okapi similarity estimation algorithm, i.e., BM11 defined below, is adopted to retrieve and rank these documents. For more information about BM11 can be referred to (Robertson et al., 1994).

$$\text{(BM11)} \qquad w = \frac{tf}{(\frac{k_1 \times d}{\Delta} + tf)} \times \log \frac{N - n + 0.5}{n + 0.5} \times \frac{qtf}{(k_3 + qtf)} + k_2 \times nq \frac{(\Delta - d)}{(\Delta + d)}$$

N: Number of items (documents) in the collection
n: Collection frequency: number of items containing a specific term
tf: Frequency of occurrence of the term within a specific document
qtf: Frequency of occurrence of the term within a specific query
d: Document length arbitrary units
Δ : Average document length
k_i: Constants used in various BM functions

To realize the achieved performance, our research was conducted based on the Chinese monolingual IR task of NTCIR4 (Kishida, et al, 2004). We tested the Okapi-based approach (the first-stage processing) and the obtained MAP (Mean Average Precision) value was about 0.18, which is close to NTCIR4 participants' average and is thus taken as the baseline. The obtained MAP value for each test topic is shown in Figure 2; and as in Figure 3, it was found that for most test topics the obtained recall rates of top 10000 retrieved documents are very high and almost close to 1. The second-stage process is, therefore base on the answer set (retrieved documents) to re-rank some relevant documents to higher position.

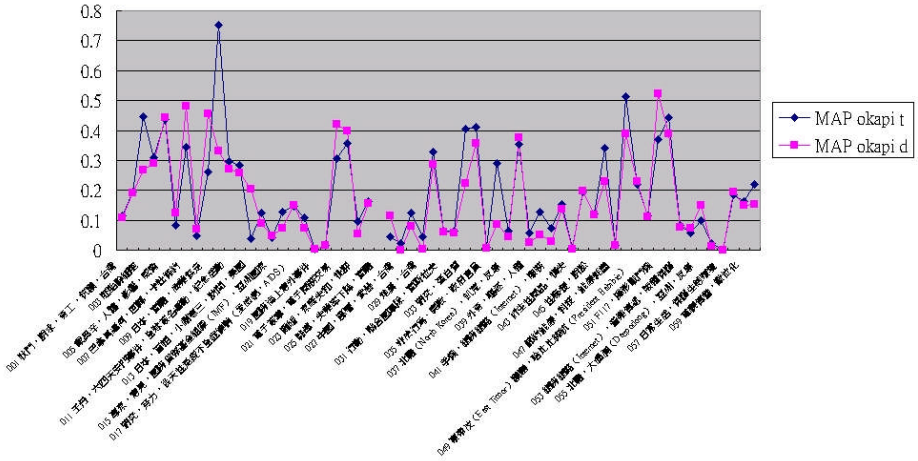


Figure 2: Obtained MAP values of top 10,000 retrieved documents using Okapi BM11 in NTCIR4, in which *t* means the result of title run and *d* is that of description run.

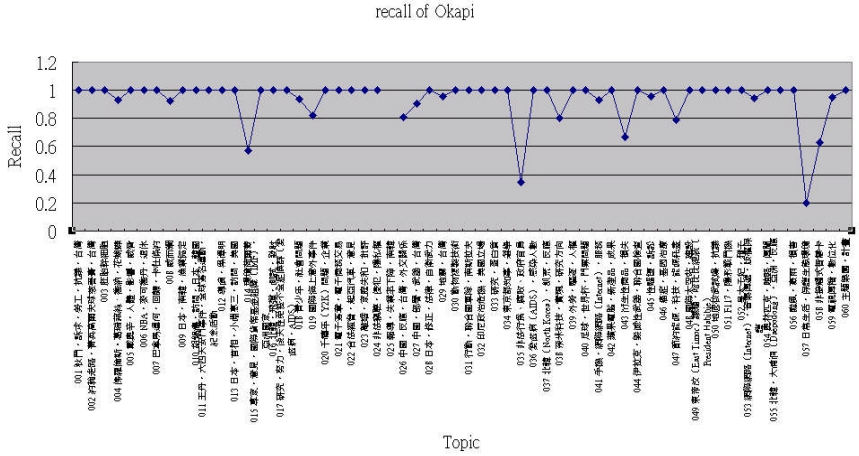


Figure 3: Obtained recall rates of top 10,000 retrieved documents using Okapi BM11 in NTCIR4, which are with respect to each test topic.

3.2 Observation after Stage One Processing

As discussed conventional Okapi algorithms treat term as independent entities and ignore semantic or locality properties of terms in documents. According to the answer set, we observed a few interesting properties of passages that can be further analyzed.

We examined the answer sets with four types of relevance in answers: highly relevant (S), relevant (A), partial relevant (B), and irrelevant (C). There were three features to be observed. An interesting finding is that relevant documents (type S, A and B) have a higher chance to contain matched query terms in their passages than irrelevant documents. The finding attracted us to do more experiments to be introduced in the next sections.

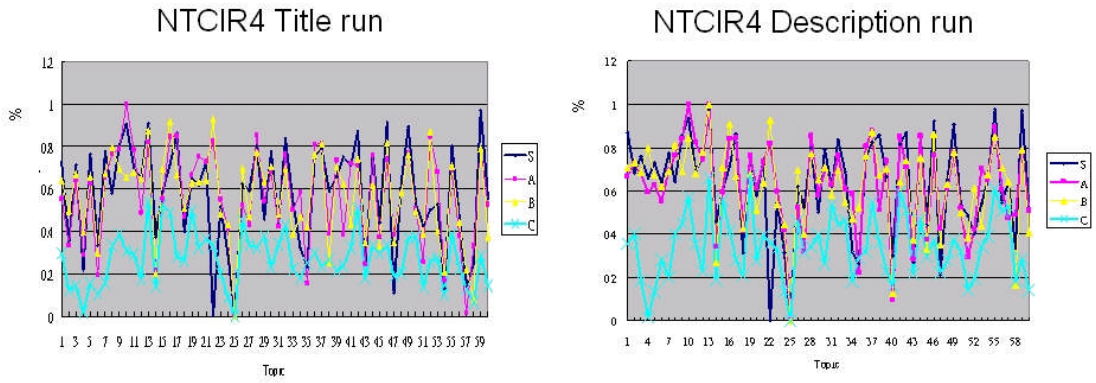


Figure 4: For most topics, relevant documents (S+A+B) get higher ratios of passages with matched query terms than irrelevant ones (C). Left: NTCIR4 title run. Right: NTCIR4 description run.

4. Passage-based Retrieval & Experiments

The second-stage process performs passage-based retrieval. The documents retrieved at the first stage will be segmented into passages via period markers in text as passage boundaries. In literature, passage is often defined as fix-length short article, and every passages are half-overlap with previous one. Passages are indexed and ranked, and the relevance of document is decided by its passage. If passage is taken as the unit to calculate scores and rank, the retrieval cost of passage become several times of that of document. In our work, we see a complete sentence separated by periods as a passage.

Query terms are main factors to effect retrieval outcome. Our Chinese segmentation tool is powered by the CKIP group, Academia Sinica. Because our test data set is a news story material, some news events use new words or longer terms that dictionary don't cover. We used the Web as the corpus to segment unknown words and extract longer terms. By sending all words in a query to Google, we can get a result that some words are frequently concatenated together in search result snippets. Adapting these combined phrases was found can remedy the lack of new terms in the dictionary.

The ranking policy is based on observation on relevant documents in which query terms are often concatenated in one paragraph. It is curious to know if using query term coverage in passages to re-score documents and change ranks of retrieval results could archive a better MAP value. We match query terms for each passage and increase the score of the document based on original score. Several passage-based ranking strategies were proposed and tested.

4.1 Passage-based Ranking

1. Ranking with Average Term Coverage of Passages

In the first experiment, the relevance value of a document was re-scored as the weighted sum of its Okapi score (the value of BM11) and the average term coverage rate of the composed passages (namely Strategy I). The term coverage rate is the percentage of the unique segmented query terms appearing in a passage, and the average term coverage rate means the average value of all passages' term coverage values. This experiment was performed to see if the addition of passage ranking score can improve the Okapi result. Figure 5 and Table 3 summarize the result of the experiment, in which the obtained MAP values are depicted with the change of the weight w from 0 (only using Okapi score), 0.01 to 0.10. The best results were obtained at w= 0.03 and 0.04, which perform better than that only using the Okapi score.

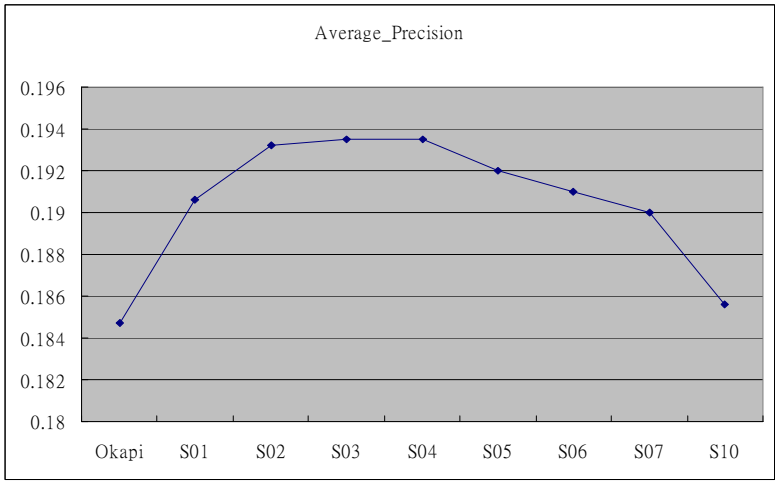


Figure 5: The MAP values obtained with Strategy I, which are depicted with respect to the increasing of the weight from 0.0 to 0.10.

Table 3 Selected 11-ponits precision rates obtained with Strategy I when w =0.01, 0.04 and 0.07

Recall	Precision						
	Okapi	Passage w=0.01	% change	Passage w=0.04	% change	Passage w=0.07	% change
0.00	0.4985	0.5199	+ 4.29	0.5360	+ 7.52	0.5319	+ 6.70
0.01	0.3729	0.3902	+ 4.64	0.3885	+ 4.18	0.3938	+ 5.60
0.20	0.2817	0.2999	+ 6.46	0.3094	+ 9.83	0.3063	+ 8.73
0.30	0.2496	0.2479	-0.68	0.2486	-0.40	0.2505	+ 0.36
0.40	0.2091	0.2147	+ 2.68	0.2193	+ 4.88	0.2074	-0.81
0.50	0.1742	0.1817	+ 4.30	0.1853	+ 6.37	0.1798	+ 3.21
0.60	0.1501	0.1530	+ 1.93	0.1574	+ 4.86	0.1487	-0.93
0.70	0.1169	0.1193	+ 2.05	0.1195	+ 2.22	0.1186	+ 1.45
0.80	0.0920	0.0944	+ 2.61	0.0928	+ 0.87	0.0923	+ 0.32
0.90	0.0680	0.0695	+ 2.20	0.0754	+ 10.88	0.0749	+ 10.14
1.00	0.0476	0.0477	+ 0.21	0.0519	+ 9.03	0.0511	+ 7.35
Avg Precision	0.1847	0.1906	+ 3.19 %	0.1935	+ 4.76 %	0.1900	+ 2.87 %

2. Ranking with Max Term Coverage of Passages

In the second experiment, the passage-based relevance score of a document is measured as the maximum term coverage of its composed passages, that is, for a document with three passages and the query with four segmented terms. If three passages contain 3, 1, 2 query terms respectively, then the max term coverage is 0.75 (3/4). That is quite simple to calculate. The relevance value of a document was then re-scored as the weighted summation of its Okapi score and the new passage-based relevance score (namely Strategy II). Figure 6 and Table 4 summarize the experimental result, in which the precision rates obtained with the addition of average term coverage to the Okapi are depicted with the change of the weight w from 0.0 to 1.0. The best result was obtained at w= 0.6, which perform slightly better than that using the Okapi score but worst than using Strategy I.

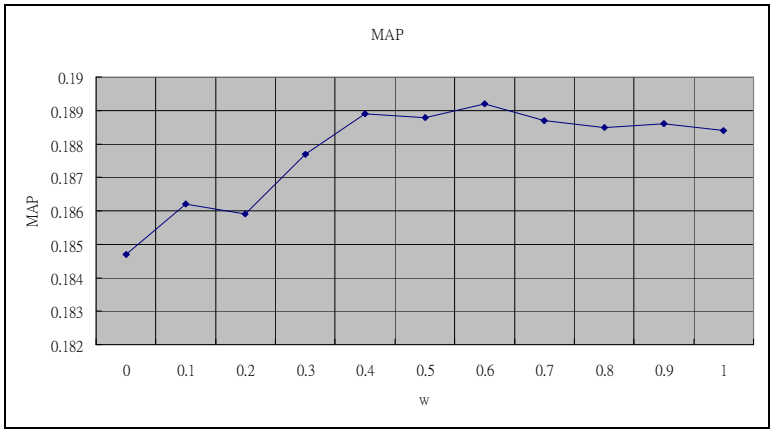


Figure 6: The MAP values obtained with Strategy II, which are depicted with respect to the increasing of the weight from 0 to 1.

Table 4: The MAP values obtained with Strategy II when w =0.4, 0.5 and 0.6

	Okapi	Passage w = 0.4	% change	Passage w = 0.5	% change	Passage w = 0.6	% change
Avg. Precision	0.1847	0.1889	+ 2.27 %	0.1888	+ 2.22 %	0.1892	**+ 2.44 %**

3. Ranking with Average Term Coverage of the Top Three Passages

In the third experiment, the passage-based relevance score of a document is measured as the average term coverage of the top three passages. The relevance value of a document was then re-scored as the weighted sum of its Okapi score and the new passage-based relevance score (namely Strategy III). Figure 7 and Table 5 summarize the experimental result, in which the MAP values obtained with the addition of average term coverage to the Okapi are depicted with the change of the weight w from 0.0 to 1.0. The best results were obtained at w= 1.0, which perform better than that only using the Okapi score.

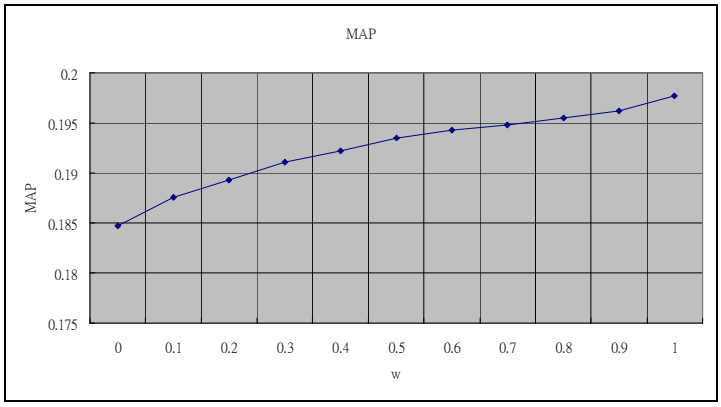


Figure 7: The MAP values obtained with Strategy III, which are depicted with respect to the increasing of the weight from 0 to 1.

Table 5: The MAP values obtained with Strategy III when w =0.8, 0.9 and 1.

	Okapi	Passage w= 0.8	% change	Passage w = 0.9	% change	Passage w= 1	% change
Avg Precision	0.1847	0.1955	+ 5.85 %	0.1962	+ 6.23 %	0.1977	**+ 7.04 %**

4. Ranking with Percentage of Passages Containing Query Terms

In the forth experiment, the passage-based relevance score of a document is measured as the percentage of passages containing query terms. For a document with three passages and the query with four segmented terms, if two of the three passages contain at least one query term, then the percentage of passages containing query terms is 0.66 (2/3). The relevance value of a document was also re-scored as the weighted sum of its Okapi score and the new passage-based relevance score (namely Strategy IV). Figure 8 and Table 6 summarize the experimental result, in which the obtained MAP values are depicted with the change of the weight w from 0.0 to 1.0. The best results were obtained at w= 0.3 and 0.2, which perform better than that only using the Okapi score.

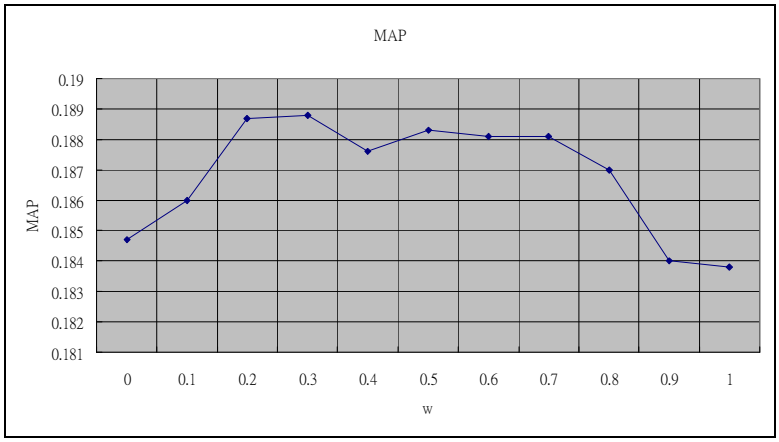


Figure 8: The MAP values obtained with Strategy IV, which are depicted with respect to the increasing of the weight from 0 to 1.

Table 6: The MAP values obtained with Strategy IV when w =0.2, 0.3 and 0.5.

	Okapi	Passage w=0.2	% change	Passage w=0.3	% change	Passage w=0.5	% change
Avg. Precision	0.1847	0.1887	+ 2.22 %	0.1888	**+ 2.22 %**	0.1883	+ 1.95 %

4.2 Ranking with "Headline" Matching

Our next group of experiments was performed to compare the results of "Headline" matching with the Okapi results and the combination of passage-based ranking scores. The NTCIR-4 document set is a news story set. News headlines normally contain keywords. We assume the query terms appearing in headlines are more critical. The relevance value of a document was re-scored as the weighted sum of its Okapi score and the headline-based relevance score (namely Strategy V). Critical keywords may appear in many sentences. If a document contains many occurrences of the critical keywords, the headline-matched query terms, the document will be assigned a higher score than that contains only other keywords. This strategy can help remove some news articles containing with query terms but irrelevant topics.

Figure 9 and Table 7 summarize the experimental result, in which the obtained MAP values are depicted with the change of the weight w from 0.1 to 0.25. The best result was obtained at w= 0.13. As can be seen, the result is better than previous set of experiments.

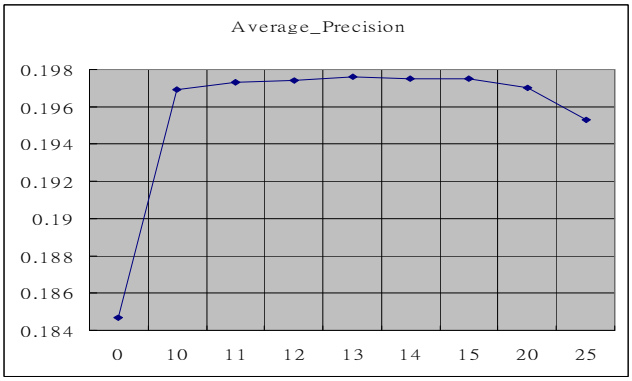


Figure 9: The MAP values obtained with Strategy V, which are depicted with respect to the increasing of the weight from 0.1 to 0.25.

Table 7: The MAP values obtained with Strategy V when w =0.1, 0.13 and 0.25.

	Okapi	HL w=0.1	% change	HL w=0.13	% change	HL w=0.25	% change
Avg Precision	0.1847	0.1969	+ 6.6 %	0.1976	+ **6.98** %	0.1953	+ 5.74 %

4.3 Combining Passage-based Ranking and "Headline" Matching

The last experiment was to compare the result of the combination of passage-based ranking and headline matching with the Okapi result. Figure 10 and Table 8 Figure 9 and Table 7 summarize the experimental result, in which the obtained MAP values are depicted with the change of the weight w from 0.01 to 0.02. The best result was obtained at w= 0.015. It is easy to see that the result is the best one that can be obtained in our experiment.

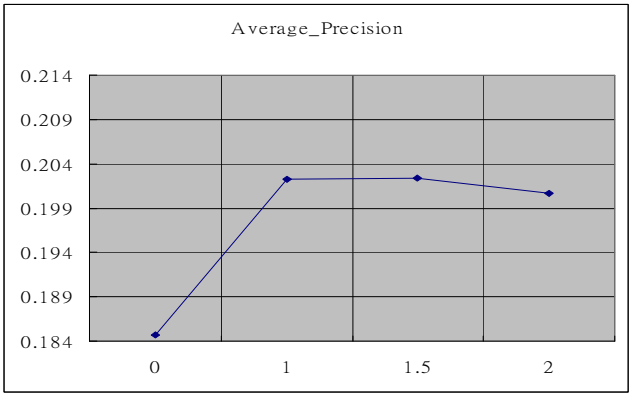


Figure 10: The MAP values obtained with Strategy VI, which are depicted with respect to the changes of the weights.

Table 7: Illustrated MAP values obtained with Strategy VI, the combination of Strategy I and V.

	Okapi	Passage w=0.01 HL 13%	% change	Passage w=0.015 HL 13%	% change	Passage w=0.02 HL 13%	% change
Avg Precision	0.1847	0.2022	+ 9.5 %	0.2024	+ **9.6 %**	0.2006	+ 8.6 %

5. Conclusion

In this paper, we have introduced a good approach to improving conventional TF*IDF methods. The approach is simple but practical. It combines the Okapi-based ranking algorithm with passage-based ranking strategies. The result also shows that using headline matching to determine critical keywords in queries is useful. A set of experiments have been conducted on the NTCIR-4 task for Chinese information retrieval. Although the proposed approach is simple, it is believed easily to be implemented and applied to the applications in industry.

6. Reference

1. Kaszkiel, M. and Zobel, J. (2001). Effective ranking with arbitrary passages. *Journal of the American Society For Information Science and Technology*, 52(4):344-364.

2. Callan, J.P. (1994). Passage-level evidence in document retrieval. In B.W. Croft & C.J. van Rijsbergen (Eds.), *Proceedings of the 17th annual international ACM-SIGIR conference on research and developments in information retrieval*, Dublin, Ireland, July (pp. 302-310), New York: ACM.

3. Kaszkiel, M. and Zobel, J. (1997). Passage retrieval revisited. In N. J. Belkin, D. Narasimhalu, & P. Willett (Eds.), *Proceedings of the 20th annual international ACM-SIGIR conference on research and development in information retrieval*, Philadelphia, PA (pp. 178-185).

4. Xiaoyong Liu, W. Bruce Croft. Language models for information retrieval: Passage retrieval based on language models. *Proceedings of the eleventh international conference on Information and knowledge management*, pp.375-382, November 2002

5. S. E. Robertson, S. Walker, S. Jones, M. M. HancockBeaulieu, and M. Gatford. *Okapi at trec*-3. In TREC-3, 1994.

6. K Kishida, K Chen, S Lee, K Kuriyama, N Kando, HH Chen, S.H. Myaeng, K. Eguchi. Overview of CLIR Task at the Forth NTCIR Workshop. *Proceedings of the 4th NTCIR*, 2004.

7. Noriko Kando, Overview of the Fourth NTCIR Workshop. *Proceedings of NTCIR-4 Workshop*, 2004.

8. Lingpeng Yang, Donghong Ji, and Li Tang. Chinese Information Retrieval Based on Terms and Ontology. *In: Proceedings of NTCIR-4 Workshop*, 2004.

9. Tetsuji Nakagawa and Mihoko Kitamura. NTCIR-4 CLIR Experiments at Oki. *In: Proceedings of NTCIR-4 Workshop*, 2004.

10. J. M. Ponte. A Language Modeling Approach to Information Retrieval. *Ph.D. Thesis, Graduate School of the University of Massachusetts Amherst*, 1998.

11. Kui-Lam Kwok, Norbert Dinstl and Sora Choi. NTCIR-4 Chinese, English, Korean Cross Language Retrieval Experiments Using PIRCS. *In: Proceedings of NTCIR-4 Workshop*, 2004.

12. K.L. Kwok. NTCIR-2 Chinese and cross language experiments using PIRCS. *In: Proceedings of NTCIR-2 Workshop*, 2002.

Web Information Extraction for the Creation of Metadata in Semantic Web

Ching-Long Yeh and Yu-Chih Su
Department of Computer Science and Engineering
Tatung University
Taipei, Taiwan
chingye@cse.ttu.edu.tw g9206033@ms2.ttu.edu.tw

Abstract

In this paper, we develop an automatic metadata creation system using the information extraction technology for the Semantic Web. The information extraction system consists of preparation part that takes written text as the input and produces the POS tags for the words in the sentences. Then we employ finite state machine technology to extract the units from the tagged sequences, including complex words, basic phrases and domain events. We use the components of an NLP software architecture, GATE, as the processing engine and support all required language resources for the engine. We have carried out an experiment on Chinese financial news. It shows promising precision rate while it need further investigation on the recall part. We describe the implementation of storing the extracted result in RDF to an RDF server and show the service interface for accessing the content.

1. Introduction

Semantic Web is an emerging technology working by building a metadata layer upon the current web and using the metadata description language to describe the resources on the WWW [1]. The metadata layer is a structured information layer, that is, the information model of RDF [2]. The service programs built upon the metadata layer can provide an efficient way for users to find the information they need according to the concepts. Furthermore, the metadata layer supports agents to provide the function of service automation for users, called Semantic Web Service [3]. The semantic metadata layer can be constructed either in a manual way using annotation tools, such as Annotea [4] or automatically using the natural language processing technology [5]. The former approach is good at higher precision rate while suffered in efficiency. The latter approach can be used to sort out the problem of the other; however, it needs very high cost to achieve similar result using the manual way.

Information extraction is a low-cost approach to natural language processing using the finite-state automata technology to extract specific noun sets and information matching specific syntax and semantic templates [6]. In this paper we employ the finite-state automata technology to extract information from the resources on the WWW to serve as the describing information of the resources, that is, the metadata. The extracted metadata is then used as a part in the Semantic Web.

Ontology, the semantic schema used to build metadata layer upon semantic web, contains the concept hierarchy of specific domain knowledge and the detail features of all concepts type. In this paper we build the templates for information extraction based on the feature structure of concepts type of ontology.

A typical information extraction system, such as FASTUS of SRI [7], ANNIE of the University of

Sheffield [8], is constructed by using a cascade of finite-state grammars. Based on such mechanism, we can identify the units of words, complex words, basic phrase group, co-reference resolution, identifying event structures *etc.*, in text. In our implementation, we use components of GATE framework [8] to develop our Chinese information extraction system. The system in sequence consists of the following modules: word segmentation, part-of-speech (POS) tagging, name entity extraction, noun group extraction and event extraction. The word segmentation module is based on matching entries in the lexicon with the input sentences. We develop a Brill POS tagger [3] in order to provide accurate POS information for latter modules. Then the succeeding extraction modules are developed by using the finite-state machine technology. The JAPE (Java Annotation Processing Engine) engine is used as the basis of the extraction modules. The task of developing the extraction functions is therefore to define the regular expressions of the respective extraction objects.

The output of the information extraction system is converted into RDF and is imported into an RDF indexing system, Sesame [9]. Therefore user can access the content of the extracted metadata through the conceptual search services interfaces provided by Sesame.

In brief, our goal is to build an ontology-driven information extraction system that processes sentences and transforms extracted data into RDF automatically. The extracted metadata is used to build a knowledge base. Services of conceptual search and semantic navigation are implemented based on the inference engine of the knowledge base. In Section 2, we describe the architecture of the Semantic Web and the related representation languages used to construct the metadata layer. In Section 3, we describe the architecture of the information extraction and integration with the Semantic Web architecture. In Section 4 we describe components of the information extraction system. In Section 5, we describe the implementation and the performance of the system.

2. Architecture and Representations of Semantic Web

In this section, we first describe the metadata layer of the Semantic Web and system architecture for managing system the layer. Then we describe the representations of the metadata layer.

2.1 Metadata Layer of Semantic Web

Semantic Web is an extension of the current Web where information is given well-defined meaning, better enabling computers and people to process in cooperation. Resources on the Web are given explicit meaning using markup language to make up a metadata layer in addition to the current information pool as summarized in Fig. 1. The Web Ontology Language (OWL) [10] provides rules for defining knowledge structures, i.e., ontology, in order that instances of knowledge can be created based on the common structures. OWL is an extension to the Resource Description Framework (RDF) [11] by adding vocabularies used to define the knowledge structures instead of the tree structures in XML documents.

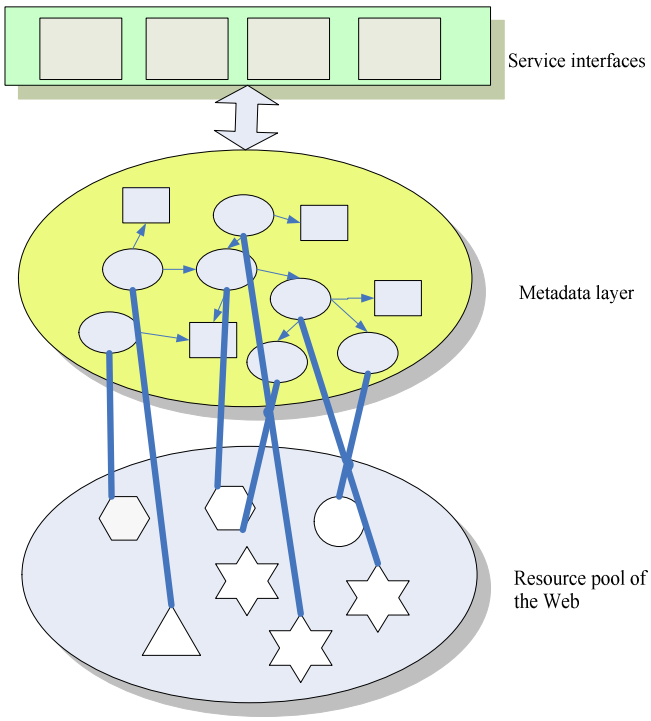


Figure 1: Architecture of the metadata layer of the Semantic Web

The Semantic Web framework can be summarized as providing a metadata layer in content-interoperable languages, mainly in RDF, which intelligent or automatic services can be made by machines based on the layer. The typical architecture for managing the metadata layer, for example, KA2 [12] and Sesame [13], consists of an ontology-based knowledge warehouse and inference engine as the knowledge-based system to provide intelligent services at the front end, such as conceptual search and semantic naviagtion for user to access the content as summarized in Fig. 2. In the backend is the content provision component, consisting of various tools for creating metatdata from unstructured, semi-structured and structured documents. In this paper, the information extraction system we developed is used as a component of the back end.

2.2 Representing the Schema and Instances of the Metadata Layer

The metadata layer of the Semantic Web as shown in Figure 1 is built on RDF. The concptual model of RDF is a labled directed graph [11], where the nodes in the graph are identified by using URIs [14] or literals representing resources and atomic values of some kinds, respectively. The label on the directed arc connecting two nodes, from a recource to another, or a resource to a literal, represents the relationship between both ends. An arc, also identified using URI, connecting two nodes is a triple of the subject-predicate-object, similar to the structure of simple declarative sentetence. Thus an RDF document can be seen as a list of triples. A triple states a fact about a subject, a resource on the Web. Since the nodes in RDF are identified using the addressing scheme of the Web, an RDF document can be easily combined with other to form an integrated description of resources on the Web. For example, the graph shown in Figure 3 represents “The author of http://www.cse.ttu.edu.tw/chingyeh is Ching-Long Yeh”.

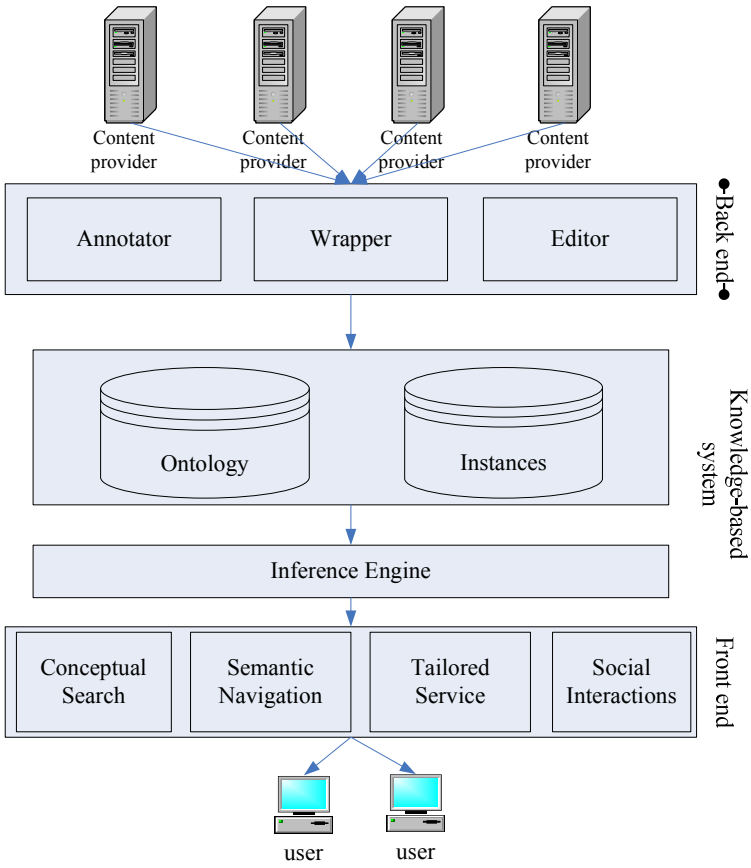


Figure 2: System architecture for managing the metadata layer of Semantic Web

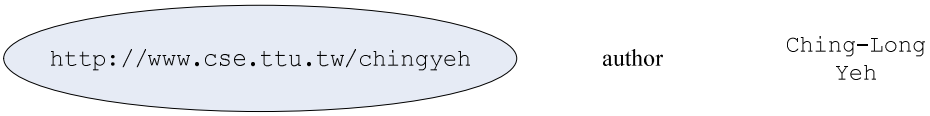


Figure 3: An RDF fragment

Representing in XML it becomes as follows.

```
<rdf:RDF
    xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
    xmlns:t='http://www.cse.ttu.edu.tw/sw'>
   <rdf:Description about="http://www.cse.ttu.edu.tw/chingyeh">
     <t:author>Ching-Long Yeh</t:author>
    </rdf:Description>
</rdf:RDF>
```

The information layer represented by using RDF is a generic relational data model, describing the relationship between resources or between resource and atomic values. The meaning of the resources can then be found in the domain knowledge base, that is, ontology. The representation of ontology in the Semantic Web is an extension of RDF, OWL [10].

3. Information Extraction System and Metadata Creation of the Semantic Web

As mentioned previously, information extraction can provide the function of creating metadata

for resources on the web. Thus we integrate the information extraction function as a part of the back end of the Semantic Web system as described in Section 2. The system we design consists of three parts: information extraction back end, ontology-based store and service front end as shown in Figure 4. The work of the back end is to extract the domain events from the relevant documents returned by search engine. The ontology-based store converts the extracted data to specific formats and stores them into the repository. We have implemented two kinds of stores; one is RDF store and the other is a frame-based store. In this paper, we focus on the former one. Finally, the service front end provides interface for user to access the extracted content.

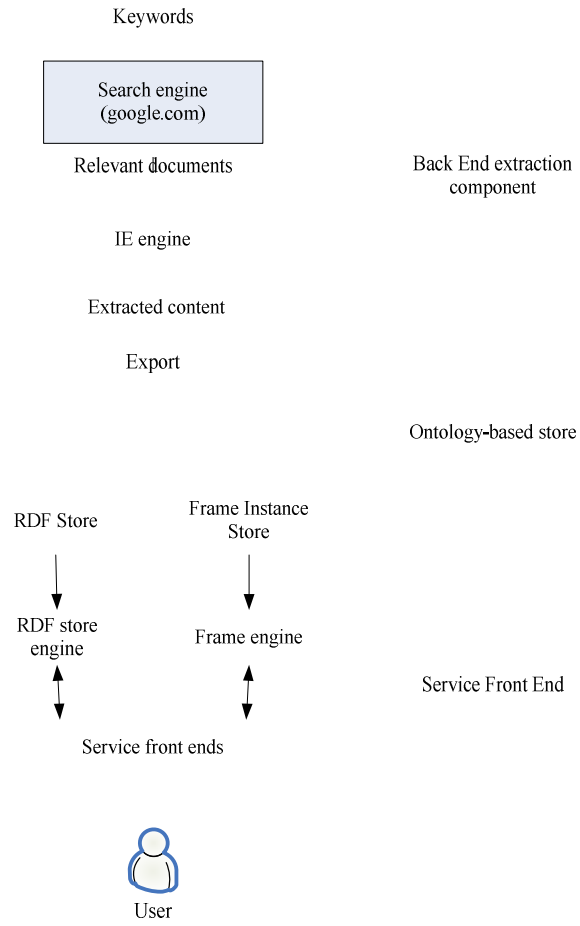


Figure 4: Integrating information extraction function with Semantic Web system

The back end of the system includes several information extraction components, including Chinese Tokenizer, Gazetteer, Sentence Splitter, POS tagger and Name-entity Transducer, as shown in Figure 5. They are used to extract specific domain events from a large number of relevant documents in a cascaded way. The relevant documents we use here are returned from calling a search engine and served as the domain data. The flow of domain events extraction starts from Chinese Tokenizer which is used to recognize the basic tokens, that is words. Gazetteer is used to recognize the special terms of domain, for example, number, day, etc. Sentence splitter and POS tagger are used to tag each token with its correct part-of-speech. Finally, the Name-entity transducer is the real work used to recognize the domain events. All processing flow is shown in the following figure.

Figure 5: The components in the information extraction system.

The central part of the system, the ontology-based store, consists of two parts: transformation mechanism and repository. The transformation mechanism transforms the original extracted data into different specific format according to the interest of users. It works as if a common interface which takes extracted domain events as input and exports the specific format data as output. After exporting the specific format data, we should have a place to store the data so that our query service has a knowledge source to query. So we should provide specific repository for specific data format to store.

The front end of the system consists of several services, including conceptual search, explore service, extraction service and clear service. The conceptual search is a search which attempts to conceptually match results with the query. It does not match by the key words, rather their concepts. The function of this service is to help user to find the data. The explore service is to explore the data according the classes and properties of ontology. The extraction service is used to extract all data from the repository. The clear is used to delete all data from the selected repository. The goal of front end is to provide several services based on semantic web for users to use.

4. Information Extraction System

Instead of building up such framework from scratch, here we adopt the reusable software component provided by the GATE framework [8] to construct the information extraction system for our purpose. Actually this is the approach having been used by ANNIE [8], a Nearly-New Information Extraction System. It consists of components used in the task of information extraction; furthermore, it provides a pipelined environment to chain all relevant components together. After all required components are added, the pipeline executes these components in a cascaded way and the results returned from current or previous components are kept so that the succeeding component can use these results to do further process. The information extraction system built by using the components in GATE provides the necessary engines; we therefore focus our attention on developing the knowledge resources necessary for each component in the system.

The information extraction system consists of five components, which can be divided into two parts, the preparation and the extraction phases as shown in Fig. 5. The preparation part is composed of the front four components, that is used to process the raw input sentences and produces word sequence tagged with their correct part-of-speech information. The other part is used to extract domain events. In the following, we describe these two parts in order.

4.1 Preparation phase

We describe in order the components in the preparation phases as shown in Fig. 5.

Chinese Tokenizer

Chinese Tokenizer is used to segment the input sentences into sequences of meaningful units, namely tokens, and identify their types such as words, numbers and punctuation. The knowledge

sources of the tokenizer include a dictionary and ambiguity resolution rules. For the former part we employ the Chinese dictionary of the CKIP [15]. For the latter part we only employ the "long-term first" heuristic rule for disambiguation.

Once a word is matched, it is assigned with a type. For example, the three tokens "籃球", "1010" and "，" are assigned the type "words", "numbers" and "punctuation". For the unmatched case from the current character by looking up the dictionary, we assign the type "words" to that character and then continue the matching work from the next character.

Gazetteer

The Gazetteer is used to define various categories of specific domain in which each category contains a large number of instances with the same concept. It is very convenient for later domain events processing. By grouping the instances of the same concept in advance, in the course of domain events recognition, we can just easily use the concept rather than the redundant words to construct the patterns. By reusing the function of ANNIE Gazetteer, we just need to focus on the definition of domain key words. For the Gazetteer function to function properly, it needs one or more `.lst` files and a `.def` file.

A `.lst` file is simply a file containing the key words of specific concept. For instance `company.lst` contains the names of company in every line of the file. The file `lists.def` is used to describe which `.lst` files we use. Except for the name of `.lst` file, we have to define the major type which represents more general concept of that file and, optionally, a minor type which represent more specific concept of that file. Here we just use the major type for our purpose. For each `.lst` file, we define the major type served as their individual concept, so we can later use the concept rather than the large number of words to recognize the domain events. This is very useful feature in the course of domain events recognition.

```
Company.lst:Company
day.lst:day
date_key.lst:date_key
hour.lst:hour
time_ampm.lst:time_ampm
production.lst:production
…
```

Sentence Splitter

The purpose of sentence splitter is to segment the full text into sentences. This component is required for the POS tagger because the processing of POS tagger is in the way of sentence by sentence. The design of sentence splitter is easy. We just need to define the symbols representing the end of a sentence. Then, segment the full text into sentences according these symbols. In this paper we the sentential mark , "。", comma, "，" and question mark "？" as the symbols of the end of sentences.

POS tagger

The POS tagger is used to tag each word recognized by the Chinese Tokenizer with its part-of-speech. By tagging each word with the part-of-speech in advance, in the course of domain events recognition, we can just use the POS rather than the redundant words to construct the patterns.

There are two steps to tag each word with part-of-speech. In order to label each word with its most-likely part-of-speech, we first need a lexicon dictionary in which each word is tagged with its most-likely part-of-speech. Then, use the initial state processor to label each word recognized by the Chinese Tokenizer with its most-likely tags by looking up the lexicon dictionary. We employ the Chinese balanced corpus developed by the CKIP [16] in this step.

Second, use the contextual rules to filter out the incorrect part-of-speech according to the contextual information. We can not find existing POS tagger to be employed in this paper; we therefore develop a POS tagger by employing Brill algorithm [5].

4.2 Domain events extraction

After the basic words and their features obtained in the preparation phase, the extraction phase further processes to obtain information about domain events. We employ the method used in FASTUS [7] to process the domain event from processing simpler units, the complex words to phrasal constituents, basic phrases, and finally domain events. All information produced by this stage is recognized by using the JAPE rules; we describe how to formulate JAPE rules at each processing step.

Complex words

The purpose of processing complex words is to identify the string of company, location, production, date, money, and year *etc*. First, we define JAPE rule by using the features produced by the gazetteer as the condition, for example, the rule of company word as below.

```
Rule: company
(
 {Lookup.majorType == Company}
):com
-->
 :com.Company = {kind = "company"}
```

It is obviously not sufficient to identify the complex words by simply using the features returned by the gazetteer. In Chinese, a noun phrase has the head-final feature [17]. In other words, the head noun appears in the final position of a noun phrase. Here we treat the feature word returned by the gazetteer as the head noun, and its preceding adjacent nouns as its modifiers. This observation is formalized as a number of JAPE rules for complex words. For example, if one or more noun connects together with the company identifier, then we recognize the preceding nouns as a company word. The company identifier is defined in the stage of Gazetteer, including “公司”, ”集團”, “航空” and so on.

```
Macro: COMPANY_IDENTIFIER
({Lookup.majorType == company identifier})
Rule: companyWithIdentifier
(
 ({Token.category == Na})+
  (COMPANY_IDENTIFIER)
):com
-->
 :com.Company = {kind = "company"}
```

In the following is another example of complex word rule for the recognition of money.

```
Rule: money1
//Priority:30
(
 (NUMBER)
 ({Token.category == PERIODCATEGORY}|{Token.string == "."})?
 (NUMBER)?
 ({Token.category == Neu})
 {Token.category == Nf}
 (MONEY_KIND)?
):money
-->
 :money.Money = {kind = "money"}
```

The recognized complex words are then used for later processing of basic phrases and domain events.

Basic phrases

The purpose of processing basic phrases is to identify several word classes, including noun group, verb, preposition, conjunction etc. For simple classes such as verb, preposition, conjunction etc, we can just use the part-of-speech tagged to each word by POS tagger to recognize them. The example using the POS to recognize the simple classes is as follows.

```
Rule: preposition
//Priority:30
(
 {Token.category == P}
):preposition
-->
 :preposition.Preposition = {kind =
"preposition"}
```

For the complicated classes like noun group, we combine determiner and other modifiers together with the head noun to form a noun phrase. For example, in the following example, the part-of-speech “Neu” and “Nf” together with the noun, we can recognize the noun group like “三本書”or “四間分公司”.

```
Rule: nounGroup1
//Priority:30
(
 {Token.category == Neu}
 {Token.category == Nf}
 {Token.category == Na}
):noungroup
-->
 :noungroup.NounGroup = {kind = "NounGroup"}
```

Domain events

In order to recognize the domain events, we first have to survey in which events we are interested from the relevant documents according to domain knowledge. After finding the event, we then define the pattern based on the form of that event. For example, we are interested in the news about which company is merged by which company. So, we look for this kind of event from the relevant documents. Once we find this kind of event, such as “明基與西門子在 7 日下午宣佈合併”, we can define the pattern like the following format.

“{Company}{Conjunction}{Company}…{Date}?...{合併}”

All keyword encompassed by the brackets are already prepared in the previous processing, so we can easily use this pattern to define our JAPE rule as following. Here we take advantage of the gazetteer to define the relevant words in advance, including “合併”, “併購” and “收購”. With this convenient way, we do not use a large number relevant word (“合併”,”併購”,”收購”) rather than the concept (MERGE) to recognize the domain event.

```
Macro:MERGE
({Lookup.majorType == merge})
Rule: pattern1
//Priority:30
(
 {Company.kind == company}
 {Conjunction.kind == conjunction}
 {Company.kind == company}
 ({Token.kind == word})*
 ({Date.kind == date})?
 ({Token.kind == word})*
 (MERGE)
):goal
-->
 :goal.Pattern = {kind = "pattern"}
```

5. Implementation

In this section we first show the implementation of domain events extraction according to the design described previously. Then, we investigate the performance of an experiment we conducted. Finally, we describe the implementation of storing the extracted result in RDF store and the services provided by the store.

5.1 Information Extraction System

The GATE framework provides a pipelined environment in which each component is executed in a sequent way [8]. By using GATE, what we need to do first of all is to select the required components, as described in Section 4, from GATE and place then in proper order. This becomes the engine of the information extraction system. We have described the function of each component in Section 4. Here we focus on the description of the knowledge source that we provide for each component.

Chinese Tokenizer

The knowledge source for the tokenizer is a dictionary. The dictionary we employ is based on the Chinese Electronic Dictionary developed by the CKIP group [15]. Since it is in XML format, we employ the DOM (Document Object Model) parser [18] to retrieve the Chinese tokens within the dictionary and store in the format required by the tokenizer. The number of Chinese tokens we use in the tokenizer is about 140000.

Gazetteer

Based on the function of GATE Gazetteer, we can just focus on the definition of domain keywords. We list the financial domain categories used in our domain events extraction as shown in Fig. 6. The left panel lists all the categories we define and the right hand side shows the list of keyword of the selected category.

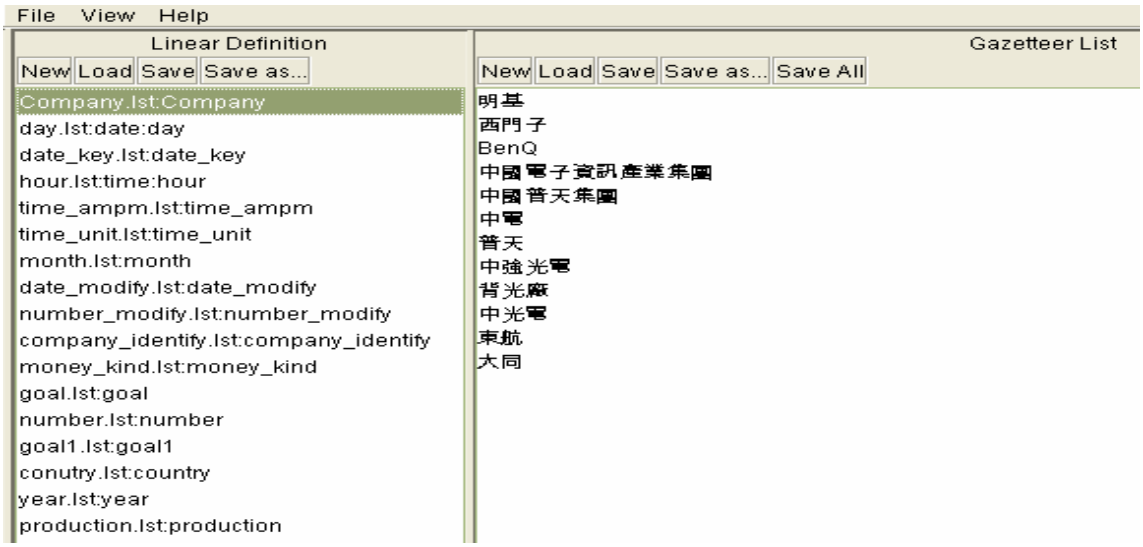


Figure 6: Definitions of the categories of domain keywords

In Fig. 7, we show the result as the document passes through the Gazetteer. We can see that several domain key words defined in our categories are found from the document. And, once a key word is found, it is attached with a `majorType` as its feature.

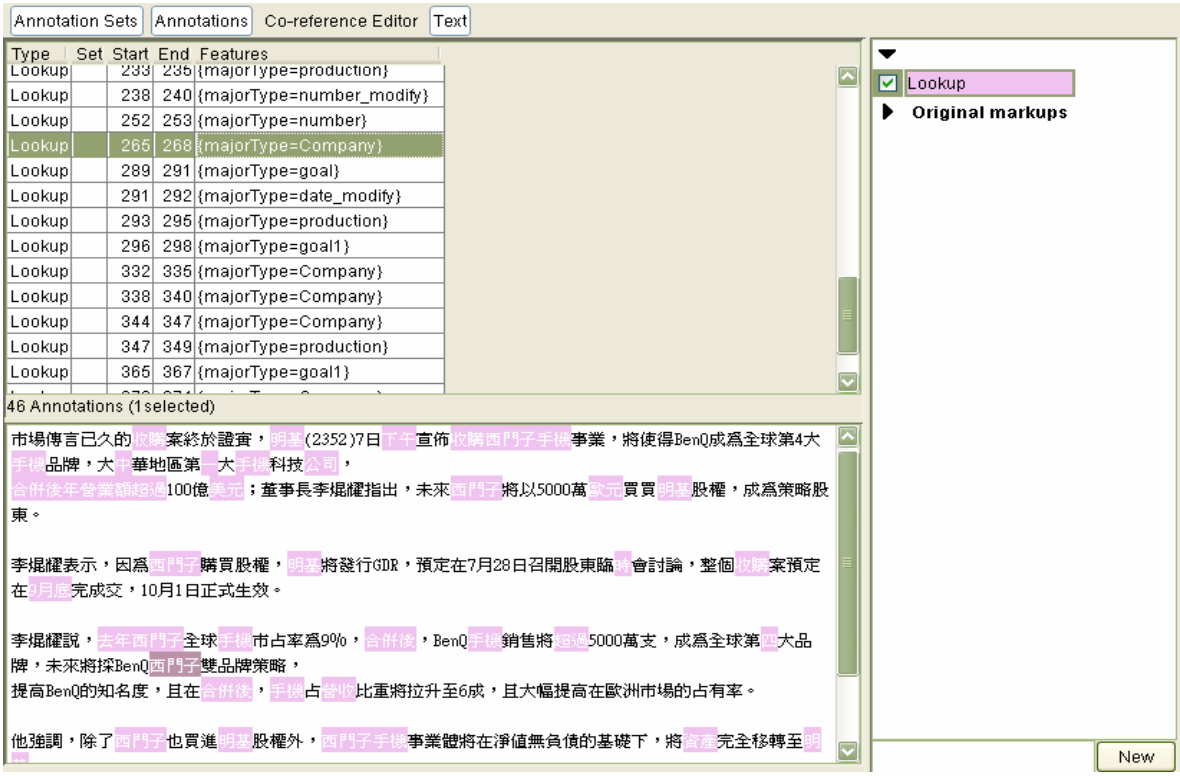


Figure7: Result produced by the Gazetteer

Sentence Splitter

Here we use the sentential mark , “。”, comma, “，” and question mark “?” in Chinese text to segment the text into sentences.

POS tagger

The GATE framework provides a Brill POS tagger [19]. In this paper we develop the rule base for the tagger by using the training algorithm of the Brill tagger [5]. We employ the Chinese balanced corpus developed by the CKIP [16] for the training and testing purposes. In our implementation, when training contextual rules on 550000 words with threshold 10, an accuracy of 95.43% is achieved. When we reduce the number of words to 55000 words, the accuracy is dropped to 94.46%. Finally, when we use only 16000 words to train the contextual rules, the accuracy is just 93.48%. The comparison is as follows.

Number of words	threshold	Accuracy (%)	Number of errors	Rule count
550000 words	10	95.43%	25469	341
55000 words	10	94.46%	3033	43
16000 words	10	93.48%	1063	8

Based on this experiment, we figure out the size of the training data has the direct impact on the accuracy. The more training data the learner use, the more contextual rules can be obtained, so, the more errors can be corrected. Next we apply these rules to the test data and see how many errors can be corrected. The test data we used here is 20000 words. The result is as follows.

Number of words	threshold	Accuracy (%)	Number of errors	Rule count
Initial state		93.7%	1280	No rule
550000 words	10	95.6%	894	341
55000 words	10	94.41%	1136	43
16000 words	10	93.76%	1268	8

Another reason which is related to the accuracy is the threshold. In the following experiment, we find that the less the threshold, the more precise the accuracy.

Brill Tagger	threshold	Accuracy (%)	Error number	Rule count
		93.34%	36797	No rule
	17	95.14%	27111	215
	14	95.30%	26254	274
	12	95.35%	25925	302
	10	95.43%	25469	341

The result of using the POS tagger is shown in Fig.8.

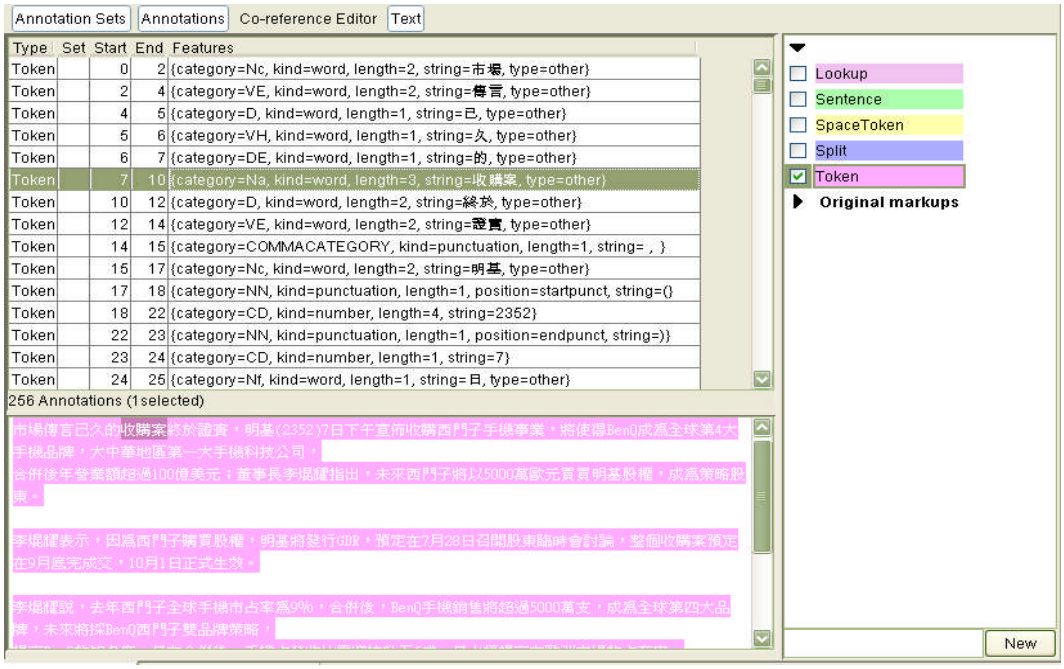


Figure: 8 Result of POS tagger

Domain events extraction

As described in Section 4, we can build up JAPE rules to extract the complex words, basic phrases and domain events from the output produced by the preceding components. The main task here is on the definition of JAPE rules. As they have been described in Section 4, here we only show the result of the implementation. Since the complex words and basic phrases are used to contribute the extraction of domain events. We therefore skip the smaller units while show the result of domain event. For example, the pattern "{Company}{Date}{收購}{Company}{Product}?{Detail}?" can be used to extract the event "明基(2352)7 日下午宣佈收購西門子手機事業" as shown in Fig. 9.

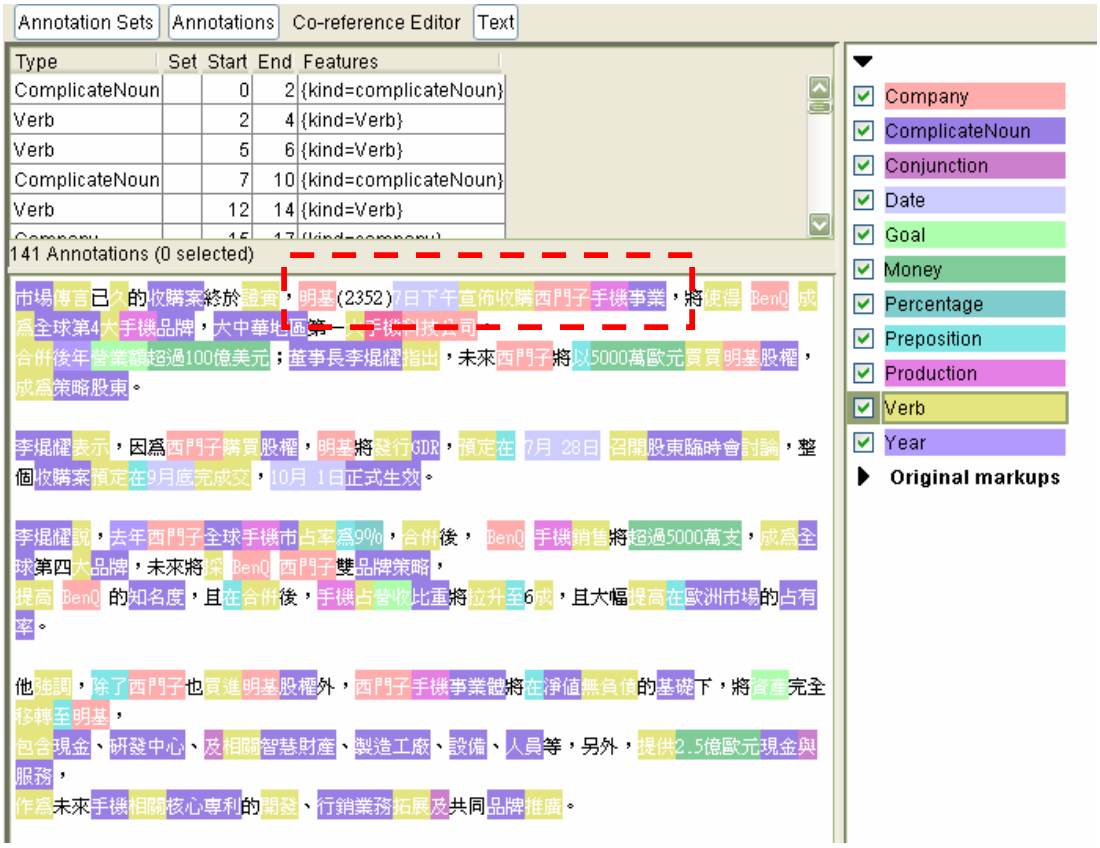


Figure 9: Result of extracting domain events

5.2 Evaluation

We first take one hundred financial news returned from the search engine, google, as our test data and use dozens of JAPE rules to recognize the specific domain event, after the processing of domain events matching, we then calculate the precision rate and recall rate of our system. We first manually extract the target events within these financial news and we obtain 120 events of interest. After the domain event matching, it returns 25 results, among which 22 are correct. So the precision rate is 88%, and the recall rate is 18%. The precision rate shows a promising result of using the information extraction system; however, the recall rate we obtain obviously needs further investigation as follows.

- First of all, the scope of JAPE rule recognition we use here is between the symbols, such as "，", "。", "?", *etc*. That is, the JAPE rule would not be able to extract events that cross the boundary of sentence symbol. for example, "中國東方航空以九億八千六百萬元人民幣，向母公司收購西北航空公司的業務". To solve this problem, we can divide our domain events recognition into several parts. For example, we can first process the recognition of one sentence, after all one sentence is processing, and we then process the recognition of two sentences, and so on.
- Second, using Part-of-speech identifier to recognize the domain events may result in error recognition. When we use the part-of-speech to recognize the domain event, some words which are irrelevant but belong to that part-of-speech may be recognized as the domain event. This will produce error domain events.
- Finally, the JAPE rules we use here is not sufficient to cover all the situations. So, if we

encounter some specific domain event that our JAPE rules do not define, we will lose it. To solve this problem, we further need much investigation into the expansion of the JAPE rules in order to improve the recall rate.

5.3 RDF store and the Services

The RDF store is used to store the RDF we obtained from the result information extraction system. In this paper, we employ an RDF store from the open source, Sesame [9]. Sesame is an ontology-based RDF store. It accepts RDF files as input and stores the input files in its indexing system. In the front end, it provides a number of application program interfaces for service programs to access the content of the RDF contents. In brief, Sesame play the functions of the ontology-based store and the front end service as shown in Fig. 4 in Section 3. The extracted domain contents are converted into RDF format and then the result can be imported into Sesame using its API. After the extracted contents are stored in the RDF store of Sesame, we then can use the query and explore service interfaces to access the content in Sesame.

Sesame provides SeRQL as its query language. For detailed description, please refer to the user guide of Sesame [20]. We give a brief description here. For example, if we want to query an RDF graph for persons who work for companies that are IT companies, the query for this information can be in the RDF graph in Fig. 10.

Figure 10: Example of query graph for SeRQL

The SeRQL path notation for the above RDF is written as below.

```
{Person} ex:worksFor {Company} rdf:type {ITCompany}
```

The parts enclosed by brackets represent the nodes in the RDF graph, the parts between these nodes represent the arcs in the graph. The SeRQL query language supports two query concepts. The first type of queries are called “select query”, the second type of queries are called “construct query”. Here we just use the select query because the form of select query is like the form of SQL, it is easy for us to use it to construct the conceptual search interface. The select clause determines what is done with the results that are found. It can specify which variable values should be returned and in what order. The from clause is optional and always contains path expressions. It defines the paths in an RDF graph that are relevant to the query. The where clause is optional and can contain additional constraints on the values represented in the path expression.

With the select query, we can extract whatever information we want to know. For example, in our financial domain, we may want to know which company is merged by which company. So we can write a simple query sentence as following to extract this information.

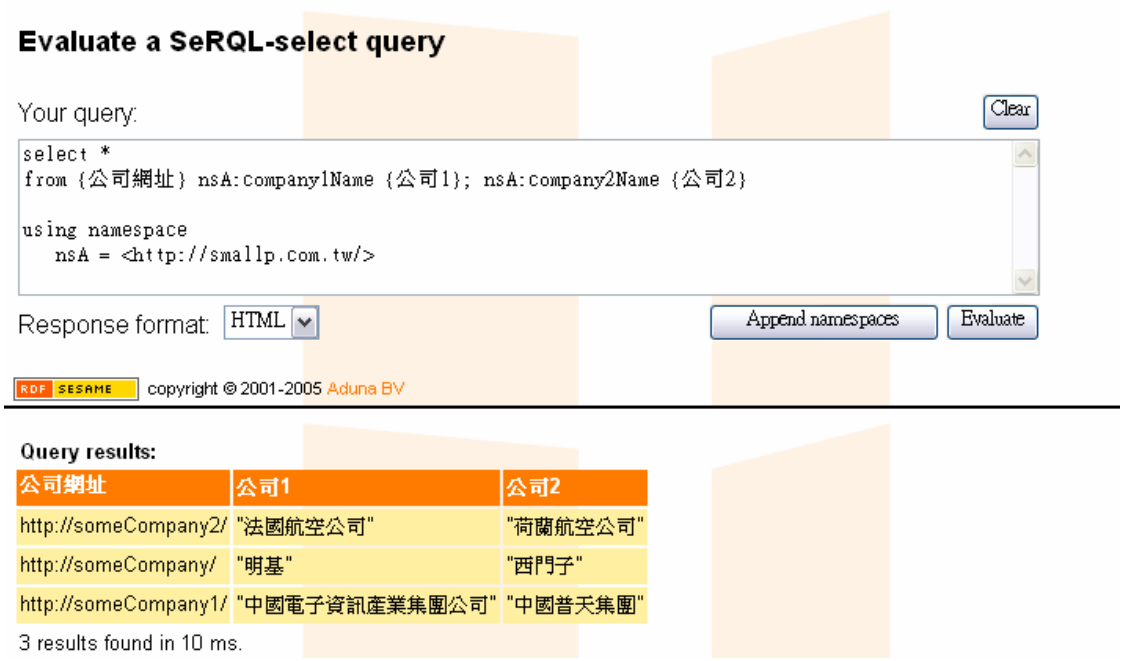


Figure 10: Example of using SeRQL of Sesame

Explore Service

The explore service can do the exploration according to the concept, that is, Class or Property. There are two methods to do the exploration. This first method is to enter the URI ourselves according to which concept you want to explore. The second method is to use the information provided by the Sesame. We can just click the URI represented in the following, and it will explore that URI for us. As shown in Fig. 11 is the result when we want to explore the property `http://protege.stanford.edu/rdfname`. We can find that when this property serves as a predicate, we can get all names of people existing in our RDF data.

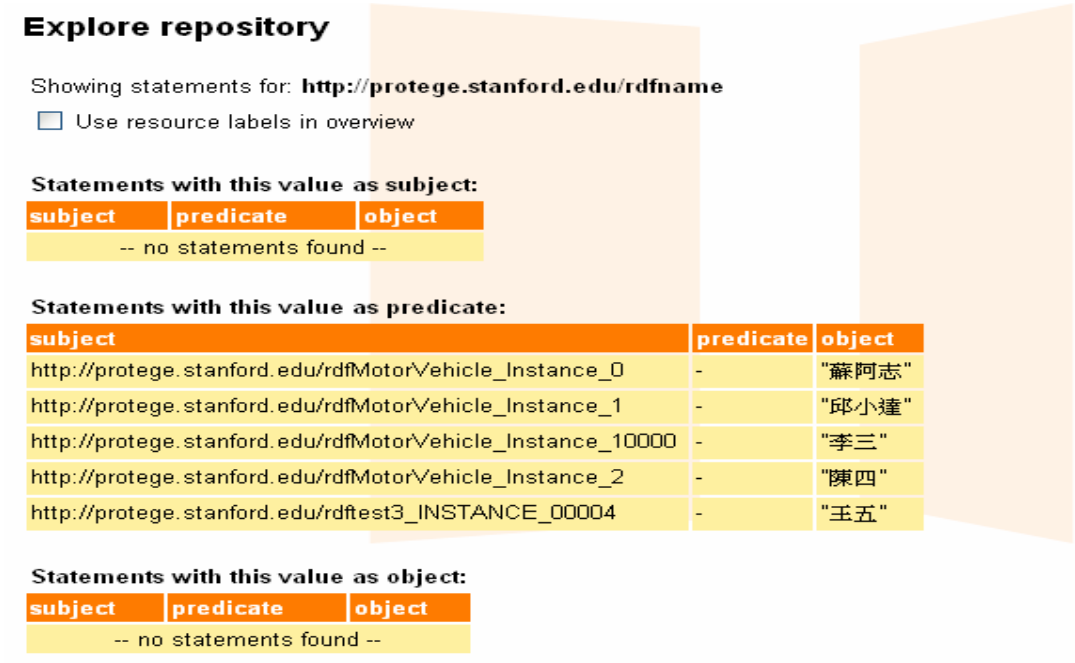


Figure 11: Example of exploring the content in Sesame

6. Conclusions

We have carried out automatic creation of metadata for the Semantic Web using the information extraction technology. The extracted information is stored in an RDF server and can be accessed through the service interfaces of the RDF server. In this paper we successfully use the components of the GATE software architecture for NLP as the processing engine to build up the system. Thus we focus our attention on the language knowledge sources required for the components. The reuse of the software components really shortens the creation of the prototype of the system. Since the system is developed mainly using the finite state machine technology, we obtain excellent system performance. We have tested the system on Chinese financial news. The result shows that most of the extracted domain events can be correctly identified. However, we still need do further investigation obtain wider coverage of extraction. In particular, we need to pay more attention on the creation of extraction rules for domain events.

Acknowledgements

We would to give our thanks to CKIP, Academia Sinica for their support of the balanced corpus.

References

[1] W3C, Semantic Web, http://www.w3.org/2001/sw/

[2] W3C, RDF Primer, W3C Recommendation 10 February 2004, http://www.w3.org/TR/rdf-primer/

[3] David Martin, *et al.*, Bringing Semantics to Web Services: The OWL-S Approach, *Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)*, 2004, San Diego, California, USA.

[4] Jose Kahan, Marja-Riitta Koivunen, Eric Prud'Hommeaux, and Ralph R. Swick, Annotea: An Open RDF Infrastructure for Shared Web Annotations**,** in *Proc. of the WWW10 International Conference*, Hong Kong, May 2001.

[5] Daniel Jurafsky and James H. Martin, *Speech and Language Processing*, Prentice Hall, 2000.

[6] Douglas E. Appelt and David Israel, Introduction to Information Extraction Technology, IJCAI-99 Tutorial, 1999, Stockholm, Sweden.

[7] Jerry R. Hobbs and David Israel, FASTUS: An Information Extraction System, http://www.ai.sri.com/~israel/fastus_brief2.pdf

[8] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.

[9] Sesame, http://www.openrdf.org/

[10]W3C, OWL: Web Ontology Language Reference. W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-ref/

[11]W3C. RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-syntax-grammar/

[12] R. Benjamins, D. Fensel, S. Decker, and Gomez-Perez. KA2: building ontologies for the Internet:

a mid term report. *International Journal of Human Computer Studies*. pp. 687-712. 1999.

[13] J. Broekstra, A. Kampman and F. van Harmelen. Sesame: a generic architecture for storing and querying RDF and RDF Schem. In J. Davies, D. Fensel and F. van Harmelen (eds.) *Toward the Semantic Web*: *Ontology-based Knowledge Management*. Wiley. 2003.

[14] T. Berners-Lee, R. Fielding and L. Masinter, Uniform Resource Identifiers (URI): Generic Syntax, IETF RFC: 2396, 1998, http://www.ietf.org/rfc/rfc2396.txt?number=2396

[15] The CKIP Group, Chinese Electionic Dictionary, http://www.aclclp.org.tw/use_ced_c.php

[16] The CKIP Group, Academia Sinica Balanced Corpus, http://www.aclclp.org.tw/use_asbc.php

[17] Charles N. Li and Sandra A. Thompson, 1981. *Chinese – A Functional Reference Grammar*, University of California Press.

[18] The Apache XML Project, Xerces Java Parser 1.4.4 Release, http://xml.apache.org/xerces-j/

[19] M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, Hong Kong, October 2000.

[20] B.V. Aduna, User Guide for Sesame, http://openrdf.org/doc/sesame/users/userguide.html.

以概念分群爲基礎之新聞事件自動摘要

劉政璋
國立交通大學資訊科學系
gis92573@cis.nctu.edu.tw

葉鎮源
國立交通大學資訊科學系
jyyeh@cis.nctu.edu.tw

柯皓仁
國立交通大學圖書館
claven@lib.nctu.edu.tw

楊維邦
國立交通大學資訊科學系
國立東華大學資訊管理系
wpyang@mail.ndhu.edu.tw

摘要. 新聞事件自動摘要乃針對敘述相似事件的多篇新聞文章編製摘要內容，其目的爲幫助讀者過濾資訊並快速瞭解事件的來龍去脈，以節省閱讀大量新聞文件的時間，主要的研究議題爲偵測不同新聞文章中相似及相異的內容，以達到過濾重複資訊的目的。本論文以概念分群(Concept Clustering)爲基礎，偵測新聞事件所要表達的語意，進而挑選涵蓋豐富語意的語句爲摘要。過程爲：1) 利用前後文關係(Context)及語意網路(Semantic Network)，描述概念詞彙；2) 使用 *K*-Means 對概念詞分群，期萃取更精確的概念，同時解決語意歧異的問題；3) 根據概念分群結果，並利用語句資訊量、語句位置及語句概念等特徵，計算每個語句之重要性，最後挑選重要性高的語句作爲摘要內容。實驗中使用 DUC 2003 (Document Understanding Conference) 所提供的新聞事件進行評估。評估結果於 ROUGE-1 指標比平均成績還好，於 ROUGE-L 指標接近最好的結果。

1. 前言

近年來，由於電腦科技的迅速發展及網際網路的推波助瀾，資訊陸續被數位化，以利於網路流傳。數位化的發展亦使得資訊量大幅增加，使用者在獲取資訊上不再受限於少數的流通管道，反而可以輕易取得大量資料。在這種現象下，困難的反而是如何過濾掉不需要及重複的資訊，使得使用者可以快速找到真正所需的資訊。爲解決前述困難，可利用摘要系統的精簡性及去重複性，減少使用者閱讀時間，幫助使用者於短時間判斷及取得重要資訊。

文件摘要技術的作法大致可分爲兩類。第一類使用專業領域知識(Domain Knowledge)分析文章中的人、事、時、地、物等要素，第二類則是以統計分析(Statistical Analysis)方法直接從原文判斷語句重要性。使用專業領域知識來達成摘要系統可有效抽取出文件內的主題，但是需要事前由領域專家介入建立領域知識，包括語言知識、文件主題背景知識等，而自動化建立領域知識的方法目前還很難突破。本論文以資訊擷取(Information Retrieval)技術爲基礎，導入語意網路(Semantic Network)，同時改良原文抽取摘要語句的方式，提出一套描述概念(Concept)且能分辨語意(Semantics)的概念偵測方法。利用不同的概念，便可找出新聞事件中具豐富概念及語意的語句，達到產生符合原文主題摘要的目的。

本文中提出以概念分群(Concept Clustering)抽取新聞事件所提及的主題(Topic)及語意，並結合傳統特徵選取法(Feature Selection)計算語句的重要性及語意涵蓋度，藉此作爲挑選摘要語句的參考依據。以下簡單說明本文所提之摘要方法的流程：1) 利用前後文關係(Context)及語意網路

(Semantic Network)，描述概念詞彙；2) 使用分群法(本文採用 *K*-Means [11])對概念詞分群，以萃取更精確的概念，同時解決語意歧異的問題；3) 根據概念分群結果，利用語句資訊量、語句位置及語句概念等特徵，計算每個語句的重要性，最後挑選重要性高的語句作爲摘要內容。

本文共分成六節。第二節介紹與多文件自動摘要相關的研究；第三節介紹結合前後文與語意網路的概念描述法，並說明如何進行分群偵測及抽取概念；第四節針對先前選出的概念找出對應的語句，並以數個特徵值進行語句的權重分析；第五節說明實驗結果的分析討論，以驗證本文所提方法的可行性；最後一節是結論與未來可繼續發展的方向。

2. 相關研究工作

本節介紹幾種多文件摘要技術。MEAD [18]接受分群過後的文件集[1]，併考量以下三個特徵：1) 語句與群中心(Centroid)的相似度；2) 語句於文件中的位置，通常出現於文件首句的語句，可加重其重要性；3) 語句與所屬文件之首句的相似度。MEAD以線性組合(Linear Combination)結合上述三種特徵，綜合評估語句重要程度。一般而言，MEAD使用的首句加重計分法，比較適用於新聞文章[2]；如果文件集是爲其他領域，例如技術類的文件，則首句加重計分法要再調整才合適。

McKeown et al. [17]認爲主題相關的文件集中，存在有許多不同的小主題(Theme)。他們的方法，分爲三個部分：1) 主題辨識(Theme Identification) [8]以語句爲最小單位，透過分群技術將文件中的主題抽取出來，同時辨識文件間相似及差異的部分；2) 資訊融合(Information Fusion) [3]將討論相關主題的段落融合，並去除重複的資訊；3) 摘要生成(Text Reformulation) 將所摘錄出來的重要字詞重新組合以產生流暢的摘要。他們主要考慮以下特徵以決定兩段落的相似度，進而利用分群法將找出主題，即相似段落的集合：

- Word co-occurrence：假如兩個段落有許多相似的字，則可視爲相似。
- Matching noun phrases：利用 LinkIt [26]判斷是否擁有互相關聯的名詞片語群組。
- WordNet synonyms：使用 WordNet [27]找出同義詞組。
- Common semantic classes for verb：判斷具有同一語意的動詞詞組。

接著利用 Information Fusion 的技術，從主題中萃取出具有代表性的詞組或片語。同時依照出現在文章中的次序，對片語排序。最後，藉由 FUF/SURGE [9]自然語言產生器生成完整語句。

MMR (Maximal Marginal Relevance) [4]適用於單文件摘要，其概念乃是對所挑選出與 Query 相關的語句重新排序，以符合具有最大相關度及最大差異度的特性。MMR-MD [10]延伸 MMR 的概念，可有效降低摘要中具有相同涵義的語句(即，減少重複性資訊)。MMR-MD 同時考慮到時間順序、專有名詞、對主題的相似度以及代名詞的 Penalty。其挑選段落的依據如下：

[1] MEAD 接受相關的文件集，以產生摘要。然此處所提及之相關文件集，實爲考慮 loosely-related documents。
[2] 此類文章通常於第一段第一句說明整篇文章的重點。因此，首句之重要性必須加重考慮。

$$MMR-MD \overset{def}{=} Arg \max_{P_{ij} \in R/S}[\lambda Sim1(P_{ij}, Q, C_{ij}) - (1-\lambda) \max_{P_{ij} \in S} Sim_2(P_{ij}, P_{nm}, C, S)]$$

公式 1: MMR-MD [10]

其中，$Sim_1(P_{ij},Q,C_{ij})$計算 P_{ij} 與 Q 的相似度，同時衡量與段落所在的文件群的相關度；
$Sim_2(P_{ij},P_{nm},C,S)$計算 P_{ij} 與 P_{nm} 的相似度，其中 P_{nm} 爲一以挑選出之段落。

MMR-MD 目的在於使摘要中的段落儘可能的相似於 Query，但其所選到的段落間要儘可能的不相似。由於與 Query 相似度高的段落，彼此之間的重複性可能也高，而與段落相似度稍低的段落，彼此之間的重複性可能也低，透過適當的 λ 值可以找到兼具主題但又不會有過多重複性的段落爲摘要。

Mani et al. [15]將文件表示成圖形(Graph)，其中，每個節點代表一個關鍵詞(Term)，節點與節點間用不同的關係連接起來，包含 1) 片語關係(PHRASE)；2) 形容詞關係(ADJ)；3) 同義關係(SAME)；4) 關聯關係(COREF)。首先，賦予每個節點一權重(Weight)，權重值初始爲該關鍵詞的 TF-IDF 值。接著，利用 Spreading Activation 演算法，透過節點間相連的連結權重變更節點的權重值，以找出與 Query 相關的節點。接著，比較兩兩文件圖形模型的相似度(Commonality)及差異性(Difference)。他們提出 FSD (Find Similarities and Differences)演算法，以找出兩圖形中相似或差異的節點。最後，透過分析 Similarities 及 Differences 集合中的關鍵詞，計算語句的重要性，並挑選重要的語句爲摘要結果。

3. 概念分群及抽取

本節說明如何以統計方法及分群技術由新聞文件集中推導出事件概念群[3]。首先介紹如何選取重要的概念詞，並利用概念詞的前後文(Context)與語意網路(Semantic Network)作爲其描述；接著說明利用分群法將概念詞分群，以導出新聞事件中的主題。

圖 1 說明本文所提方法之架構。步驟一爲前置處理；步驟二挑選具有代表性的名詞及名詞片語當作候選的概念詞；步驟三根據候選概念詞之前後文及事先建立的語意網路，描述該候選概念詞，以得到一向量表示式；步驟四針對候選概念詞作分群，可得到概念相似之概念群；步驟五依據語句中關鍵詞的資訊，將語句對應到概念群中，得到語句與概念群的關連；步驟六根據語句與概念群的關聯，同時考量文章結構的關係，計算 3 個與語句相關及 2 個與概念群相關的特徵值；步驟七則依據步驟六計算之特徵值，以線性組合的方式，計算位於同一概念群中語句的重要性，該重要性可作爲摘要語句挑選的依據。

[3] 概念爲單一或多個字詞所組成的集合。此集合可視爲一個概念性的描述，並定義該概念的範圍。透過此集合，可作爲系統理解概念語意的媒介。

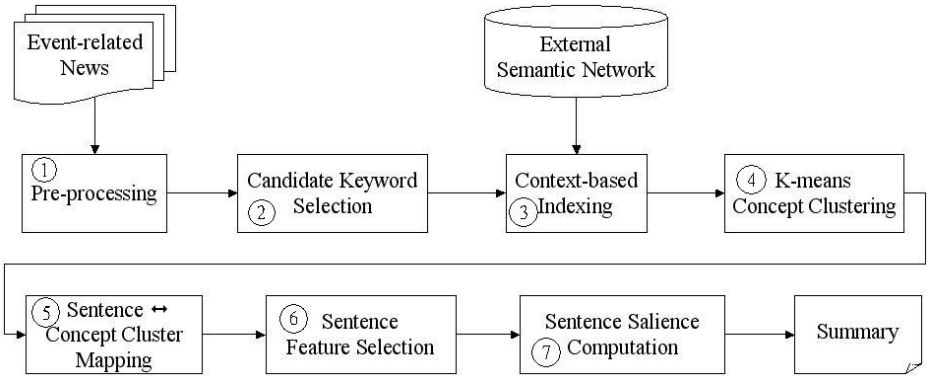


圖 1: 系統架構圖

3.1. 結合前後文與語意網路的概念描述法

首先，進行前置處理(Preprocessing)，其作用爲避免雜訊干擾，降低統計數據的參考性。此步驟包含斷詞切字(Tokenization)、詞性標記(Part-Of-Speech)、詞幹還原(Stemming)、小寫化(Lowercasing)、刪除停用字(Stopword Removing)及片語化(Chunking)等。本文中，斷詞切字及詞性標記採用 NLP Processor [20]；詞幹還原採用 Porter 演算法[24]；片語化則利用統計方法計算詞性組合機率，以辨別是否爲可組合片語；停用字的部份針對 DUC 2003 所提供的文件集設計，共有 309 個字，其中絕大多數爲介係詞、指代詞、連詞及助詞。

前置處理後僅保留名詞(Noun)及名詞片語(Noun Phrase)作爲可能的候選概念(Concept Candidate)，原因乃是名詞比其他詞性含有更多語意[1] [13]。本論文同時計算每個候選詞的 *tf-idf* 值[28]，進一步過濾不具代表性的字詞。最後，再由概念候選詞中挑選一般名詞、複數名詞、專有名詞、複數專有名詞等字詞，作爲最後所保留的概念候選詞集合。

接著說明如何導出概念候選詞的表示法。[2]提到絕大多數描述同一事件所伴隨出現的字詞，其語意皆很相似。[5]亦提到除考慮單一字詞的重要性外，更不能忽略出現在重要詞彙前後文的影響力；例如，condemn(譴責)及 intensively(強烈)經常一起出現，此兩關鍵詞可用來描述彼此。基於這個想法，本研究利用候選概念詞的前後文描述該字詞。作法上以出現在候選概念詞前後分別爲 *N* 及 *M* 個字作爲描述字彙集合，同時限制挑選的範圍爲一個完整的語句。在 *N* 與 *M* 的設定上，由於在[5]中提及人類的短暫記憶通常爲 7±2 個字詞，因此，實作上設定 *N* 及 *M* 各爲 5。取最小值最主要是要讓前後文的涵蓋範圍小一點，使相鄰的不同概念在描述的內容不致於有過多重複，以免影響到概念分群的結果。

表 1 爲下例中「the U.S. Embassy」的描述法，以 several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday 作爲索引詞(Indexing Description)。此描述概念的方式是希望利用前後文的關係，讓概念的語意更加明顯，以方便進行分群的時候能更精確計算兩兩概念的相似度。

BONN, Germany (AP) _ **German police** raided **several locations** near **Bonn** after receiving **word** of **a terrorist threat** against **the U.S. Embassy**, but **no evidence** of **a planned attack** was **found**, **officials** said **Wednesday**.
Source: d30005t\APW19981104.0772.xml

表 1: 以前後文描述後選概念範例[4]

Concept	Indexing Description	Length
the U.S. Embassy	several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	11

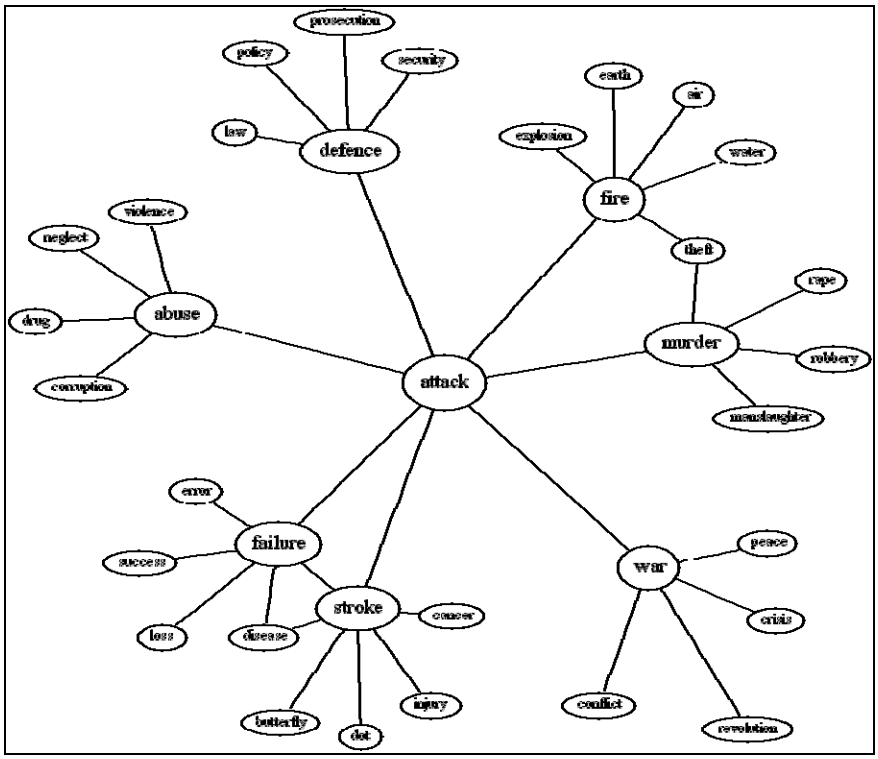


圖 2: 以 attack 爲中心之語意網路範例[12]

爲了加強描述字詞的語意，本研究在以前後文描述概念時亦加入語意網路。我們使用 Informap [12]建立新聞事件的語意網路，Infomap 以共現 (Co-occurrence) 的原則來判斷字與字之間語意的相關性。首先，依照字出現的頻率選訂出語意基本字(Content-Bearing Words)，再訂出一個可調式範圍(Window)，在這個範圍內的每一個字伴隨語意基本字一起出現的頻率，就把這些頻率定在共現矩陣。Infomap 利用共現矩陣，並使用奇異值分解(Singular Value Decomposition)降低字向量的維度， 最後計算兩兩詞間的相似度，並建立語意網路。圖 2 爲一以 attack 爲中心的語意網路範例，透過語意網路，可找出與 attack 語意相關的字詞。例如，直接相關的字詞有 defense、abuse、failure、stroke、war、murder 及 fire。

本論文提出兩種整合語意網路於前後文描述的方法。第一種方法是在前後文中只取與概念於語意網路中有關聯字詞，以美國大使館(the U.S. Embassy)此一概念爲例，於表 1 中，描述字詞的集合包含 Bonn、word、found、officials、Wednesday 等五個字彙，然而，透過語意網路的連接分

[4] 爲方便說明，此例及本文中所提及之 Indexing Description 皆列出未經過前置處理的關鍵詞。

析，發現「the U.S. Embassy」與上述五個字並沒有相關，因此在表 2 中便去除此五個字彙。由此可看出加入語意網路後，可消除多餘的字詞，使得描述概念的字詞更精確。

表 2: 加入語意網路方法 1 的範例

Concept	Indexing Description	Length
the U.S. Embassy	several locations, receiving, a terrorist threat, the U.S. Embassy, no evidence, a planned attack	6

第二種方法則是希望能夠突顯與概念語意有關的字詞，且不至於影響到原本描述字詞的組成，我們保留了原始用以描述概念的所有前後文字彙，但加重在語意網路上與概念相關的字彙。以美國大使館(the U.S. Embassy)此一概念爲例，由於 several locations, receiving, a terrorist threat, the U.S. Embassy, no evidence, a planned attack 爲語意網路中與 the U.S. Embassy 有直接相關的關鍵詞，因此在表 3 中將其 TF-IDF 值加上一常數 X，以達到加重權重的目的。

由上述兩例可知，方法一與方法二的差別爲，方法一根據語意網路刪除不重要的索引詞，而方法二則是保留所有的索引詞，但加重在語意網路中與概念相關詞之權重值。另外，方法一的描述雖然比較能夠貼近概念的語意，但其描述的字彙分佈比較散，且去除了不在語意網路內的字彙，使得描述的字彙數目少於原本的描述字彙，因此，方法一雖然描述精準但是會有描述字彙不足的情形，連帶會影響到之後在分群以及後來計算特徵的權重。

表 3: 加入語意網路方法 2 的範例

Concept	Indexing Description
the U.S. Embassy	several locations (5.1705+***X***), Bonn (5.1705), receiving (2.9733+***X***), word (5.1705), a terrorist threat (5.1705+***X***), the U.S. Embassy (2.9733+***X***), no evidence (5.1705+***X***), a planned attack (5.1705+***X***), found (3.5611), officials (4.4773), Wednesday (2.9733)

3.2. 利用分群技術抽取主題概念

本文中分群的對象，是經過 3.1 處理後之概念向量，分群方法則採用 *K*-Means [11]。考量新聞事件可再細分爲地點、對象、影響結果等特性，分群的結果可視爲文件中所提及的主題概念。

概念分群之後，便要將語句與概念群作連結，以期找到能夠代表每個語句的概念群(亦即，該語句所要表達的語意及相關主題)。本文提出兩種對應方法。第一，判斷語句中的字詞出現於哪個概念群中的字數最多，則歸類到該概念群。第二，判斷語句中的概念出現在哪個概念群中的字數最多，則歸類到該概念群。公式 2 爲語句對應到概念群的判斷依據。

簡單的說，第一種方法只單純判斷語句中有多少字出現在該概念群裡，概念群中原本只包含概念，然而本文亦嘗試把描述概念的字彙也加進概念群裡；這樣的作法是希望能夠增加語句對應到字詞的數量，避免一句話裡只有少數幾個字詞出現在概念群內，且對應的字詞數量越多，越容易判斷語句屬於哪一概念群。第二個方法判斷語句中的概念在哪個概念群中，由於概念是以編成向量的方式做 *K*-Means 分群，每個向量都可以找出與所屬概念群的相似度，也就是離中心點的距離。當語句裡有向量出現在概念群之中時，會以該向量離中心點的相似度當作該語句跟此概

念群的相似度。

(1)$SIM_{s,i}$ = Words Match
= sim (Match_Word, $Cluster_j$) / L_of_S

(2)$SIM_{s,i}$ = Concepts Match
= sim (Match_Vector, $Cluster_j$) / L_of_S

Match_Vector：concept vector included in this sentence
sim (Match_Word, $Cluster_j$)：number of word appear in $cluster_j$
sim (Match_Vector, $Cluster_j$)：distance between vector and centroid of $cluster_j$
L_of_S：length of the sentence

公式 2: 語句對應到概念群的方式

比較上述兩個方法，方法一的對應由於把描述概念的字彙也加入對應的條件，因此幾乎文件集內的每一個語句都可以找到對應到的概念群，造成了每一個概念群內的語句數量多，但是語句的語意可能不是與概念群的概念高度相似，造成此現象的原因可能爲只對應到描述概念的字彙，並不是對應到概念本身。方法 2 的對應則可以有效的過濾掉語意不符合概念群的語句，雖然對應後每個概念群包涵蓋的語句數目較少，雖然剩下的語句數量較少，但是再經由後面的特徵選取時，亦可有效提升選取適合摘要語句的效率。

4. 語句語意權重摘要

透過概念群及其中概念字詞與語句關鍵詞的相似關係，可將每個語句對應至語意相近的概念群。然而，位於同一概念群中的語句彼此語意近似，仍需要透過其他條件以判斷哪個語句最能代表該概念群。由語句特徵挑選重要語句的方法在很多研究中被提出來，藉由抽取不同的特徵，可以整合這些特徵以判斷語句的重要程度[16]。本文考量 3 個與語句相關及 2 個與概念群相關的特徵計算位於同一概念群中語句的重要性。

4.1. 語句相關特徵

- TF*IDF

  考慮語句中所有字詞的 TF*IDF 總和，並除以語句長度以正規化(Normalization)。

$$S_{tfidf} = \left(\sum_{i=n}^{m} TF \times IDF_i \right) \Big/ sentence\_length$$

- 語句於文件中出現的位置

  位於首句或尾句的語句通常具有關鍵性的語意資訊[5]。因此，當語句位於此位置時，則加重此語句的權重。

- 語句與所屬的概念群的相似度

  3.2 節提到兩個不同的對應方式，分別爲比較語句與概念群中共同出現的字彙數量及比較語

句所包含概念與概念群的相似度。本文針對此兩種對應方式，提出不同計算相似度的方法。

方法一：採用所對應到的字彙數目計算相似度，並除以語句長度作正規化，如公式 3。然而，由實驗中發現以這種方式來計算相似度，會發生有很多語句所對應到的字彙數量是一樣的情形，導致這個方法所計算出的權重不具有辨別性。

$$S_{sim} = match\_words_i / sentence\_length$$

match_words: 計算與概念群 *i* 內有多少字彙是一樣的
i: 語句所對應到的概念群 *i*
sentence_length: 語句的長度

公式 3: 相似度特徵計算方法 1

方法二：相似度的計算取決於向量對應的概念群與其中心點的距離，如公式 4。此方法可比較哪些語句比較接近該概念群的中心點。在多維度的向量中，使用歐基理得距離(Euclidean Distance)可更精確地找出哪些向量接近中心點，每個語句將可以更清楚地分出代表概念群的重要性。

$$SIM_{s,j} = \frac{1}{\sum_{i \subset s} \left(dis\tan ce(concept_i, cluster_j) \right) \times L_s}$$

concept: 語句有對應到概念群的概念
distance(concept, cluster): 取向量到概念群中心的距離
L_s: 語句長度

公式 4: 相似度特徵計算方法 2

4.2. 概念群相關特徵

- 概念群內含的概念多寡

包含越多的概念數量，表示原文件集提到的許多概念都在同一個群。當包含越多概念的群，其權重應該越高[5]。

- 概念群與中心點的距離

分群的結果，依照向量的分佈情形可以找出全部向量的中心點。每一個概念群中越靠近中心點的給予越高分。在中心點附近的概念群，越有可能涵蓋其他概念群的意思，在順序上應該要比其他遠離中心點的概念群要重要，亦能加強涵蓋性越大的概念群重要性。

4.3. 語句重要性

綜合上述的五個特徵可以得到一個權重總和，如公式 5。「C_{length}」爲概念群內的向量個數；「S_{tfidf}」爲語句內字彙的 TF*IDF 總和；「$C_{distance}$」爲語句所屬概念群距離全體向量質心的距離倒數；「$S_{location}$」爲語句所在位置；「S_{sim}」爲語句與所屬概念群的相似度。

$$sentence\ \_\ weight\ = \alpha(C_{length}) + \beta(S_{tfidf}) + \gamma(C_{distance}) + \theta\ (S_{location}) + \lambda\ (S_{sim})$$

公式 5: 計算語句權重總和公式

5. 實驗結果分析與評估

自動摘要的成效評估，可分爲直接(Intrinsic)與間接(Extrinsic)評估兩種方式[16]。直接評估需先定義出一組理想的摘要準則或答案，然後跟系統取出的摘要做比較。間接的方式則無須具備理想的摘要答案，而是評估自動摘要的結果在其他相關應用的成效。本摘要系統使用的測試文件集爲 DUC 2003 (Document Understanding Conferences 2003) [6]，文件內容是英文的新聞文件，分成 30 個新聞事件，每個事件中約有 10 篇相同主題的新聞，DUC 2003 並請不同的專家對同一類別作三篇摘要。評估方法是將系統自動產生的摘要與 DUC 2003 的專家所作出的摘要比較，每個事件的摘要以 100 字爲上限。效能評估採用 ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [22]，主要比較的項目爲 ROUGE-N、ROUGE-L，分別代表「自動摘要有多少 N 字詞與人工摘要一樣」及「自動摘要與人工摘要有多少字彙是出現在同一語句」。

本文所提出的方法有諸多變數需要最佳化以調整系統效能，表 4 列出所有可能的變數。

表 4: 實驗變數說明

步驟	變數	說明
描述概念	前後文長度	描述字彙的長度(即 *N* 與 *M*)
	加入語意網路方法	加入語意網路後描述概念字彙的方法
分群	分群數量	*K*-means 分群法需要先設定 *K* 值
	語句對應概念群方法	如何判斷語句所屬的概念群
語句重要性	權重比例調整	五個特徵以何種比例計算才能挑出最適當的語句

首先對權重比例進行最佳化，先選擇該變數的原因是希望之後的實驗都可有一個最佳的權重比例。調整的方法是先變換一個變數，同時固定其他四個。最後我們所調整出計算語句重要性之特徵權重比例 $\alpha:\beta:\gamma:\theta:\lambda$ 爲 1:5:5:1:8。

圖 3 調整的變數是概念向量的長度，也就是用來描述概念所用的前後文長度。評估的結果發現 ROUGE-1 最高情形出現在向量長度爲 11 之內，ROUGE-L 最高出現在向量長度爲 9 之內，與[5]提到的資料吻合。亦即，依照人類書寫以及閱讀習慣在看到某個字時，會記憶到前 7±2 個字彙，這區間的字也最爲相似，實驗結果也比其他超過區間的長度爲高。圖 4 爲調整分群數量變數的實作評估，由圖 4 得知在 5 群時 ROUGE-1 的分數最高，在 20 群的時候 ROUGE-L 的分數最高，因此之後會分別利用 5、20 作爲分群的數量。圖 5 爲調整語意網路關係權重的結果，在最好的情況下，加入語意網路可以比沒有加入語意網路改善約 7%。這個數據顯示出適當地加入語意網路可以有效地提升摘要品質。結果中也顯示分群數目在 5 群、20 群時互有高低，不過我們只取最高值，因此在這一結果中決定將語意網路加重之常數值 *X* 爲 1，分群數目則設定爲 5。

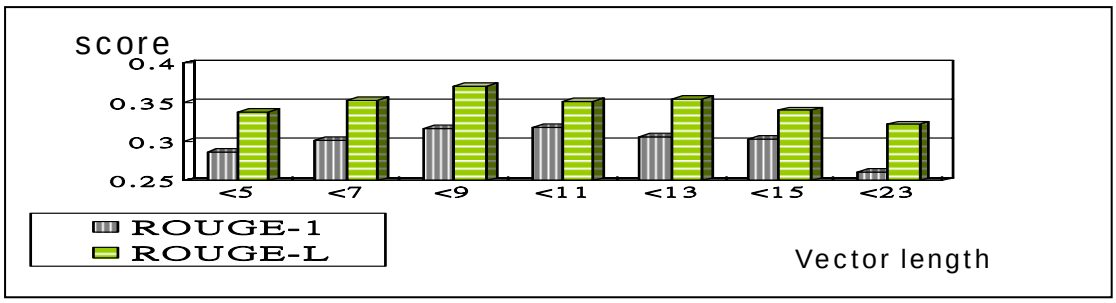


圖 3: 調整概念向量長度變數

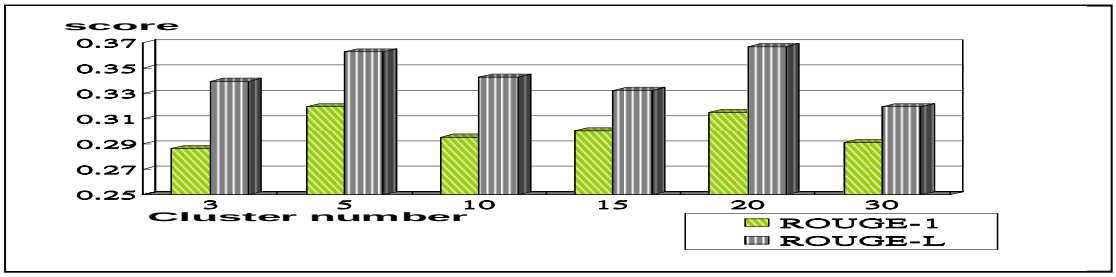


圖 4: 調整分群數量變數

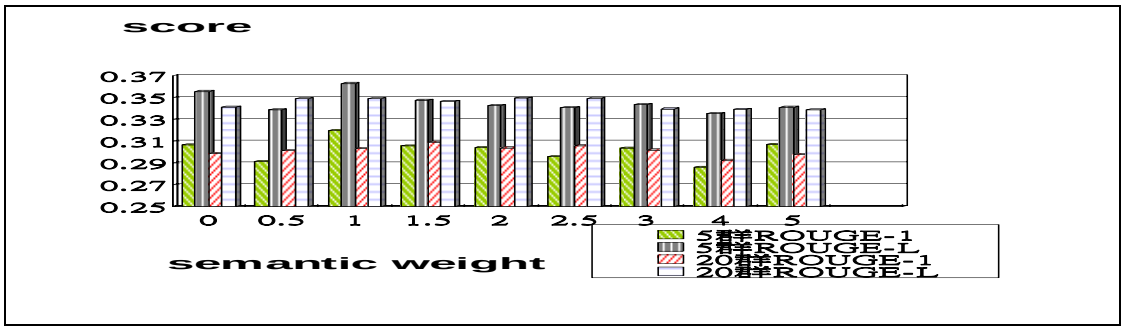


圖 5: 調整加入語意網路變數

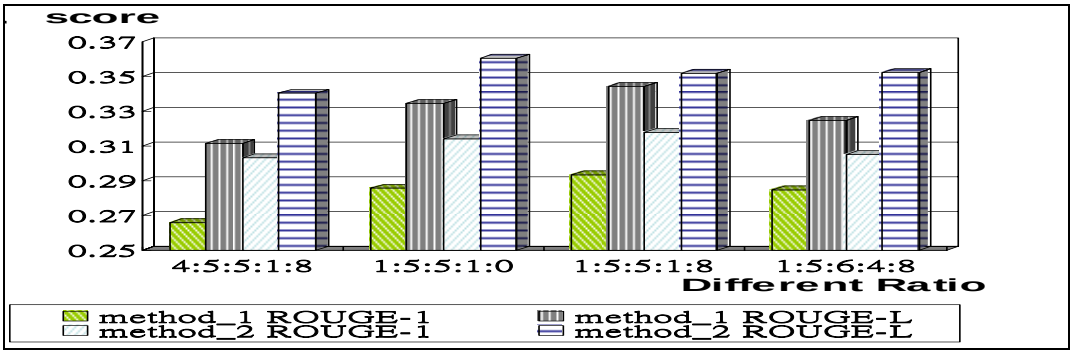


圖 6: 兩種加入語意網路方法的比較

圖 6 比較在 3.1 中提出的兩個加入語意網路的方法，方法一是只用有出現在語意網路上的字彙來描述概念，方法二是使用語意網路來決定是否要增加描述字彙的權重。可以發現方法二在各種變數的情況下都比方法一要好，最極端的情況下可以相差 19.6%。推估原因有二：第一，以方法一描述概念時，描述的字彙會比較少，因爲描述概念的字彙必須與概念共同出現在語意網路

中；第二，描述的字彙可能會離所要描述的概念距離過遠，在方法二中用來描述的字彙距離概念都在 4 個字之內，第二點在之前的實驗也說明了使用距離過遠的字彙來描述效果並不好。圖 7 中比較語句對應到概念群的兩個方法。方法二使用向量距離來決定語句該對應到哪個概念群，方法一是只比對出現的字彙數量來決定對應到哪個概念群。從圖 7 可以觀察出在不同的特徵比例下，使用方法二的效果較好。

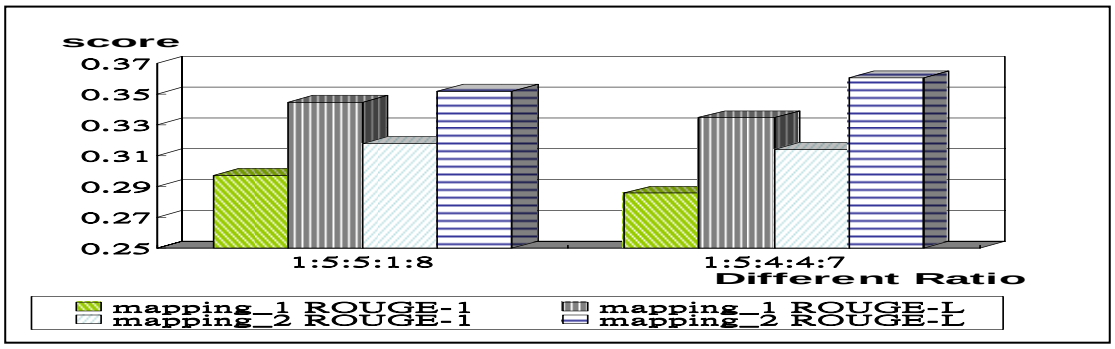


圖 7: 兩種語句對應概念群方法的比較

表 5 列出參加 DUC 2003 的其他系統及專家建立的摘要使用 ROUGE 進行評估的分數。「average of human summarizers」這列的數值是以人工針對三十個新聞事件做出的摘要，經過 ROUGE 評估工具所得到的成績，亦即評估專家間對於摘要內容看法的一致性。由此可知，以人工建立的摘要所得到的召回率(Recall)約在 40%左右，這個數值也代表不同專家對相同的文件集所摘要的內容及觀點不盡一樣。「best system」代表的是參加 DUC 2003 年摘要比賽的結果中最佳的數據，「worst system」則是代表比賽結果中最差的數據。「average of all DUC 2003 systems」則是所有參加 DUC 2003 系統的平均值。本論文所設計的系統在以 ROUGE-1 評比時位於平均以上，但是當考慮 ROUGE-L 時，本文所提的方法則非常接近於最佳的系統。

表 5: ROUGE 分數比較 (部分數值取自 DUC 2003 [6])

Summarizer	ROUGE-1	ROUGE-L
our system	0.32404	0.381149
average of human summarizers	0.4030	0.4202
best system in DUC 2003	0.36842	0.38668
average of all DUC2003 systems	0.31102	0.34652
worst system	0.23924	0.28194

6. 結論與未來研究方向

本文提出兩個新的想法，第一爲使用前後文資訊及語意網路描述隱藏在文件中的概念；第二爲對擷取出的概念進行語意分群，以解決語意歧異、語意重複的問題。同時，以概念分群爲基礎，並考量語句特徵，來計算語句的重要性，以挑選重要性高的語句作爲摘要內容。本研究所提出的技術有下列幾項特點：1) 以詞頻爲基礎，無須事前訓練；2) 透過共現矩陣建立語意網路，無須專家以人工建立；3) 分群可擷取重要主題概念，並對應語句與概念群關聯；4) 特徵選取包含一般性特徵(Surface Feature)以及加入概念分群的語意特徵。

未來，我們希望針對以下幾點作改進。首先，本論文以擷取語句爲基礎，然而測試文件集中大多爲長語句，以產生 100 字內的短摘要而言，僅能挑選到約 5 個語句，若考量語句壓縮，可納入更多語句，增加摘要內容的多元性。第二，*K*-Means 分群會因爲 *K* 值的不同而影響摘要品質，未來可考慮不同的分群法，如階層式分群(Hierarchical Clustering)。最後，片語化的過程，仍然會有相似的片語卻被當成不一樣的片語，例如 Saudi dissident Osama Bin Laden 與 Bin Laden 皆爲恐怖份子首腦賓拉登，但是沒有經過關聯及指代(Anaphora)處理會被誤認是不一樣的名詞，因此若加入指代關係處理，相信對於以名詞當作候選概念的擷取方式，可以增加準確度。

致謝

本研究由國科會計畫 92-2213-E-009-126-及部份由 93-2213-E-009-044-補助。

參考文獻

[1] R. Angheluta and R. De Busser and M.-F. Moens, "The Use of Topic Segmentation for Automatic Summarization," In *Proceedings of the ACL-2002 Post-Conference Workshop on Automatic Summarization*,2002.
[2] R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization," *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 1997 Page(s): 10 – 17.
[3] Barzilay, R., McKeown, K. R., & Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Conference on Association for Computational Linguistics* (*ACL'99*), College Park, Maryland, MD, USA (pp. 550-557).
[4] Carbonell, J., & Goldstein, J. (1999). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'98*), Melbourne, Australia (pp. 335-336).
[5] F. Chen and K. Han and G. Chen, "An Approach to Sentence-Selection-Based Text Summarization," *IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, (TENCON '02)* Volume1, Oct. 2002 Page(s):489- 493.
[6] DUC 2003 (Document Understanding Conferences). Available at http://www-nlpir.nist.gov/projects/duc/guidelines/2003.html.
[7] Elhadad, M. (1993). Using argumentation to control lexical choice: a functional unification implementation. *Ph.D. Thesis*, Department of Computer Science, Columbia University, New York, NY, USA.
[8] Eskin, E., Klavans, J., & Hatzivassiloglou, V. (1999). Detecting similarity by applying learning over indicators. In *Proceedings of the 37th Conference on Association for Computational Linguistics* (*ACL'99*), College Park, Maryland, MD, USA.
[9] FUF/SURGE. Available at http://sal.jyu.fi/Z/3/FUF-SURGE.html.
[10] Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings the ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA (pp. 40-48).
[11] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2000.
[12] Information Mapping Project. Available at http://infomap.stanford.edu.
[13] W. Lam and K. S. Ho, "FIDS: an intelligent financial Web news articles digest system**,"** *IEEE Transactions on Systems, Man and Cybernetics, Part A,* Volume 31, Issue 6, Nov. 2001 Page(s):753 – 762.
[14] C. S. Lee, Z. W. Jian and L. K. Huang, "A Fuzzy Ontology and Its Application to News Summarization," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics : Accepted for future publication* Volume PP, Issue 99, 2005 Page(s):859 – 880.
[15] Mani, I., & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2), 35-67.
[16] D. McDonald and H.C. Chen, "Using sentence-selection heuristics to rank text segment in TXTRACTOR," *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries*, Portland, Oregon, USA, 2002 Page(s): 28 – 35.

[17] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence* (*AAAI'99*), Orlando, FA, USA (pp. 453-460).
[18] MEAD. Available at http://tangra.si.umich.edu/clair/mead.
[19] U. Y. Nahm and R. J. Mooney, "Text Mining with Information Extraction," In *Proceedings of the AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.
[20] NLPprocess- Text Analysis Toolkit. Available at http://www.infogistics.com/textanalysis.html.
[21] Robin, J. (1994). Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design, implementation and evaluation. *Ph.D. Thesis*, Department of Computer Science, Columbia University, New York, NY, USA.
[22] ROUGE. Available at http://www.isi.edu/~cyl/ROUGE/.
[23] C. N. Silla Jr. and C. A. A. Kaestner and A. A.Freitas, "A Non-Linear Topic Detection Method for Text Summarization Using Wordnet," *Workshop of Technology Information Language Human (TIL'2003)*, 2003.
[24] The Porter Stemming Algorithm. Available at http://tartarus.org/~martin/PorterStemmer.
[25] L. Vanderwende and M. Banko and A. Menezes, "Event-Centric Summary Generation," In *Document Understanding Conference at HLT-NAACL*, Boston, MA, 2004.
[26] Wacholder, N. (1998). Simplex NPs clustered by head: a method for identifying significant topics in a document. In *Proceedings of Workshop on the Computational Treatment of Nominals*, *COLING-ACL*, Montreal, Canada (pp. 70-79).
[27] WordNet. Available at http://wordnet.princeton.edu/.
[28] 陳莉君(2003),"線上個人化參考文獻系統," 碩士論文, 國立交通大學資訊科學研究所, 新竹, 2003.
[29] 曾元顯, "中文手機新聞簡訊," 第十六屆自然語言與語音處理研討會, 台北, 2004 年 9 月 2-3 日, 頁 177-189.

中文句子相似度之計算與應用

鄭守益　梁婷
國立交通大學資訊科學系
{gis93540, tliang}@cis.nctu.edu.tw

摘要

近年來受惠於國內外各項語料庫資源的建置及網際網路上的大量中文語料，使電腦語文輔助教材的涵蓋層面日趨廣泛。因此如何產生大量且具高品質之輔助教材日益受到許多自然語言處理研究者的重視。有鑑于此，本論文提出以中文句子相似度爲基礎的研究與應用。相似度的計算乃考慮句子的組合及聚合性。我們實作此一應用，並提出解決未知詞的語意計算問題的方法。實驗結果顯示系統的檢索 MRR 值可以提升到 0.89 且每一檢索句皆可找到可堪用之例句。

1. 緒論

句子是可完整表達語意的基本單位[21]，也是語法的具體表現。因此，在語言學習中，學童若是學會了各種句型，也就學會了隱含在句型中的語法規則。藉由語言學家的歸納整理[14]，我們知道句子的結構並不是詞語的隨意組合，而是依照一定的「語法規則」。根據[15]，語法規則可進一步分爲「組合規則」及「聚合規則」。組合規則是指語法單位的橫向組合，例如，「我」、「買」、「書」這三個詞彙可以組合成「我買書」，但卻不能組合成「書買我」。當詞組合成結構之後，將具有語法意義，並使得整體結構的意義大於個別詞彙的意義總和，例如：「綠」、「葉」這兩個詞各自有其意義，但組合之後則形成了「綠」修飾「葉」的語法意義。

至於聚合規則是指在句子中，每個位置的語法單位都有其適合替換的詞語集合，例如，在「我買書」這個句子裡，「我」可以替換成「你」，但「買」卻不能替換成「花」。句子中的聚合替換規則可以視爲詞彙的語義替換問題，例如：語義同屬植物的「花」、「草」可以互相替換。

句型在學習語法時十分重要，因此融合語法變化的「句型練習」就成爲國小學童語言學習時的一個重要活動[18]。國語習作是現行國語課程的輔助教材，主要供國小學童課後練習使用，而習作的內容中幾乎每課都有「造句」、「照樣造句」、「替換語詞」等句型的練習 [16]。然而，由於習作中所提供的例句數量不多，再加上國小學童不論在閱讀的文章數量及習得的詞彙數量皆有所不足，因此，本研究之目的爲設計一有效率之句子相似度計算方法，以自動擷取國小學童句型練習中的「照樣造句」所需的範例例句。我們將句子相似度定義爲計算兩個句子之間的語法規則相似度，也就是說如果兩個句子的語法組合及聚合規則越相似，則其相似度越高。

句子相似度計算可依照語句的分析深度分成兩種方式。一種是基於向量空間模型的方法，把句子當成詞的線性序列，因此語句相似度衡量機制只能利用句子的表層資訊，即組成句子的詞的語義資訊。由於不加任何結構分析，這種方法在計算語句之間的相似度時無法考慮句子整體結構的相似性。例如在[8] [20]是以比對相同辭彙來計算相似度，對於句子之中，普遍存在的同義或近義詞之間的取代及比對，並沒能有效的解決。在[9]則提出搭配語義詞典檢索，並分配字義權重，以解決單純語義匹配的問題；但是，只使用語義詞典檢索來作爲相似度的計算依據，而沒有考慮到句子內部的結構和詞彙之間的相互關係，因此準確率並不理想。在[11]中提出使用編輯距

離的方法，其規定的操作模式，並不完全適用於整體句義相似的計算，也缺乏同義或近義詞替換的設計。另一方面，使用統計之語言模型的方法 [6]則需要建置大量的訓練語料。在[2][4]中結合了語義詞典檢索方法及傳統編輯距離方法[10]的優點，並利用 HowNet[5]和《同義詞詞林》[7]兩種語義辭典，以計算辭彙之間的語義距離，同時賦予不同編輯操作不同的權重，因此具有較好的輸出結果。由於其方法是基於同義詞典，來進行語義判定，因而衍生出未知詞及專有名詞語義判定的問題。另外。檢討其所使用的編輯操作權重，篩選候選句的計算方式，及評估輸出結果的方法，仍有改進的空間。

另一種方法則是對語句進行結構的句法與語義分析，並在分析結果的基礎上進行相似度計算，例如[17][19]先對被比較的兩個句子進行深層的句法分析找出依存關係，並在依存分析結果的基礎上進行相似度計算，但目前的語義依存句法分析器的準確率只有 86%，因此造成依存分析的結果並不準確，導致句子的核心詞無法正確判斷，因而產生了錯誤的計算結果。

在本論文中，我們提出以聚合規則相似度和組合規則相似度來設計並實作中文相似句子擷取系統。我們使用兩個句子中所含的詞彙之同義或近義詞來計算聚合語義的相似度，以及改良式編輯距離計算的方法，並設計新的權重配置比例、候選句篩選原則。在語義計算過程中，加入詞性標記資訊，以節省語義計算的次數，最後使用語義相似度矩陣，將所輸出的參數加以正規化，以取代人工評分的方法。

由於本論文所提之「句子相似度」可應用於學童句型練習中「照樣造句」所需之範例例句，操作方法即是按照原來句子的句型造句，例如：輸入「今天看到一幅畫」，輸出「昨天想到一個人」，因此只需要計算詞的線性序列相似度，而不需用深層的樹狀結構分析。此外，本研究將同時使用全域匹配(Global Alignment)及局部匹配(Local Alignment)的策略，求取兩句在全句和部分句段的結構相似度。

2. 聚合結構相似度

我們定義句子的聚合結構相似度爲兩個句子之間的詞語是否可使用同義或者近義詞替代。例如：「我愛你」與「你喜歡哥哥」就是一對聚合規則相似的句子。本研究改良並採用，以語義爲基礎的編輯距離演算法，來計算句子之間的聚合規則相似度。重新考慮編輯操作代價，及使用上下文資訊以解決未知詞及專有名詞語義判定問題，並利用網路語料來改進因資料稀疏而無法有效進行詞義比對的計算問題。

2.1 語義相似度計算

一般的編輯距離指的是，從一個句子變爲另一個句子，所需要的最小編輯操作的步驟數。傳統的編輯操作共有「保留」、「插入」、「刪除」和「替換」四種。以下圖爲例：（ ：代表刪除；○ 代表保留）

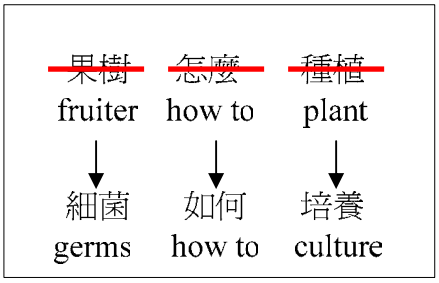


圖 1(a)：傳統編輯距離

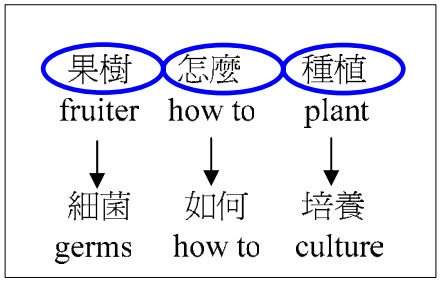


圖 1(b)：語義型編輯距離

從上面的計算過程可以看出，若僅使用編輯距離的方法，則計算出的語義距離和的實際情況將有許多差距。就語義而言，詞語之間的編輯操作代價應當各不相同。例如，上述兩句範例，雖然字面上的詞彙都不一樣，但若細細探究其中的涵義，可以發現其中的詞彙所扮演的文法角色及上位詞的語義內涵，有一定程度的相似。此外若在檢索目標的句子或短語的詞彙之中，加入具有修飾功能的詞彙，其語義也具有相似性。例如「果樹怎麼成功種植」與「細菌應如何快速培養」可視爲相似句。基於以上的觀點，本研究採用編輯距離的改進演算法[2]，即以辭彙爲基本的計算單位，同時以 HowNet 和《同義詞詞林》作爲語義距離的計算資源，以涵蓋更多的中文詞彙。

在《同義詞詞林》中，將詞彙按照語義關係的遠近親疏，賦予了一或多個語義代碼。按照樹狀的層次結構把所收錄的詞條分別歸類。同一層的詞語其關係有詞義相同或相近，或詞義在真實世界中有很強的相關性。例如：「大豆」、「毛豆」和 「黃豆」在同一層；這些詞不同義，但相關。從樹狀結構來看，《同義詞詞林》有五層結構，越靠近根節點，語義的概念越抽象。具體的詞彙，只分佈在節點末端。利用《同義詞詞林》來計算詞與詞之間的語義距離，可視爲單純的代數操作。但詞義的操作代價，應隨同義詞典的級距分歧度加大而增加，而非等量的增加。因此我們定義 X、Y 兩詞之間的詞義距離如下：

$$Dist(X,Y) = \underset{x\in X, y\in Y}{Min}\ dist(x,y) \qquad (1)$$

其中 x,y 爲分屬於 X，Y 兩詞之語義集合，根據《同義詞詞林》的結構，其計算公式定義如下：

$$dist(x,y) = \mathrm{Csim}(x,y) + (ld * \alpha) \qquad (2)$$

$$\mathrm{Csim}(x,y) = [((|n-5|*(4-n))/10) + 0.1] \qquad (3)$$

Csim(x,y)是指兩詞在同一棵語義結構樹之中，且兩詞的詞義從第 n 層結構開始有所不同；而 *ld* 爲該兩詞彙在個別的句子中的位置差距，α 爲系統定義的同義詞位移編輯代價。由於同義詞在距離相對詞語的位置超過三個以上時，其語義角色就已經產生變化，例如：「我對你很好，對不對？」句中的「對」這個詞，雖然，在句子中出現了兩次，但其語義已然不同。爲將詞語的位移控制在三以內，於是我們以計算同義詞林第一層語義代價除以三，將 α 設爲 0.3。

另外，我們認爲在詞語中進行插入或刪除等操作，將有可能影響並改變句子的整體意義及結構，因此這些操作將有較高的操作代價。我們定義爲：若進行刪除或插入操作，則操作代價應等同於兩詞不同義的代價，因此，我們以 n=0 代入公式（3）計算而設定爲 2.1。

HowNet 中同義詞的定義爲具有相同的英語譯文（W_E）和語義定義（DEF）的辭彙，其操作

代價設定爲 0.1。例如「愛」和「喜歡」，其簡化詞條如下：

表 1：HOWNET 同義詞舉例

NO	W_C	G_C	W_E	DEF
514	愛	V	love	FondOfl喜歡
89949	喜歡	V	love	FondOfl喜歡

在系統的計算過程中，先比對在 HowNet 中，兩詞是否爲同義詞，若是則兩詞之操作代價爲 0.1，若否則比對《同義詞詞林》並引用(1)作爲決定操作代價之依據。

2.2 未知詞詞義處裡

我們定義在 HowNet 及《同義詞詞林》中未收錄的詞彙稱爲未知詞。我們先在現有的語料庫中搜尋包含該未知詞的句子，並使用上下文資訊的相似度來協助判斷兩個詞語的相似程度，設定前後文的詞窗個數爲三個鄰近詞，並用共現值 I 來抽取相關度高的上下文詞組，其計算公式如下：

$$I(X_u, Z_w) = \log \frac{f(X_u, Z_w)/N}{(f(X_u)/N)(f(Z_w)/N)} \quad (4)$$

其中 N 表示語料詞數量，X_u 爲未知詞，Z_w 爲位於 X_u 前後的 3 個詞的任一詞彙，$f(X_u)$， $f(Z_w)$ 分別表示 X_u，Z_w 在語料庫中出現的次數，$f(X_u, Z_w)$ 表示詞 X_u，Z_w 一起出現的次數。經過實驗測試我們將共現值門檻值定爲 7，挑選出的關聯詞將作爲 X_u 的詞義集合。假設查詢句和目標句中分別有未知詞 X，Y，且他們的關聯詞分別是 $x_1, x_2 \cdots x_m$ 和 $y_1, y_2 \cdots y_n$ ，則同樣的我們可利用公式(1)來建立X和Y的相似矩陣M如下：

$$M(X,Y) = \begin{bmatrix} Dist(x_1, y_1), Dist(x_1, y_2) ..., Dist(x_1, y_n) \\ ..... \\ Dist(x_m, y_1), Dist(x_m, y_2) ..., Dist(x_m, y_n) \end{bmatrix} \quad (5)$$

再利用公式(6)計算出 X 和 Y 之間的語義相似度 S(X,Y)：

$$S(X, Y) = \frac{\sum_{i=1}^{m} Min[Dist(x_i, y_1), Dist(x_i, y_2), ...Dist(x_i, y_n)]}{m} \quad (6)$$

在未知詞詞義選取處理時，若無法獲得關係詞作爲語義相似度的判斷時，我們將使用網路語料作爲關係詞的查詢來源，本研究使用 Google 作爲查詢的搜索引擎，其步驟如下：

步驟 1: 查詢 Google 首頁，得知目前全部的待查網頁數量，作爲 N 值。

步驟 2: 使用未知詞 X_u 作爲搜索詞，查出 $f(X_u)$，並選取前 10 個摘要內容，作爲鄰近詞的抽取對象。

步驟 3: 將抽出的含有未知詞 X_u 的句子，進行斷詞，並進行鄰近詞抽取，詞窗個數爲三個鄰近詞。

步驟 4: 將未知詞 X_u 及鄰近詞一同作爲搜索詞，送進Google分別查出 $f(Z_w)$ 及 $f(X_u, Z_w)$。

步驟 5: 利用公式(4)，並篩選出超過門檻值的詞。

步驟 6: 將鄰近詞組代入語義相似度計算矩陣，計算關鍵詞對的語義相似度。

3. 組合結構相似度計算

如前所述，中文句子相似性的計算考量如下：

（1）任意句子中的詞組，其詞性角色的排列不可任意錯置，但可容許有限度的局部置換，例如：「我(N)吃(Vt)了(ASP)一(DET)碗(N)麵(N)」，不可寫成「我(N)一 (N)麵(N)吃(Vt)了(ASP) (DET)碗」，後者明顯不合文法；但「一(DET)碗(N)麵(N)是(V)我(N)吃(Vt)了(ASP)」，卻可以說的通。

（2）句子中的詞與詞之間，具有可插入適當空隙的特性，例如：名詞的前面應可容許插入相關的形容詞，「我(N)吃(Vt)了(ASP)一(DET)碗(N)麵(N)」，也可寫成「我(N)吃(Vt)了(ASP)一(DET)碗(N)很(Dfa)燙(VH)的(DE)麵(N)」。

由於全域匹配在計算上，將會考慮查詢句的詞性標記的整體的序列，因此可充分反映上述的第(2)項特點；而局部匹配，則只考慮由查詢句的詞性標記序列末端往前回溯的最佳子序列，因此可充分反映上述的第(1)項特點。我們將使用動態規劃演算法，分別計算句子與句子之間的，整體及局部相似度後，再依一定比例加權計算：

設A,B爲兩中文句之詞性標記序列，分別表示爲：A:{$a_1,a_2,\cdots a_n$};B:{$b_1,b_2,\cdots b_m$}，序列中之任一詞性標記 a_i和b_j，$a_i \subseteq A$，$b_j \subseteq B$，”-“爲序列中因不匹配而插入之間隙(gap)，$\sigma(a_i,b_j)$表示a_i和b_j比較時的分數值，我們定義爲：

$\sigma(a_i,b_j)$=2, for all $a_i=b_j$

$\sigma(a_i,b_j)$= -1, for all $a_i \neq b_j$

$\sigma(a_i,b_j)$= $\sigma(-,b_j)$=-1

我們利用 SMITH WATERMAN 所提出的全域相似匹配演算法[12]來找出匹配句，並以公式(7)計算出以詞性標記爲主的兩句相似度值，其中 l 爲兩序列比對時之最大長度。

$$GSim(A,B)=\frac{Score(A,B)}{n+m} \quad (7)$$

$$Score(A,B)=\sum_{i=1}^{l}\sigma(a_i,b_i) \quad (8)$$

另外一方面，我們利用改良式的 SMITH WATERMAN 算法[1]來找出局部相似的候選句。其計算匹配的路徑不需要到達矩陣的盡頭，如果某種匹配的分數值不會因爲增加匹配的數量而增加時，這種匹配就是最佳的。其相似度值爲：

$$LSim(A,B)=\frac{Score(A,B)}{n+m} \quad (9)$$

$$Score(A,B)=Max\{C[i,j]\} \quad (10)$$

其中 $Max\{C[i,j]\}$ 爲計算矩陣中分數最高的數值。

4. 混合式的中文句子相似度的計算應用系統

綜合上述所提的組合及聚合結構相似度計算，我們提出了一個混合式的中文句子相似度的計算應用系統(系統架構圖見圖 2)，在進行句義相似度計算時，主要分爲以下步驟：

步驟 1:利用[13]進行中文句斷詞並自動標注詞性標記。

步驟 2:同義詞擴展：爲了使候選句能更具有多樣性及提高系統的召回率，因此我們對斷詞之後的各個辭彙進行同義詞擴展。本系統使用 HowNet 語義詞典作爲詞擴展的資源。

步驟 3:候選句檢索：我們假設，如果一個候選句中所含的詞語，與查詢句的相同詞或同義詞越多，就越有可能是我們要擷取的相似句。因此，我們設定候選句的標準爲：候選句的詞數不能大於檢索句的 3 倍，而符合的詞數不得小於檢索句詞數的三分之一，並按照句子權重由大到小排序，選擇前 100 句作爲候選句。

步驟 5:句子聚合結構相似度計算

步驟 6:句子組合相似度計算

步驟 7:分別依各項不同的計算數值，取前 10 句候選句，作爲答案。

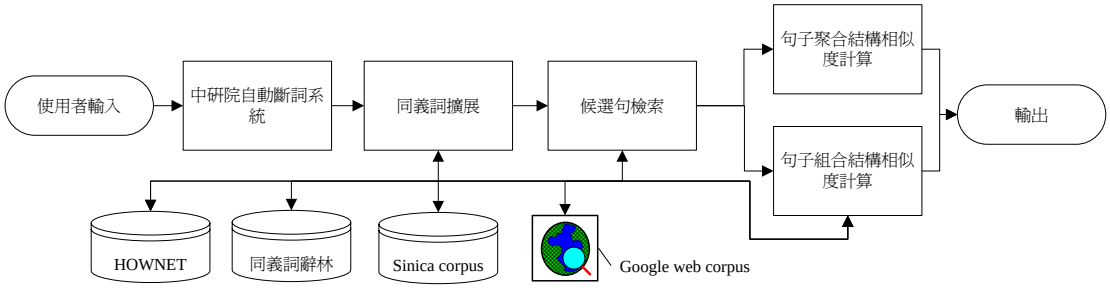


圖 2 系統架構圖

5. 實驗與分析

我們採用中央研究院平衡語料庫 3.0 版，作爲系統的候選句及查詢句的語料庫，其中包含了 500 萬已標記的中文語料。我們從中隨機選取包含 5 到 8 個詞彙的短句 100 個作爲查詢句。

本論文設計成四種不同的實驗做比較：

(1) BaseLine：以[2]中所提的詞彙作爲計算單位的動態規劃編輯演算法。

(2) M1：在詞義判斷過程中，利用語料庫的上下文資訊，來處理未知詞。

(3) M2：在語義判斷過程中，加入網路語料上下文資訊，來處理未知詞。

(4) M3：利用本論文所提的組合及聚合規則來計算相似度。

又依選擇候選句的指標不同，使用 MRR(Mean Reciprocal Rank)分別測試其對於選擇正確的候選句的影響：

(1) OC (Operation Cost): 使用原有的編輯距離作爲抽取候選句的標準。設 n,m 分別爲候選句及查詢句的長度：

$$OC = \sum_{i=1}^{n} \sum_{j=1}^{m} dist\ (x_i, y_j) \qquad (11)$$

(2) NOC (Normalized Operation Cost): 使用候選句及查詢句中，句子所含的詞數進行正

規化：

$$\mathrm{NOC} = \frac{OC}{Max(n,m)} \tag{12}$$

(3) SWR (Semantic Weight Ratio):傳統上多數句子相似度的評分標準都是以編輯距離操作代價作爲句子相似度的評分標準，但是這樣的分數會因爲句子的長度不同，而造成長句往往分數會高出許多。然而將原始的編輯操作代價進行正規化，亦無法避免因編輯距離的不同，而給予較客觀的相似度評估。因此，本論文乃設計在語義計算過程中，所產生的編輯操作代價，依照正負相關系數的門檻值，切分成正相關係數及負相關係數，再透過與原始編輯距離的計算，產生出詞語語義貢獻度SWR(Semantic Weight Ratio)。其計算方式如下：設S_q, S_t爲兩中文句詞彙序列，P爲所有編輯距離操作代價之分數總和，Q爲所有負相關係數總和[1]，則

$$\mathrm{SWR}(S_q, S_t) = \frac{P-Q}{P} \text{ ， } 0 \le \mathrm{SWR}(S_q, S_t) \le 1 \tag{13}$$

其值越接近 1 則表示S_q, S_t句中所含的相似詞語越多。

(4) PCRC (POS Construction Related Coefficient): 結合全域及局部的匹配相似度，作爲判斷候選句及查詢句之間的結構相似度，經實驗將兩項數值的比重設定如下：

$$\mathrm{PCRC} = (0.6 \times LASim(A,B)) + (0.4 \times GASim(A,B)) \tag{14}$$

(5) CSSS (Combine Semantic and Structure Similarity): 結合語義及語法結構，作爲抽取候選句的標準，因本系統主要將應用於國小學童的照樣造句的活動之上，因此將比較偏重於結構方面的相似度，因此將兩項數值的比重設定如下：

$$\mathrm{CSSS} = (0.4 \times \mathrm{SWR}) + (0.6 \times \mathrm{PCRC}) \tag{15}$$

以上的各項標準所篩選出的候選句集合，我們使用人工方式以 MRR 值來評定其效能。使用此值能測量出系統產生出第一個語義最相近的例句的平均名次。若第一個結果即爲最佳匹配，則分數爲 1，第二個匹配分數爲 0.5，第 n 個匹配分數爲 1/n，若無匹配的句子，則分數爲 0。最終的分數爲所有得分之和。另外我們還觀察各項實驗中，找不到例句的查詢句的數量變化，我們使用「NON」來表示其數值。

另外，我們在使用上下文資訊，進行語義相似度計算時，使用詞性標記資訊，以減少計算的雜訊。例如：在「張三站在椅子上」，「飛彈上有字」這兩句都有「上」(Ncd（位置詞）)。因此，在使用上下文資訊計算詞意相似度時，我們將不考慮下列詞性的詞：

Da（數量副詞）、Caa（對等連接詞）、Cbb（關聯連接詞）、Nep（指代定詞）、Neqa（數量定詞）、Nes（特指定詞）、Neu（數詞定詞）、Nf（量詞）、Ncd（位置詞）、Nd（時間詞）、Nh（代名詞）、P（介詞）、Cab（連接詞）、Cba（連接詞）、Neqb（後置數量定詞）、DE（的，之，得，地）、I（感嘆詞）、T（語助詞）、SHI（是）、V_2（有）

以下爲各個模組採用 CSSS 標準所產出的範例，查詢句爲：「世上還有癡心的人嗎？」

[1] 本研究設定之門檻值爲操作代價大於同義詞林第三層語義代價，也就是以n=2 代入公式(3)而設定爲 0.7。

表 2 查詢結果範例

查詢句	模組	排序	候選句	CSSS	MRR
世上還有癡心的人嗎？	M3	1	音樂真的有那麼深的殿堂嗎？	0.71	1
		2	你有足夠的耐性嗎？	0.68	*
		3	我還有追求幸福的權利嗎？	0.63	*
	M2	3	有這樣子的人啊？	0.58	*
		4	中國也有瓷器嗎？	0.55	0.25
		5	屈辱的生，英勇的活。	0.53	*
	M1	4	屈辱的生，英勇的活。	0.52	*
		5	中國也有瓷器嗎？	0.50	0.2
		6	唉唷，還有巧的呢！	0.49	*
	BaseLine	7	限電方式也有雙贏的？	0.45	*
		8	中國也有瓷器嗎？	0.44	0.125
		9	並且也發表您的看法嗎？	0.43	*

5.2 實驗結果與分析

圖 3 及圖 4 顯示了 MRR 值在四個實驗模組的分佈情況。從圖顯示，相對於其他的模組，M3(MRR 值平均皆大於 0.7)可以有效提昇相似候選句的選取。並且也不會因爲使用不同的篩選模組而降低候選句的品質。另一方面，實驗也顯示所設計設的 CSSS，其 MRR 值平均大於 0.68。相對於其他篩選標準，CSSS 可以控制候選句的品質，並且可以將相似句的 rank 值提升。由於 CSSS 的 MRR 值顯示了正相關，跟 PCRC 及 OC 比較起來，當同時考慮語義相似度時，它可以改善 MRR 值到達 0.89。

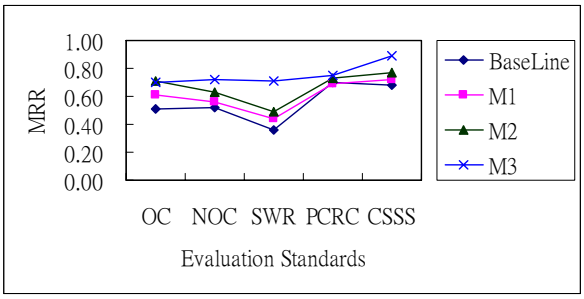


圖 3: 四個實驗模組的 MRR 值

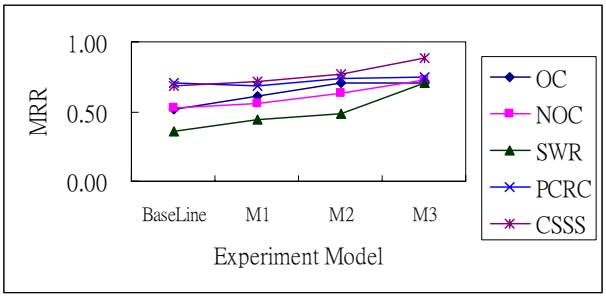


圖 4 不同系統之 MRR 值變化

另一方面，圖 5 及圖 6 顯示 NON 值，在使用不同的條件句篩選標準情況下，四個實驗模組的分佈情況。從圖顯示，M3 的 NON 平均值皆小於 2 (如果同時使用 PCRC 或 CSSS 則可以下降到 0)。這意味 M3(相對於其他模組而言，)可以更有效的抽取出相似的候選句。另一方面，PCRC 及 CSSS 的 NON 值平均小於 3.5，因此相對於其他模組而言， 在抽取候選句的時候如果同時考慮語義則可以將 NON 的值降到 0。

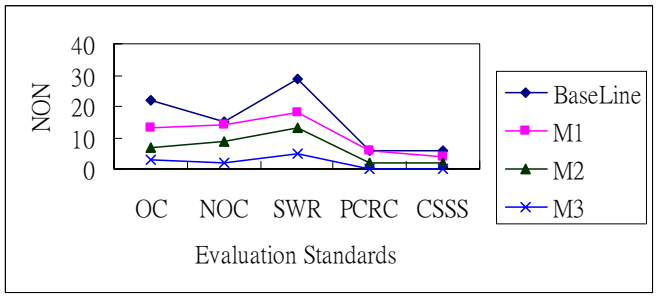


圖 5: 不同實驗指標之 NON 值變化。

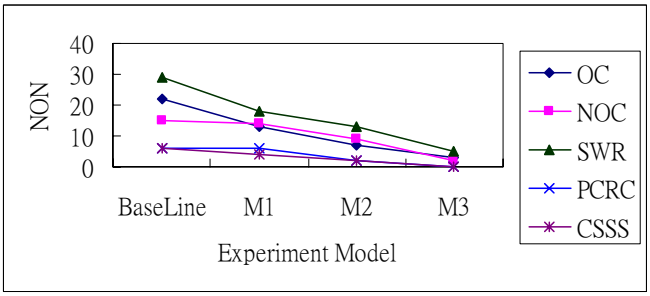


圖 6: 不同系統之 NON 值變化。

6 結語

在本論文中，我們提出新的中文句子相似度計算策略，並應用於中文習作中的範例句產生之自動化。此例句產生系統可自動從語料庫中抽取相似句以作爲學童練習造句時之參考。此系統主要有如下之改良：

(1) 改良語義計算所使用之編輯距離計算方式，加入限制同義詞或近義詞位移的操作代價，以解決詞語因重複出現，而造成語義權重判斷錯誤的問題。

(2) 使用上下文資訊之相似度，作爲判斷詞義相似的標準，以解決辭典未收錄詞彙的詞義

判斷問題。

(3) 爲了解決資訊稀疏的問題，在現有的資料庫無法提供有效判斷詞義的上下文資訊時，將採用 Web Corpus 來輔助。

(4) 使用詞性標記資訊協助判斷詞義，去除不含有效語義判斷成分的詞類，並減少相似度比對時的計算量。

(5) 使用全域相似度匹配及局部相似度匹配，並結合詞性標記，加權計算句子之間的組合結構相似度。

(6) 改良並設計新的句子相似度計算公式，結合句子的聚合及組合相似度，並可依照系統的需求，機動調整權重，以符合使用者的需求。

致謝

我們感謝中央研究院資訊科學所詞庫小組提供之線上斷詞系統 (http://ckipsvr.iis.sinica.edu.tw/)。

參考文獻

[1] Altschul, S.E., Gish, W.: Local alignment statistics, Vol. 266. Methods Enzymol (1996) 460-480

[2] Che, W. X., Liu, T., Qin, B., Li, S.: Similar Chinese Sentence Retrieval based on Improved Edit-Distance, Vol. 14(7). High Technology Letters (2004) 15-20

[3] Chen, K.J., Ma,W.Y.: "Unknown Word Extraction for Chinese Documents,"Proceedings of COLING 2002, pages 169-175

[4] Chatterjee, N.: A Statistical Approach for Similarity Measurement Between Sentences for EBMT, Proceedings of Symposium on Translation Support Systems. 2nd Indian (2001)

[5] Dong, Z. D., Dong, Q.: HowNet, http://www.keenage.com (1999)

[6] Li, S., Zhang, J., et al.: Semantic Computation in Chinese Question-Answering System. 2002, Journal of Computer Science and Technology, 17(6): 933

[7] Mei, J.J. et al.: TonYiCi CiLin－thesaurus of Chinese words (同義詞詞林), Shangwu Yinshuguan (商務印書局香港分館), Hong Kong (1984)

[8] Nirenburg, S.: Two Approaches of Matching in Example-Based Machine Translation, Proc. TMI-93, Kyoto, Japan, 1993

[9] Qin, B., Liu, T., Yang, W., Zheng, S., Li, S.: Chinese Question Answering System Based on Frequently Asked Questions, Journal of Harbin Institute of Technology May (2003)

[10] Ristad , E. S., Yianilo, P. N.: Learning string-edit distance. Vol. 20(5). IEEE PAMI (1998) 522

[11] Ristad, E. S., Yianilo, P. N.: Learning string-edit distance. 1998, IEEE PAMI, 20(5): 522

[12] Simith, T. F., Waterman, M. S.: Identification of Common Molecular subsequence, Vol. 147. Journal Mol. Biol. (1981) 195-197

[13] 中央研究院線上斷詞系統，http://ckipsvr.iis.sinica.edu.tw/

[14] 謝國平，《語言學概論》台北：三民書局，2002 年 頁 195

[15] 葉蜚聲，徐通鏘，「語言學綱要」，台北：書林，2001 年，頁 97-106。

[16] 蔡米凌，「國小三年級學童作文句型結構之分析研究—以嘉義地區爲例」，國立嘉義師範學院國民教育研究所碩士論文，1997 年。

[17] 穗志方，「語句相似度研究中的骨架依存分析法及其應用」，北京大學博士學位論文，1998 年。

[18] 陳玫秀，「學前兒童國與句型結構之分析研究」，國立師範大學特殊教育研究所碩士論文，1990 年。

[19] 李彬，劉挺，秦兵，李生，「基於語義依存的漢語句子相似度計算」，電腦應用研究，2003 年。

[20] 我國簡易刑事判決的製作輔助系統 (Decision support for criminal summary

judgment)，第七屆人工智慧與應用研討會論文集 (TAAI'02)，178-183。台中，台灣，15 November 2002 年。

[21] 胡百華，「華語的句法」，台北：阿爾泰，1984 年。

日本學生學習華語的聲調偏誤分析:以二字調爲例

張可家　　陳麗美
高雄師範大學華語文教學研究所　　國立成功大學外文系
visit@pchome.com.tw　　leemay@mail.ncku.edu.tw

摘要

外國人在學習華語的過程中，聲調是最感陌生和不易掌握的難點。朱川（1994）提出許多日本學生學華語遇到二字詞時，不論這兩個音節原調是什麼，一律誤讀爲「升降格」。何平〈1997〉認爲日本學生常常起音度掌握不好，在聽辨上聲（三聲）與去聲（四聲）、陽平（二聲）和上聲（三聲）、陽平（二聲）與去聲（四聲）方面非常困難。往往區別不開陽平和上聲，十分容易把上聲當作陽平。造成上述問題的原因是因爲受到母語音韻系統的影響。標準日語（東京話）的高低重音特性是呈現在各個音節的高低起伏，而且第一個音節和第二個音節高低一定相反。本文以語音分析軟體 PRAAT 分析聲調變化，並討論日本學生在學習華語和以華語爲母語者聲調的差異。研究結果發現，本研究中兩位日本學生在唸華語二字調時，第一個音節爲第二聲和第三聲的錯誤率最高，而且聲調偏誤率集中在 2─1、2─4、3─4 的二字調詞組。

一、前言

華語語音的最大特性之一就是聲調，華語的聲調特性是每個音節都有固定的聲調，不但有高低之分，還有升降曲折之分。聲調的聲學特性主要是由音高來決定的。但是聲調高低不是絕對的，而是相對的，也就是說不同的人發音的調域不同，同一個人在不同時候發同一個音時其中的音高也不完全相同，一般來說，男人音高比女人低，成人音高比兒童低。因此所謂的相對性就是說話者本身的調域內高低穩定性。

有些教學者建議以單音節詞練習來彌補華語學習者的聲調錯誤，但是實際教學中可發現連調的練習更重要，特別是二字調的練習。通常華語學習者經過反覆的練習幾乎都可以掌握到華語單音節的四聲，但是一旦讓學生唸兩個音節以上的詞語，其可懂度和自然度就大打折扣了。造成這樣的原因在於學習者和教學者往往忽略了在語流中的聲調的配合的模式：如變調的規則和連調模式。在二字調、三字調、四字調甚至五字調中，尤其應該以二字調爲主。就如同朱川〈1997〉所說，語言教學應該有個界線不能漫無目的，他認爲連調教學應以二字調爲主，基於以下兩個原因。第一點、二字調幾乎囊括了語流中各種單字調的連接方式，所以二字調是基礎。第二點、現代華語多以雙音節詞爲多。所以二字調的練習可以避免只練習單音節詞的錯誤，也可以解決大部分連調的聲調問題。

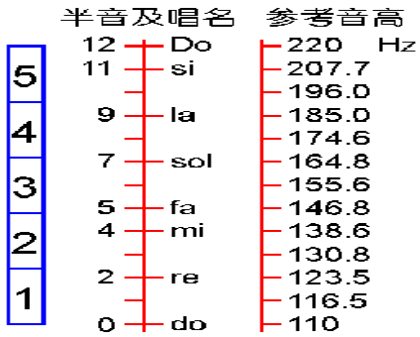
在說話的連續過程中，華語聲調的連續變化對學習華語的學生相當困難。在實際教學中也發現，日本學生在學習華語時，往往因爲在表達時將聲調說錯了而造成溝通上的困難。本文以華語雙音節詞來探討日本學生在學習華語時聲調偏誤的現象，尤其在哪些聲調會產生困難。並探討日語的重音對於日本學生學習華語聲調的負遷移效果，以期能對華語文發音教學上有所助益。

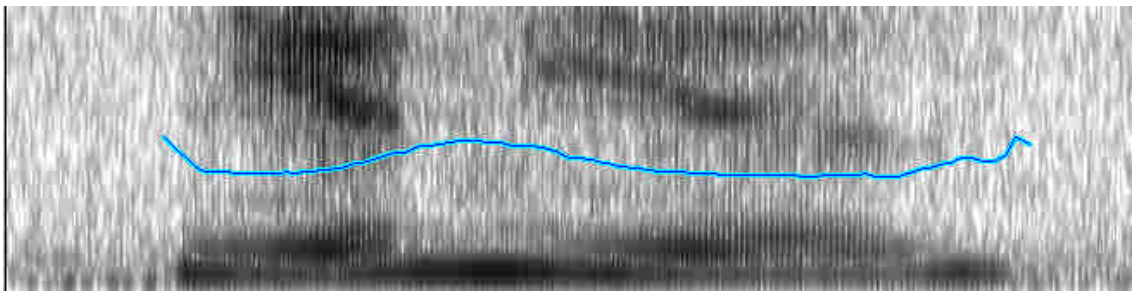
二、文獻探討

2-1 華語和日語的語音特性

華語的每個音節都有固定的聲調，字音的高、低、升、降是由聲帶顫動的快慢決定的〈見表一〉。華語聲調的構成不僅僅取決於音高，還取決於音高之間變化所造成的平調、升調、降調和曲折調型。華語的聲調跟音強、音長也有一些關聯。音強指聲音的輕重或強弱，譬如華語的輕聲和音強有關。目前描寫和紀錄聲調調值最簡便、有效的方法是趙元任先生所提出的五度制標調法。五度制標調法是把聲調的音高分爲五度，並用一條豎線四等分，確立五個座標點，自下而上用數字 1、2、3、4、5 分別表示低、半低、中、半高、高五度音高。然後用線條的方式從左至右把各個聲調的具體音高變化標出來，通過聲調高低升降或曲折變化顯示各個聲調的具體調值。吳宗濟〈1992〉提出華語二字連讀變調調型在表達時，當前後兩個聲調的升降起伏相連接時，會盡量使兩個聲調的連接處平滑些，如前調的尾高而後調的頭低。這種前後相互遷就的流程還帶著一些寧低勿高（第二個音節的高度會比第一個音節的高度還低一點）的趨勢，譬如當兩個陽平相連時，後一個陽平調就會變成低升〈見圖一〉，當兩個去聲相連時，兩個音節的去聲互相影響起音都變低了〈見圖二〉。

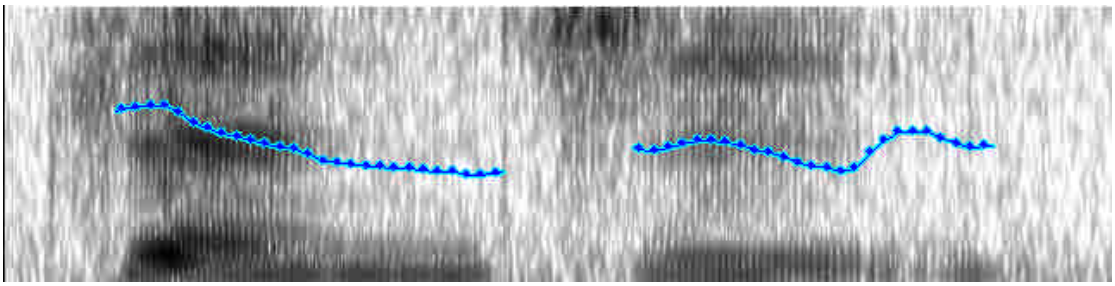
表一 參考音高頻率





年 年

圖一、以華語爲母語者發 2-2 音，藍色是標示聲調



政　　治

圖二、以華語爲母語者發 4-4 音

日語語音特性中和華語聲調有相關的就是重音（アクセント， accent）。渡邊弘史（2003）提到：全世界的語言類型中有兩種重音，一種是「強弱重音」（つよさアクセント），它是利用嗓音的強弱來辨別語詞，另一種是「高低重音」（たかさアクセント），它是利用嗓音的高低來辨別語詞。日語就是屬於「高低重音」（以下簡稱爲重音）。王壽雲〈1997〉、重松淳〈2001〉提到日語的「高低重音」可分爲「平板型」（へいばんしき）和「非平板型」〈也稱爲起伏型きふくしき〉兩種。王〈1997〉還提到日語某些「拍」要高讀，某些「拍」要低讀。日語的「拍」是指發音的長度。「拍」和音節不同，如：しんぶん（新聞）有兩個音節，可是有四拍，日語的高低重音是表現在拍當中，而不是在音節中。

「平板型」和「非平板型」的區別在於「平板型」沒有「重音核」（アクセントかく）的音節，即單字內部無由高音到低音的音程變化。而非平板型則有重音核。「非平板型」又可分爲三種：1.頭高型 2.中高型 3.尾高型。簡明日漢辭典（2002）中以金田一式標記法在日語單詞後標記⓪、①、②、③、④、⑤等記號說明音高的變化。⓪型表示除第一拍低讀外，第二拍以後都高讀，並延續到後續一拍的助詞，這種音調屬「平板型」，如：はし⓪（端）、ともだち⓪（友達）。①型表示第一拍高讀，第二拍以後一律低讀，該音調屬頭高型，如：ねこ①（猫）。②型表示第一拍低讀，第二拍高讀，以後又低讀，後續的一拍助詞也要低讀，該音調中除兩拍的單詞屬尾高型外，三拍或三拍以上的單詞都屬中高型，如：かわ②〈川〉、のみもの②（飲み物）。③型表示第一拍低讀，第二拍和第三拍高讀，以後又低讀，後續的一拍助詞也低讀。該音調除三拍的單詞屬尾高型外，四拍或四拍以上的單詞都屬中高型，如：おとこ③（男）、みずうみ③（湖）。④型表示第一拍低讀，從第二拍開始直到第四拍都高讀，以後又低讀，該音調除第四拍的單詞屬尾高型外，五拍或五拍以上的單詞都屬中高型，如：おとうと（弟）、わたしぶね（渡し舟）。⑤型表示第一拍低讀，從第二拍開始直到第五拍都高讀，以後又低讀，該音調除五拍的單詞屬尾高型外，六拍或六拍以上的單詞都屬中高型，如おしょうがつ（お正月）、たんさんガス（炭酸ガス）。

根據上述型態可以歸納出標準日語（東京話）的重音有以下幾點特性：一、一個單詞中，只能出現一個高讀部分（一拍或幾拍連在一起），絕不會出現兩個高讀部分。二、一個單詞中，第一拍和第二拍的高度必不相同。若第一拍爲高讀，第二拍必是低讀，若第一拍低讀，第二拍必是高讀。三、華語聲調第一聲和第三聲的音高變化在日語高低重音中是空缺的。

日語重音和華語聲調表面上看來都是音程的高低變化，然而日語重音表現在一個單詞內拍和拍之間的高低關係，即以各個拍爲單位的音高變化。而華語的聲調則表現在各個音節內部間的高低變化，即以各個語素爲單位的音高變化。華語共有四種基本調值〈輕聲除外〉---高平調、中升

調、低降升調、全降調。以調類來看則有---陰平、陽平、上聲、去聲，以調號來標示則爲「55」、「35」、「214」、「51」〈參考表二〉 。

表二 趙元任先生創制的五度制聲調符號

調類：陰平	陽平	上聲	去聲
調值：高平 55	中升 35	降升 214	全降 51
例 ：媽**(mā)**	麻**(má)**	馬**(mǎ)**	罵**(mà)**
音長：次短	次長	最長	最短

2-2 日本學生學習華語聲調的偏誤

朱川〈1994〉認爲日本學生在學習華語時產生三個明顯的偏誤，一、聲調偏平，二、連詞誤讀，三、輕聲重讀。許多日本學生遇到華語二字詞時，不論這兩個音節原調是什麼，一律誤讀爲「升降格」，如：把「春風」說成「純風」，把「方便」說「房便」等。造成誤讀的原因與日語的「音便」（爲日文漢字）有一定的關係。日語「音便」其中的一種表現就是日語在構成雙音節以上的詞會出現一些音高變化。也就是說無論原來日語的音高型式爲何，當兩個詞合成爲一個詞之後，只允許保留「中高式」的音高變化。例如わせだ□（早稲田）原音高爲前高型，だいがく□（大学)原音高爲尾高型 ,但兩個詞組合後音高形式則變成中高型的わせだだいがく□(早稲田大学)。這是因爲日語語音特性不允許在一個詞中有二次的音高變化。也就是說日語合成詞只有一個音高的高峰。

吳〈1992〉建議華語聲調教學可以按照音位系統的概念，把華語的四個聲調分爲有區別性的四種調型特徵，即高〈H〉、低〈L〉、升〈R〉、降〈F〉。陰平是四個調域內最高的，所以訂爲「高」。上聲在連讀時，只有兩種調型，一種爲後半上，這與陽平相同爲高升調，所以可以把這種高升與陽平歸爲一類。另一種爲前半上爲低降調，因此將它定於「低」的這類。陽平和去聲分別爲升、降兩類。(見表三)

表三 二字調連讀變調的區別特徵 〈吳宗濟，1992〉

調類	陰陰	陰陽	陽上	陰去
特徵	HH	HR	HL	HF
調類	陽陰	陽陽	陽上	陽去
特徵	RH	RR	RL	RF
調類	上陰	上陽	上上	上去
特徵	LH	LR	LL → RL	LF
調類	去陰	去陽	去上	去去
特徵	FH	FR	FL	FF

何平〈1997〉 提出日本學生在聽辨上聲與去聲、陽平和上聲、陽平與去聲方面非常困難。特別是陽平和上聲一起聽時，日本學生往往區別不開，十分容易把上聲當作陽平。何平還提到日本學生常常起音度掌握不好，如華語的陰平【55】是高平調，學生由於起音過低而發成半高平調【44】或中平調【33】。華語的陽平【35】是高升調，學生由於起音低而發成中升調【24】或低升調【13】。華語的上聲【214】是低降升調，學生常常發爲升調【14】或【35】。華語的去聲【51】是全降調，學生由於起音不高，常常讀成低降調【31】或【21】。

三、研究方法及步驟

3-1、研究對象

本研究對象爲華語文能力爲初中級的兩位日本學生，這兩位受試者都是國立中山大學華語中心的學生，學習華語時間 3~6 個月不等。本文將兩位受試者分別編號爲受試者甲、乙。其中受試者甲華語能力較受試者乙高。本研究對象還有一位以華語爲母語受試者丙，作爲本實驗的對照組。受試者甲、乙、丙均以同樣的方式唸二字調詞表。

3-2、研究步驟

本研究分爲三個部份，第一部分是先進行自然的訪談，以收集自然的語料來輔助詞表語料，第二部份請受試者依照詞表唸出。爲求本研究的客觀性，詞表共分成 A、B 兩卷，其中內容完全一樣只是順序上重新排列過。詞表的設計主要是參考朱川（1997）（見附錄一）。但本實驗詞表分組與朱川不同，目的在於方便語音分析。

3-3、詞表的設計方法

華語有四個聲調，若放到二字調中排列組合，可得出 16 種排列關係，分別再加上輕聲，則共有 20 種排列。本研究中將這 20 種排列分爲 A、B、C、D、E 五組。1、2、3、4、5 分別表示華語聲調中的陰、陽、上、去和輕聲。分組如下：

A：1-1、1-2、1-3、1-4

B：2-1、2-2、2-3、2-4

C：3-1、3-2、3-3、3-4

D：4-1、4-2、4-3、4-4

E：1-5、2-5、3-5、4-5

詞表中的內容分別將上述各組音〈A、B、C、D、E 組〉重新排列組合，以預防受試者的規律性預期心理。另外，本研究的目的是要測試發音能力，並不是要測試受試者的認字能力，所以詞表還附上注音符號〈bpmf〉，而且詞表中的雙音節詞皆屬於初級的詞彙。在開始進行錄音前，先讓受試者不限時間來熟悉測試詞表，但旁人不可告知受試者正確發音。正式錄音時，受試者若說錯了，也可以讓他自己進行修正，但是旁人不能加以協助。第三部份則是請以華語爲母語者做聽覺測試，將甲、乙受試者聲調上的偏誤分別挑出來。接著再以語音分析軟體 PRAAT 分析並探討日本學生學習華語聲調偏誤現象。

3-4、分析方法

首先研究者以聽覺辨識甲、乙兩位日本學生在發（A、B、C、D、E）五組音時的聲調偏誤，並判斷哪一組音較常出現錯誤。再分別將甲、乙兩位受試者的偏誤率進行高低順序排列。最後則分別將兩位日本學生聲調偏誤的狀況，依 A、B、C、D、E 五組分別探討，並以語音分析軟體 PRAAT 加以測量，與以華語爲母語者的聲調進行對比分析。

四、研究結果

4-1、聲調錯誤率

從表四中可看出甲、乙兩位受試者在聲調上的偏誤主要集中在 B、C 組音上，也就是集中在第一個音節爲二聲和三聲的詞組。

表四 甲、乙兩位受試者聲調偏誤率

前＼後	1		2		3		4		5(E)	
受試者	甲	乙	甲	乙	甲	乙	甲	乙	甲	乙
1(A)	0.333	0.333	0.2	0	0.3	0.3	0.149	0.285	0.2	0.2
2(B)	**0.4**	**0.8**	0.25	**0.75**	0	**0.5**	**0.4**	**1**	**0.667**	0.333
3(C)	**0.375**	0.375	0.286	**0.625**	0	**0.5**	**0.4**	**0.8**	0.25	**0.75**
4(D)	0.333	0	0.333	0	0.167	0	0.2	**0.4**	0	0

由於甲、乙兩位受試者華語程度不盡相同，所以偏誤率高低是取個人的平均值。受試者甲偏誤率平均爲 0.336。對受試者甲來說 0.336 以上爲高偏誤率。受試者乙平均值爲 0.4。對受試者乙來說 0.4 以上爲高偏誤率。受試者乙 C 組音中（3-3）偏誤率偏高，推斷受試者乙變調規則沒有掌握好。

甲受試者的偏誤率高低順序：

1. 2-5
2. 2-4、3-4、2-1

乙受試者的偏誤率高低順序：

1. 2-4
2. 3-4、2-1
3. 3-5、2-2
4. 3-2
5. 2-3、3-3

雖然從受試者甲、乙的偏誤率高低順序中，看不出兩人之間偏誤率高低分佈的相同點，但是發現兩個有趣的現象。一是比對兩人的高偏誤率落點，發現在 2-1、2-4、3-4 有重複的現象。再者， 2-3 和 3-3 的錯誤率都完全相同，這在受試者甲及受試者乙皆如此。這主要是 3-3 變調後

爲 2-3，所以這兩組產生的結果應該是相同。但是乙受試者發 3-3 的錯誤率高達 0.5，由此可見乙受試者對於變調規則並沒有掌握好。以下是甲、乙兩位受試者聲調個別偏誤的情形。

4-2 甲、乙兩位受試者聲調偏誤情形

A 組音：

受試者甲

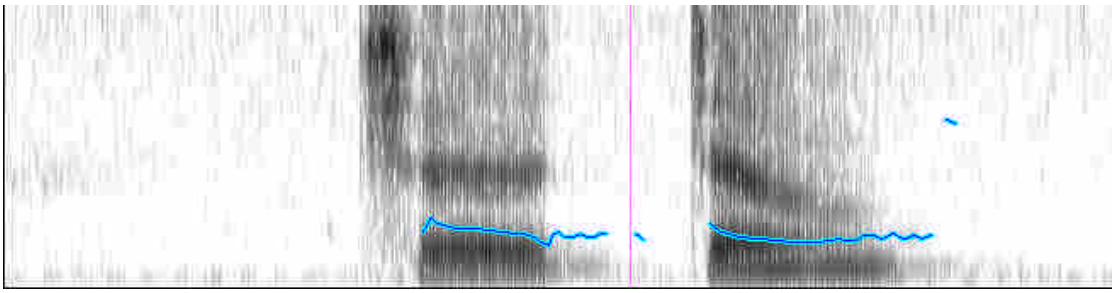
標準 / 誤讀	1-1	1-2	1-3	1-4
	1-4	2-2	2-3	4-4

受試者乙

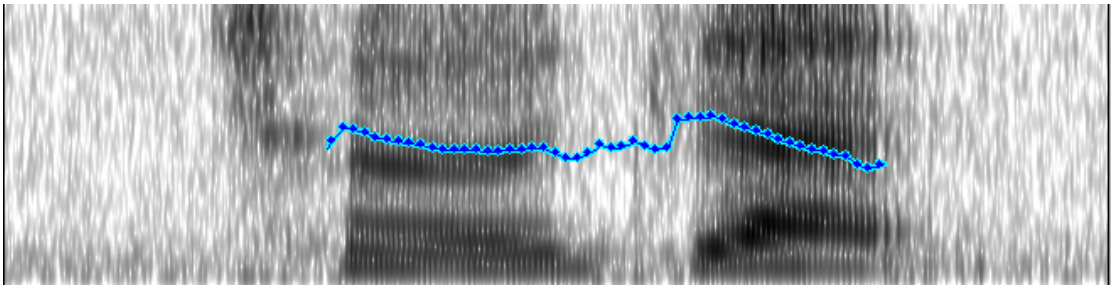
標準 / 誤讀	1-1	1-2	1-3	1-4
	4-2	---	1-3/(1-2)	1-1
				1-1

在 A 組音中，在受試者甲語料有朱川所提出的「升降調」偏誤模式，其中在 1-2、1-3 中分別誤發爲 2-2、2-3。在兩位受試者中也發現了趙麗君（2003）所提的日本學生在發一聲時常常會錯發爲二聲或四聲的現象。受試者乙在發 1-4 音時總是誤發爲 1-1，推斷是四聲降不下去，所以則維持第一音節音高發爲一聲。如下圖。

車站(受試者乙) 1-4 ⟶ 1-1(誤讀)



車站(以華語爲母語者)



B 組音：

受試者甲

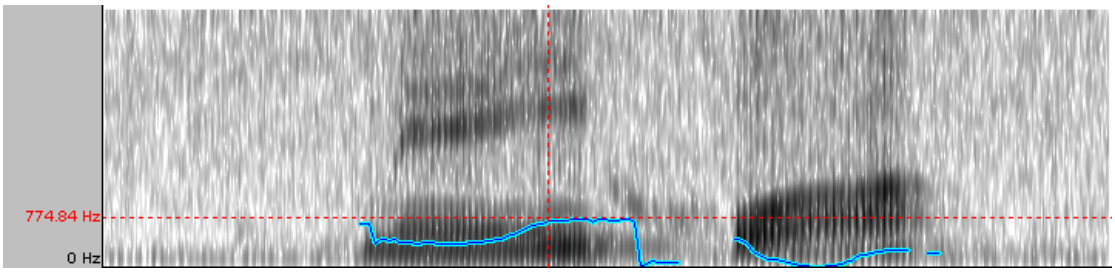
標準 / 誤讀	2-1	2-2	2-3	2-4
	1-1	2-3	---	1-4
	1-1(2-2)	3-1(2-1)		1-4

受試者乙

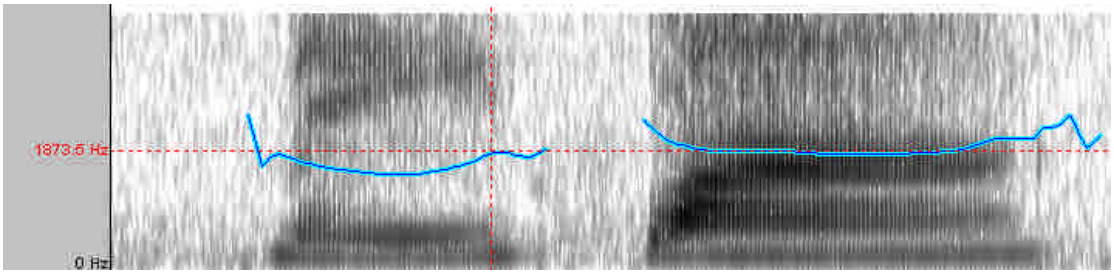
標準 / 誤讀	2-1	2-2	2-3	2-4
	1-1	1-1	1-2/(1-3)	1-4
	2-3	1-1/(2-2)	1-3	1-4
		1-2/(2-2)		2-3/(2-4)
		1-2		1-4
		1-2		
		2-1		

在 B 組音中，不論是受試者甲還是受試者乙，二聲普遍發不上去，而錯發爲一聲〈但是音高較以華語爲母語者低〉。趙麗君提出就算第一個音節二聲發正確了，第二個音節的音還是會因爲同化作用〈Assimilation〉錯發爲三或四聲。在本研究中，發現除了上述提出的誤讀現象，第二個音節還有錯發爲一聲的情況。在 B 組音中第一個音節正確但第二音節錯誤的音共有 4 個（如果不算糾正後的音），分別是在甲受試者的 2-2 組音錯發爲 2-3；乙受試者的 2-1 組音錯發爲 2-3；2-2 組音錯發爲 2-1；2-4 錯發爲 2-3。根據上述發現，兩位受試者在 B 組音都常有錯發爲 2-3 的現象。造成這樣的誤讀是由於依日語重音特性在一個單詞中只允許一個高峰，而且高讀之後就一定要低讀。而 2-1 這組音在二聲結束後還維持高讀，對於日本學生來說，是較困難的（如下圖）。

梅花(受試者乙) 2-1 ⟶ 2-3(誤讀)



梅花(以華語爲母語者)



C 組音

受試者甲

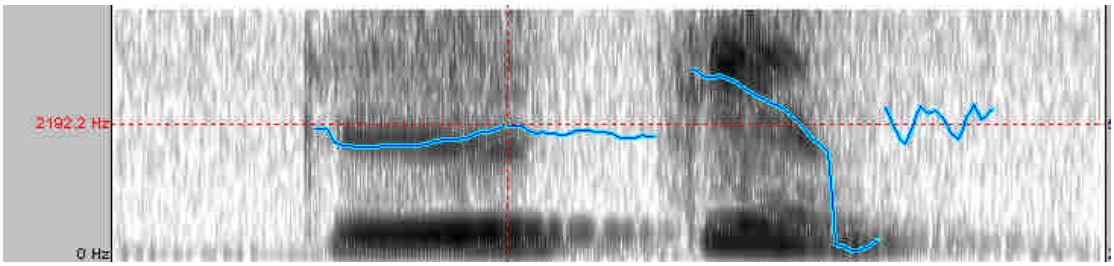
標準 / 誤讀	3-1	3-2	3-3	3-4
	2-1	3-1	----	2-1
	4-1	3-3/(2-3)		2-4
	1-1/(2-1)			

受試者乙

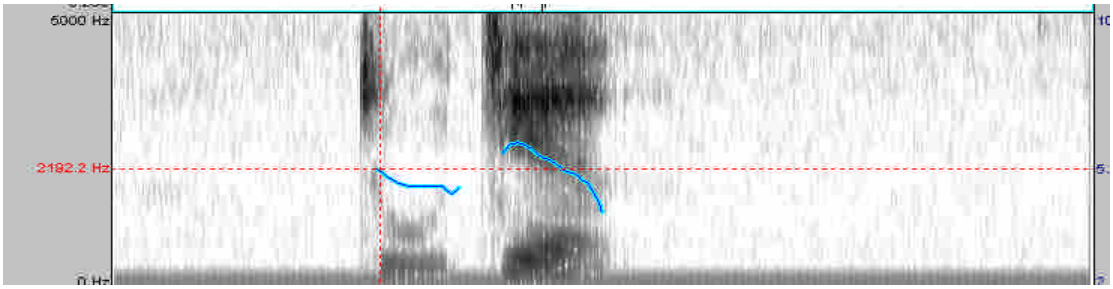
標準 / 誤讀	3-1	3-2	3-3	3-4
	4-1	2-3	1-3	2-1
	1-1	1-2	1-3	2-4/(2-3)/1-3
	1-1	1-1/1-2	1-3	2-1
		2-1		1-4
		2-2/1-1		

在 C 組音中發現第一個音錯發爲一聲和二聲居多。日本學生在發三聲時常錯發爲上升調。但由於日本學生本來在發二聲時就常常發不上去，所以在這組音中也可以看到錯發爲一聲的情形。由受試者甲發 3-3 音錯誤率爲零情形看，甲受試者對於三聲變調規則完全掌握。同時甲受試者發 2-3 音，也發現錯誤率爲零。但是乙受試者在發 3-3 和 2-3 時偏誤率均爲 0.5，可知乙受試者對三聲變調的規則掌握情況還不純熟。在下圖 3-4 這組音裡，我們發現以華語爲母語者在發「主」是發 214 裡的 21 的時長較長，後半上 14 的時長只有一點點。而日本學生則相反，在發 214 裡的後半段的 14 的時長相當長，而前半段的 21 時長則相當短，所以我們聽起來日本學生在發這個音時就像是發二聲。

主見(受事者甲) 3-4 ⟶ 2-4(誤讀)



主見(以華語爲母語者)



D 組音

受試者甲

標準 / 誤讀	4-1	4-2	4-3	4-4
	1-1	----	1-3	1-4

受試者乙

標準 / 誤讀	4-1	4-2	4-3	4-4
	---	2-1	---	1-4
				4-1

在 D 這組音裡，我們發現日本學生在這組音裡偏誤率最低，出現很多零錯誤率。甲受試者第一個音以錯發爲一聲爲多，符合朱川提出的誤讀爲「升降調」的特性。另外輕聲的調值因爲是由前一個音節的聲調決定，所以以另外一個章節來探討。

4-3、甲、乙受試者 E 組音聲調偏誤情形

E 組音

受試者甲

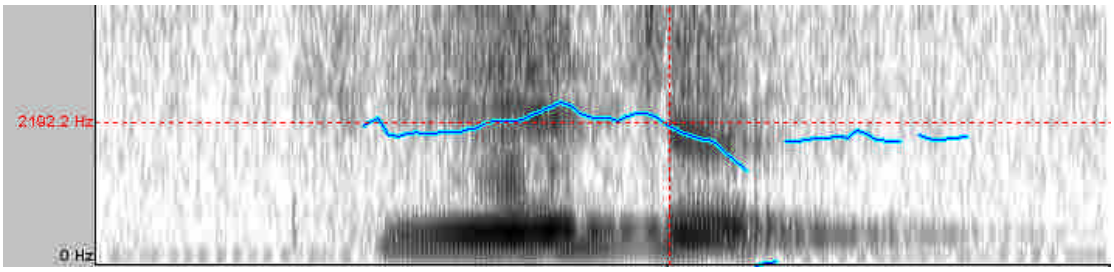
標準 / 誤讀	**1-5**	**2-5**	**3-5**	**4-5**
	2-5	2-1	2-5	----

受試者乙

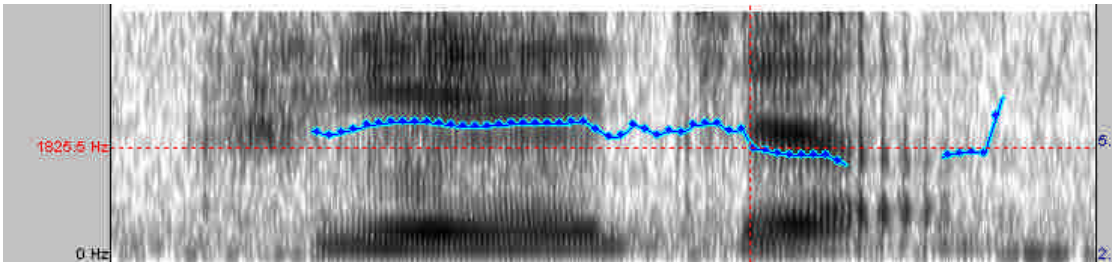
標準 / 誤讀	**1-5**	**2-5**	**3-5**	**4-5**
	2-5	1-5	2-5	----
		3-5	1-5	
			1-5	

朱川提到輕聲在語音分析中常常被處理爲變調現象，因爲它的調值由它前面音節的聲調決定。在陰平及陽平後輕聲音節音高爲 31，.在上聲後爲 4，在去聲後爲 1。而且輕聲音節的時長一般來說較短。輕聲的音高雖然總的來說都是又輕又短，但卻不是一成不變，它總是隨著前一個音節末尾的趨勢而改變自己的音高。甲、乙受試者常見的偏誤情形如 1-5 的音均誤發爲 2-5。以下將一一檢視以華語爲母語者和以華語爲第外語者在二字調中輕聲的差異。

接著(受事者甲)1-5 ⟶ 2-5(誤讀)

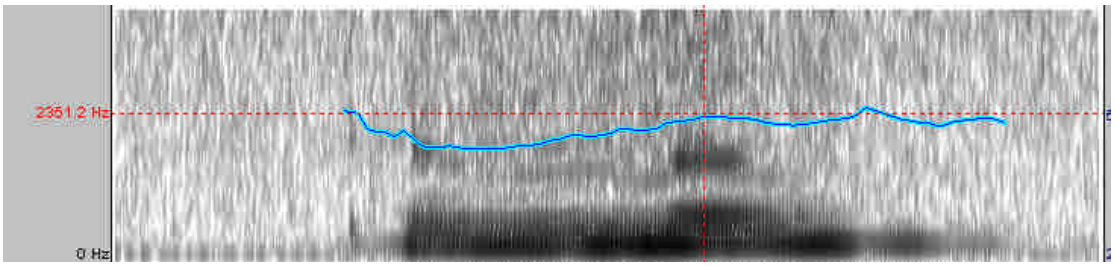


接著(以華語爲母語者)

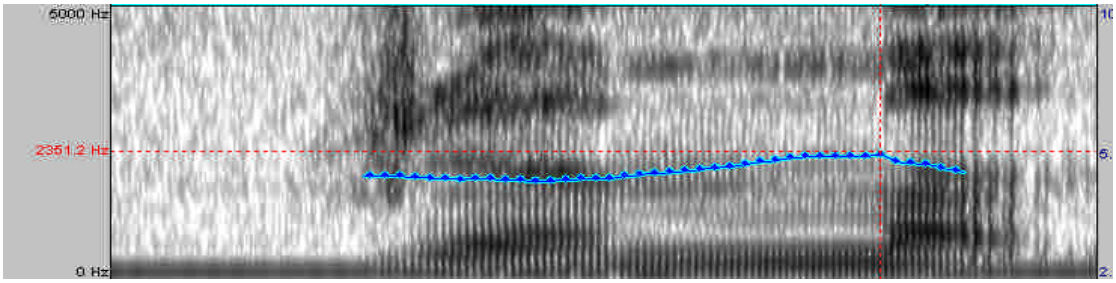


在第一個音節是一聲的情況下，我們發現日本學生在發這類音時，下降的動程較以華語爲母語者大。

人們(受事者甲) 2-5 ⟶ 2-1(誤讀)

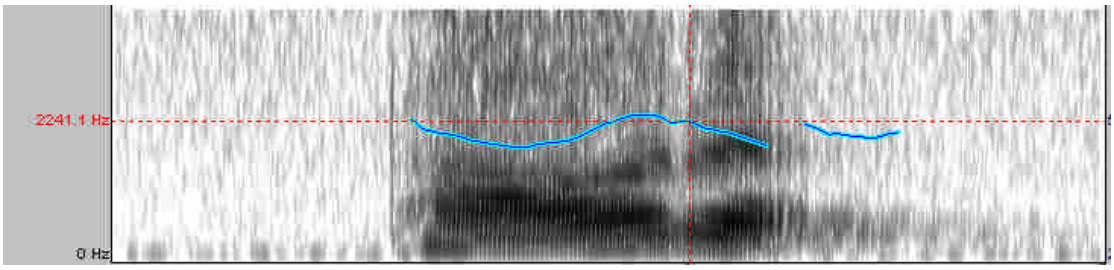


人們(以華語爲母語者)

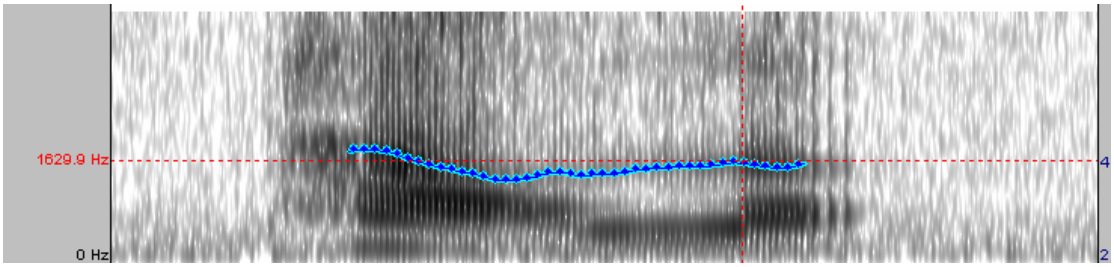


在「人們」的「們」，我們發現以華語爲母語者發音時較平滑，且時長短。而日本學生發輕聲時長明顯過長。

跑了(受事者甲) 3-5 ⟶ 2-5

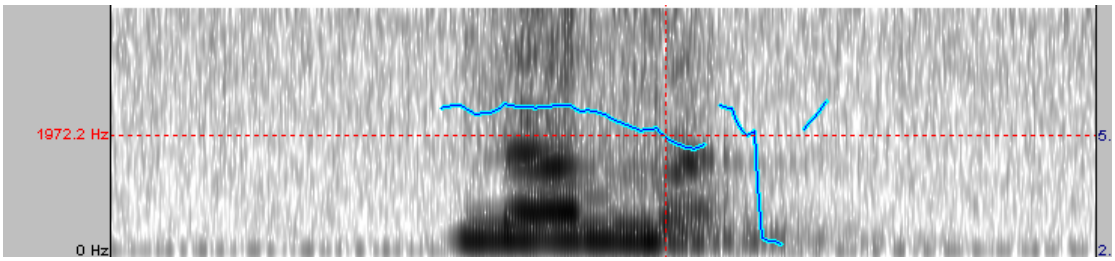


跑了(以華語爲母語者)

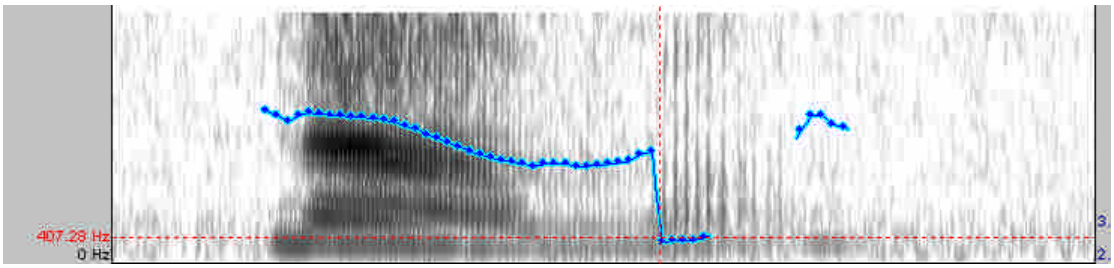


以華語爲母語者在發這類音時，在第二個音節的結尾有些微的上揚，而日本學生由於第一個音錯發爲二聲，所以在第二個音節的輕聲發現調型是呈現下降趨勢。

那麼(受事者甲) 4-5 ⟶ 1-5(誤讀)



那麼(以華語爲母語者)



在這組音中，也可以發現日本學生在第二個音節頻率明顯過高。

五、結論

從以上的分析中發現，甲、乙兩位受試者在聲調上的偏誤主要集中在 B、C 組音上，也就是集中在第一個音節爲二聲和三聲的詞組。兩位受試者的高偏誤率均集中在 2-1、2-4、3-4 組音。再者， 2-3 和 3-3 的錯誤率都完全相同，這在受試者甲及受試者乙皆如此。

在 A 組音中，日本學生在發一聲時常常會錯發爲二聲或四聲，但是錯發爲二聲的情況較錯發爲四聲情況多。受試者甲在 1-2、1-3 中分別誤發爲 2-2、2-3。在 B 組音中，日本學生二聲普遍發不上去，而錯發爲一聲〈但是音高較以華語爲母語者低〉，就算第一個音節發正確了，第二個音節的音也會因爲同化作用〈Assimilation〉錯發爲三或四聲或一聲。譬如，甲受試者的 2-2 組音錯發爲 2-3。乙受試者的 2-1 組音錯發爲 2-3；2-2 組音錯發爲 2-1；2-4 錯發爲 2-3。在 C 組音中發現第一個音錯發爲一聲和二聲居多，日本學生在發三聲的時候常錯發爲升調二聲，但由於日本學生本來在發二聲時就常常發不上去，所以在這組音中也可以看到錯發爲一聲的情形。受試者甲在發 2-3 和 3-3 都呈現在零偏誤率，顯示甲受試者對於變調規則已經確實掌握。但乙受試者在發 3-3 和 2-3 時偏誤率偏高，可知乙受試者對三聲變調的規則掌握情況還有待加強。日本學生在 D 這組音裡偏誤率最低，在甲受試者中發現在 D 組音中第一個音節以錯發爲一聲爲多。而在 E 組音中，日本學生普遍發音時長過長，甲受試者和乙受試者在 1-5 組音均錯發爲 2-5。甲受試者在 2-5 組音中錯發爲 2-1，主要因爲第二個音節的時長過長，所以聽起來像一聲。在 3-5 組音中，兩位受試者都有錯發爲 2-5 的現象。但甲、乙兩位受試者在 4-5 組音均爲零偏誤。綜上所述，由於華語聲調和日語高低重音兩者的作用不同，藉著對比分析找到規律，可幫助華語學習者掌握正確的聲調變化。

附錄 一

普段話している速さと声の大きさで、下記の文章を読んでください
請用平常說話的速度、聲音、大小，唸出下面的字。

Subject No._____
姓名：_____
中国語の勉強時間：_____
Date：_____

1.新ㄒㄧㄣ村ㄘㄨㄣ、便ㄆㄧㄢˊ宜ㄧˊ、求ㄑㄧㄡˊ情ㄑㄧㄥˊ、新ㄒㄧㄣ春ㄔㄨㄣ、知ㄓ心ㄒㄧㄣ。
2.心ㄒㄧㄣ事ㄕˋ、知ㄓ識ㄕˋ、粗ㄘㄨ布ㄅㄨˋ、車ㄔㄜ站ㄓㄢˋ、高ㄍㄠ興ㄒㄧㄥˋ。

3.會ㄏㄨㄟˋ話ㄏㄨㄚˋ、政ㄓㄥˋ治ㄓˋ、立ㄌㄧˋ論ㄌㄨㄣˋ、無ㄨˊ賴ㄌㄞˋ、求ㄑㄧㄡˊ救ㄐㄧㄡˋ。
4.電ㄉㄧㄢˋ梯ㄊㄧ、汽ㄑㄧˋ車ㄔㄜ、不ㄅㄨˋ開ㄎㄞ、半ㄅㄢˋ天ㄊㄧㄢ、大ㄉㄚˋ家ㄐㄧㄚ。

5.年ㄋㄧㄢˊ年ㄋㄧㄢˊ、飛ㄈㄟ機ㄐ、癡ㄔ心ㄒㄧㄣ、廚ㄔㄨˊ房ㄈㄤˊ、學ㄒㄩㄝˊ習ㄒㄧˊ。
6.方ㄈㄤ法ㄈㄚˇ、工ㄍㄨㄥ廠ㄔㄤˇ、開ㄎㄞ始ㄕˇ、公ㄍㄨㄥ里ㄌㄧˇ、發ㄈㄚ展ㄓㄢˇ。

7.點ㄉㄧㄢˇ滴ㄉㄧ、女ㄋㄩˇ家ㄐㄧㄚ、恍ㄏㄨㄤˇ惚ㄏㄨ、主ㄓㄨˇ張ㄓㄤ、呂ㄌㄩˇ家ㄐㄧㄚ。

8.事ㄕˋ實ㄕˊ、四ㄙˋ十ㄕˊ、事ㄕˋ實ㄕˊ、動ㄉㄨㄥˋ畫ㄏㄨㄚˋ、道ㄉㄠˋ路ㄌㄨˋ。

9. 制ㄓˋ止ㄓˇ、住ㄓㄨˋ防ㄈㄤˊ、淘ㄊㄠˊ汰ㄊㄞˋ、評ㄆㄧㄥˊ價ㄐㄧㄚˋ、無ㄨˊ奈ㄋㄞˋ、

10.收ㄕㄡ拾ㄕˊ、梅ㄇㄟˊ花ㄏㄨㄚ、直ㄓˊ接ㄐㄧㄝ、花ㄏㄨㄚ肥ㄈㄟˊ、積ㄐㄧ極ㄐㄧˊ。

11.提ㄊㄧˊ交ㄐㄧㄠ、資ㄗ源ㄩㄢˊ、休ㄒㄧㄡ息ㄒㄧˊ、國ㄍㄨㄛˊ家ㄐㄧㄚ、打ㄉㄚˇ算ㄙㄨㄢˋ。

12起ㄑㄧˇ重ㄓㄨㄥˋ、舉ㄐㄩˇ重ㄓㄨㄥˋ、主ㄓㄨˇ見ㄐㄧㄢˋ、離ㄌㄧˊ開ㄎㄞ、表ㄅㄧㄠˇ示ㄕˋ。

13.字ㄗˋ紙ㄓˇ、回ㄏㄨㄟˊ答ㄉㄚˊ、明ㄇㄧㄥˊ年ㄋㄧㄢˊ、年ㄋㄧㄢˊ級ㄐㄧˊ、課ㄎㄜˋ本ㄅㄣˇ。

14.廚ㄔㄨˊ房ㄈㄤˊ、可ㄎㄜˇ能ㄋㄥˊ、檢ㄐㄧㄢˇ查ㄔㄚˊ、以ㄧˇ前ㄑㄧㄢˊ、辦ㄅㄢˋ法ㄈㄚˇ、

15.解ㄐㄧㄝˇ決ㄐㄩㄝˊ、桌ㄓㄨㄛ子˙ㄗ、椅ㄧˇ子˙ㄗ、旅ㄌㄩˇ行ㄒㄧㄥˊ、起ㄑㄧˇ床ㄔㄨㄤˊ、
16.航ㄏㄤˊ海ㄏㄞˇ、結ㄐㄧㄝˊ果ㄍㄨㄛˇ、沒ㄇㄟˊ有ㄧㄡˇ、牛ㄋㄧㄡˊ奶ㄋㄞˇ、啤ㄆㄧˊ酒ㄐㄧㄡˇ。
17.表ㄅㄧㄠˇ演ㄧㄢˇ、揍ㄗㄡˋ了˙ㄌㄜ、爸ㄅㄚˋ爸˙ㄅㄚ、可ㄎㄜˇ以ㄧˇ、了ㄌㄧㄠˇ解ㄐㄧㄝˇ、

18.代ㄉㄞˋ表ㄅㄧㄠˇ、電ㄉㄧㄢˋ影ㄧㄥˇ、接ㄐㄧㄝ著˙ㄓㄜ、叉ㄔㄚ子˙ㄗ、杯ㄅㄟ子˙ㄗ。

19.覺ㄐㄩㄝˊ得˙ㄉㄜ、名ㄇㄧㄥˊ字˙ㄗ、人ㄖㄣˊ們˙ㄇㄣ、飽ㄅㄠˇ了˙ㄌㄜ、跑ㄆㄠˇ了˙ㄌㄜ。
20.湊ㄘㄡˋ了˙ㄌㄜ、那ㄋㄚˋ麼˙ㄇㄜ、哪ㄋㄚˇ裡ㄌㄧˇ、水ㄏㄨㄟˇ果ㄍㄨㄛˇ、你ㄋㄧˇ們˙ㄇㄣ。

21.老ㄌㄠˇ子˙ㄗ、腦ㄋㄠˇ子˙ㄗ。

參考書目

王壽雲（1997）。日語聲調及其讀音模式。福建外語，2 期，24-26。

朱川（1994）。漢日超音質特徵對比實驗。華東師範大學學報，1 期，85-86。

朱川（1997）。外國學生華語語音學習對策。北京：語文出版社。

何平（1997）。談對日本學生的初級華語語音教學。語言教學與研究，3 期，49-50。

吳宗濟（1992）。現代華語語音概要。北京：華語教學出版社。

重松淳 （2001）。日本人學習華語聲調方面的一些問題和解決方法。對日華語教學國際研討會論文集。中國社會科學出版社，172。

渡邊弘史（2003）。日語重音之理論與語音分類。吳鳳學報，11 期，283-286。

趙麗君（2003）。有針對性的對日本學生進行語音教學。雲南師範大學學報，1 卷，3 期，66-67。

劉文祥、馬金森、鄭玉和、李紹庚、黃瑞金、王希時（2002）。簡明日漢辭典。台北：大新書局。

FAST：電腦輔助英文文法出題系統
FAST：Free Assistant of Structural Tests

陳佳吟[1] 柯明憲[2] 吳紫葦[1] 張俊盛[1,2]
{g936727, g936339, g936704}@oz.nthu.edu.tw；
jschang@cs.nthu.edu.tw
[1]清華大學資訊系統與應用研究所
[2]清華大學資訊工程研究所

摘要

近年來電腦輔助教學應用於自動產生試題系統的研究，在自然語言處理領域裡異軍突起，成爲現今熱門探討的重點。綜觀目前的自動出題系統，大都著重字彙、克漏詞、閱讀測驗題型，並沒有針對英文文法爲考量的相關研究。本論文提出一個以網路爲本的作法，可以自動產生英文文法測驗考題。這個句法樣式爲本的做法，涉及對閱讀的文章進行詞性分析、基本片語分析等等，再根據多重的句法樣式，擷取具備特定句法樣式的句子，形成題目、答案、誘答項目。我們也根據這個做法，製作了電腦輔助線上文法測驗系統雛形，FAST (Free Assessment of Structural Test)。實驗的結果顯示，FAST 能將七成以上搜集得來的英文句子，自動生成文法考題。因此，FAST 運用適性化數位學習的潛力很大。

Key Words：電腦輔助教學應用，英文文法，電腦自動出題

一、緒論

電腦輔助語言學習（Computer Assisted Language Learning，CALL）興起於 1950 年代。近年來，電腦輔助產生試題（Computer Assisted Item Generation，CAIG）也開始受到重視，逐漸成爲電腦輔助語言學習領域的研究課題。藉由電腦科技的輔助，透過演算法的計算，電腦輔助產生試題系統能產生大量、豐富多樣的試題，補強人工出題費時費力的缺點，同時亦能提高試題的信度與效度。此外，系統有潛力可以支援適性化出題（adaptive testing），經由提供難度不同的試題，協助受試者漸進式學習。

縱觀現有電腦輔助產生試題系統與相關研究，出題重點多著重於英文字彙測驗、英文克漏詞測驗與英文閱讀測驗，英文文法測驗則相當缺乏。然而非母語的學習者（Non-native speaker，NNS）學習語言上，文法佔有一定的重要性，因此，我們深感利用電腦輔助、自動產生英文文法試題的必要性和重要性。

Larsen-Freeman（1997）曾指出文法並非單方向的死板規則，其所包含的面向包括型態（form）、意義（meaning），與用法（use）三面向。所以英文文法測驗的目的雖在檢視受試者對於英文句子結構是否瞭解，受試者的字彙能力、片語能力、閱讀能力與寫作能力與受試者作答文法測驗能力卻有一體兩面的影響，因而單純一道英文文法測驗試題，不僅測驗受試者是否融會貫通文法概念，也測驗受試者是否具備上述各項能力。

文法測驗的目的在於評量學習者對某一語言基本的結構與詞彙順序的掌握程度，通常以選擇題的形式爲多。最好題目所採用的句子，必須具備了語言學習者，特別是非母語的學習者，不容易掌握的困難結構，而且最好句子是真實的，意思完整的。在選擇項目方面，必須和挖空位置前後的幾個詞彙，看起來似是而非，容易讓一知半解的學生產生混淆，選擇錯誤的項目，因而達到區隔學生文法能力的測驗目標。

因此，人工或自動化的出題，都需要有一套做法，才能達到考試合理的效度和信賴度。若是隨意選取文章的句子作爲考題的主體，且擷取其中任意一個字詞或片語，作爲標準答案，再安排任意的選擇項目，並不能達到我們預期的目標。另外，考試題目也不宜集中在單一的題目類型，使得受

測者有預期心態，無法測驗出真正的語言能力。

因爲文法的結構不在少數，爲了平衡地測驗整體的文法能力，必須選取各種的句子，作爲考題。而在選擇項目之上，又必須設計適當的選項，作爲誘答項目。以上的這些因素，都使得文法出題的過程步驟繁複，必須耗費極大的人力與時間。若是能夠利用電腦自動出題，並保持出題品質，勢必能有效提升文法測驗出題的效率，以強化自主學習、學習評量。

一道完整的英文文法選擇試題（multiple-choice test）必須包括：一句獨立（context-independent）且語意清楚的完整句子、挖空文法概念的題幹（stem）、被挖空的文法概念則爲該試題的答案，而其餘的三個選項則爲該題的誘答選項（distractor）。例 1 爲一道完整的英文文法選擇試題的範例。

（1） Reading is to the mind _____ exercise is to the body.
(A) so
(B) that
(C) as
(D) what

在目前語言學習與語言測驗學理上，很著重真實的語言，因此測驗的句子，最好是來自語言使用的真實情境，而不是教師爲了教學的目的所造出來的一些簡化的句子。而誘答項目應該和標準答案很相似，以避免一知半解的學生只要比較選項，就可以猜測到正確答案。例如，我們可以利用網路上，自由開放的內容（如 Wikipedia 網路百科全書），擷取到像圖一的文章。由文章中，我們不難發現第三句的結構，開始是一個完整的主要子句，接著一個形容詞片語之後，最後出現一個現在分詞引領的 reduced clause。這樣含有中插元素的句型，很容易讓受測者，混淆了中插元素前後部分的關係，是很常見文法考題的樣式。因此，我們很容易將此一特定句法樣式的句子，轉換成爲例 2 所示的選擇題：

（2） There are six subspecies of the fox, each unique to the island it inhabits,
_____ its evolutionary history.

(A) to reflect
(B) and reflect
(C) reflecting
(D) that reflect

而誘答選項，是同一動詞（reflect）的不同形式，to reflect、and reflect、 reflecting。

直覺上，利用簡單的句法樣式（如例 3），不難產生如例 2 的文法測驗選擇題。

（3） CLAUSE, * ADJ *, X/VBG *.

➔

CAUSE, * ADJ *, _____ *. (A)
(A) X/INFINITIVE
(B) and X
(C) X/VBG
(D) that X

爲了讓選擇題所採用的句子文義完整，不待審視句子前後的文脈，即可瞭解題意而進行作答，所選取的句子，應該自給自足，最好不要含有代名詞、連接詞、定指名詞片語等。像例 4 的句子含有代名詞（it）就不太適合採用製作成試題。

（4）It is the smallest fox species in the United States.

The Island Fox is a small fox that is native to six of the eight Channel Islands of California. It is the smallest fox species in the United States. There are six subspecies of the fox, each unique to the island it inhabits, reflecting its evolutionary history. Introduced diseases or parasites can decimate Island Fox populations. Because Island Foxes are isolated they have no immunity to parasites and diseases brought in from the mainland and are especially vulnerable to those domestic dogs may carry. A canine distemper outbreak in 1998 killed approximately 90% of Santa Catalina Island's fox population. In addition, Golden Eagle predation and human activities decimated fox numbers on several of the Channel Islands in the 1990s. Four Island Fox subspecies were federally protected as an endangered species in 2004, and efforts to rebuild fox populations and restore the ecosystems of the Channel Islands are being undertaken.

圖一 不收費的維基百科在 2005 年 7 月 10 日的有關島狐（island fox）主題文章，來源 http://en.wikipedia.org/wiki/Island_Fox

我們分析了大量的托福（Test of English as a Foreign Language，TOEFL）測驗的文法考題、準備托福考試的參考書，得到一組有效的出題句法樣式。根據這些句法樣式，我們提出一個文法測驗題的自動出題方法。爲了評估該做法的可行性，我們也製作了一個名爲 FAST（Free Assessment of Structural Test）的電腦輔助自動產生英文文法試題系統，試題形式與測試之文法重點比照托福測驗，試題來源則落實網路爲本（Web as corpus）的觀念，取自網路文字資源，我們利用自然語言處理技術，建立電腦輔助自動產生英文文法試題。我們以 Wikipedia 共 2,000 多篇的網頁文章，做出 2,000 多題的選擇題。FAST 能成功地將四分之三蒐集而得的網路資訊，自動產生出英文文法試題，而產生的考題中，有超過四分之一的試題，文法觀念的測驗方向與模擬托福題目相同。初步的實驗之後，我們發現藉由電腦輔助，的確可以減少出題的人力負擔，並加快出題速度。

在接下來的第二節，我們將探討現有電腦自動出題系統與相關研究，並在第三描述 FAST 系統自動文法出題的機制。第四節將介紹 FAST 系統，而第五節將提出未來發展與結論。

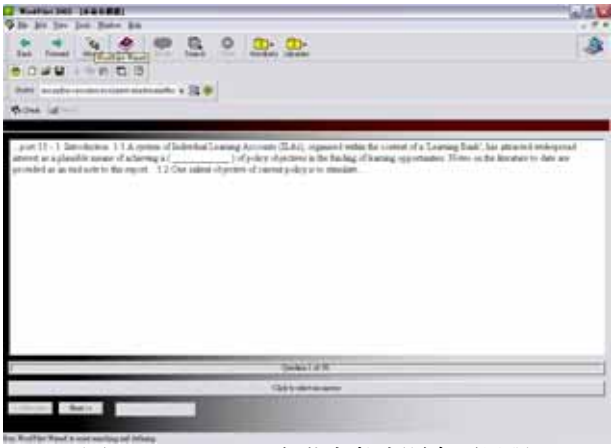
二、相關研究

Warschauer (1996) 提出電腦輔助語言學習發展 30 年來的狀況。最近十年的發展，可以參考 Wichmann, Fligelstone, McEnery 和 Knowles（1997）編輯收納 21 篇論文，探討如何利用語料庫於語言教學與測驗。

和本文最相關的研究中，Wilson (1997) 提出利用一般的語料庫來產生練習題作爲電腦輔助語言學習的活動。所產生的題目包括文法練習題或考題，例如在一個挖空的句子中填入正確的分詞型態「startled」 或 「startling」 。Milton（1998）報告在線上英語學習系統中，自動出題系統 WordPilot 的設計與功能。WordPilot 提供自動出考題的功能，老師可以挑選出數個容易誤用的片語或單字，讓 WordPilot 自動產生如例 5 的選擇題，提供學生自我檢測練習。

（5） … Title: Sulfur and Nitrogen Emission Trends for the U.S. – Introduction INTRODUCT (_____) the substance metabolized by industrial activities, fossil fuels are most significant, both in quantity and by the variety of chemical that are mobilized. Industrial consumers take fuels as inputs a …

WordPilot 介面呈現的方式，如圖二所示。



圖二　WordPilot 自動字彙出題介面呈現

儘管 WordPilot 系統本身已提供完整自動字彙測驗的出題功能，然我們發現此系統仍有很大的改善空間，我們彙整出數項其設計不足之處：

(i) WordPilot 系統自動出題的試題內容，直接取自於系統蒐集而成的資料庫文章，並未加以修飾及調整。如例 5 所示，WordPilot 系統自動產生的英文字彙試題內容頗爲粗糙，且其內文亦不完整，試題的句子長度也不一致，使得自動出題的適度性大打折扣。

(ii) WordPilot 系統自動出題的誘答題，並未真正達到「誘答」的功效，其誘答的單字選項，只是由系統隨機亂數選出，因此出現不少無效誘答選項（useless distractor），影響到試題的難易度（item facility）與試題的鑑別度（item discrimination）。

(iii) 因試題的誘答選項爲系統隨機選出，造成誘答選項多寡不一致的問題，而過多（如 15 個）或過少（如 2 個）的誘答選項，都大爲降低該系統自動出題的專業性。

最近 Mitkov 和 Ha（2003）利用自然語言處理技術擷取英文句子主要關鍵詞，搭配 WordNet 作關鍵詞的替換，建立英文閱讀測驗自動出題系統。Mitkov 和 Ha 將考題實施於真正的考試，並利用測驗理論的指標，證明他們的作法的確可以縮短出題的時間，增進考題的信度與效度。王俊弘、劉昭麟、高照明（2004）以自然語言處理的統計和 selectional preference 技術，搭配詞（collocation）概念生成誘答選項機制，產生自動英文克漏詞試題系統。Liu, Wang, Gao 和 Huang（2005）提出透過詞彙語意解析技術，可針對特定詞彙的某一語意，自動產生詞彙考題的作法。Sumita, Sugaya, Yamamoto（2005）根據 *Item Response Theory* （IRT），提出自動產生填空題的作法，來測驗學生的英語能力。誘答項目是由同義詞網路辭典中選取，再經過查詢驗證確認。實驗結果顯示考題的確可以測驗出非母語學生之英語能力，而母語學生幾乎都可以得到滿分。

和前人的研究不同，我們真正分析真實的文法考題，瞭解文法測驗的目標與策略，透過自然語言處理技術，完成英文文法試題自動出題系統。我們針對考題句子的長度樣式、誘答選項數量、誘答題的設計方式、試題的呈現與互動方式都有深入的探討，也實際製作了雛形系統。

三、FAST 自動產生考題的作法

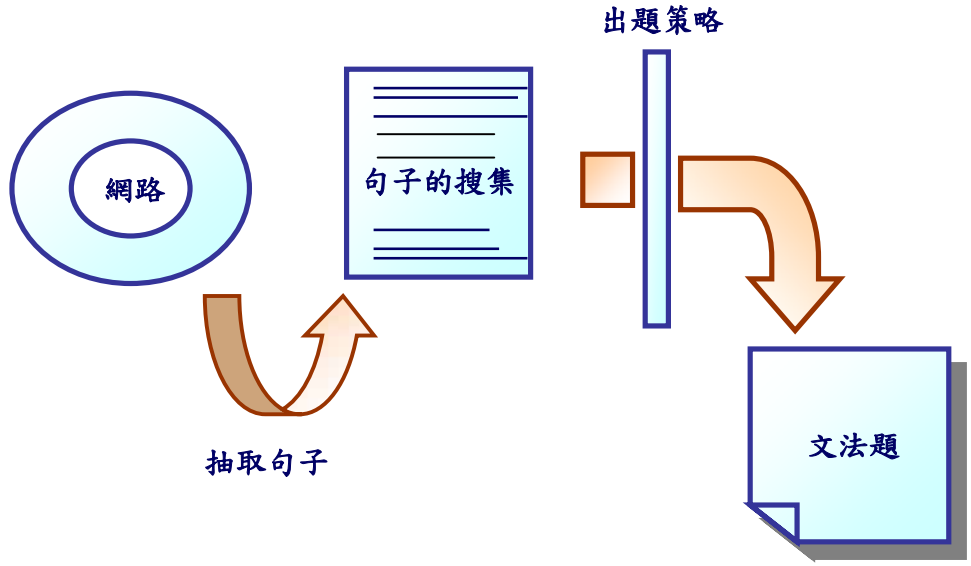
爲了使 FAST 系統能產生和托福文法測驗高相似度的英文文法試題，我們分析近五年來（1998 年～2002 年）托福模擬試題的語料庫（http://Acezh.com），透過統計，歸納托福文法試題題型，並經由人工檢視，分析托福文法考題內容。這些統計數據和分析結果將有助於我們研發 FAST 系統應注意的條件和限制。我們研究分析的托福模擬試題語料庫共有 1279 道文法試題，其中 497 道（39%）爲如例 6 所示的傳統四選一單選文法試題，782 道（61%）爲如例 7 所示的文法改錯題。目前 FAST 系統出題題型以傳統四選一的單選文法試題爲主。

（6）Most doctors of the Colonial period believed _____ was caused by an imbalance of humors in

the body.
(A) in disease
(B) that disease
(C) of disease
(D) about disease

（7）During the Harlem Resaissance the 1920's, much African American
(A)
writers, artists, and musicians came to Harlem in New York City,
(B)
creating a cultural center there.
(C) (D)

經由人工檢視，我們察覺托福文法試題內容不僅符合語意完整且獨立的條件，還具備教育類型（learned genre）文章中，敘述客觀事實的特徵。舉例而言，例 6（選擇題）、例 7（改錯題）托福文法試題，其考題內容皆闡述真實事件或敘述永恆不變的真理。有鑑於此，我們遂決定利用似教科書或百科全書的文體類型，作爲出題句子的來源（例如 Wikipedia）。就題目的長度而言，我們透過統計托福模擬試題發現，托福文法試題長度分佈在 7 到 34 個英文字。
FAST 系統的自動出題過程，整理如下圖三：



圖三　FAST 系統自動出題流程

在本節中，我們將敘述自動出題的方法，包括如何分析並撰寫試題產生句法樣式（3.1 節）、自動蒐集網路文章（3.2 節）、題型與題幹的選擇（3.3 節）、以及產生誘答選項（3.4 節）。

3.1 分析並撰寫試題產生句法樣式

在設計系統之前，有必要將各式各樣的文法題依題型分門別類，再針對不同的題型來編寫程式。由於我們對英文文法題的出題概念缺乏系統性的了解，若是貿然將搜集來的托福考題依憑觀查分類，將顯得雜亂無章，不但欠缺學理的依據，更增加了實作上的困難，因此我們參考了 Pamela J. Sharpe 所著的「How to Prepare for the TOEFL」一書，該書將托福的常考文法題型整理成九大類，共五十種基本文法重點。

得到此一有系統的分類法之後，仍然存在著另一個問題，亦即如何讓電腦將英文句子自動歸

類，換句話說，也就是如何讓電腦決定某一英文句子是否適合出題，如果是，又適合哪幾種出題題型。我們觀察發現，Sharpe 的分類依據基本上是詞性及結構。因此，Tagger 所提供的資訊，便足以讓我們將分類規則撰寫成出題策略程式，使電腦具備自動將句子歸類的能力。我們使用了能將英文句子自動原型化及標示詞性的工具，將句子以句法上的詞性表達，藉以分析歸類句子與考題的類型。舉例來說，例 8 一句中，透過詞性標注後，會得到如例 9 的詞性。因此，我們就可以知道句中的「have come」的詞性是「hv vbn」的詞性，也就是 have 加上過去分詞的意思，而「to bring」的詞性是「to vb」，可看出此句帶有過去分詞及不定詞。如此一來，系統就能判斷這樣的英文句子，至少可應用在兩種題型，也就是現在完成式和不定詞的觀念。需要特別注意的是，只要句子在處理後，表面型態(surface pattern)符合於某些分類的要求，我們就將其歸類其中，這是因爲無論該句在嚴謹的傳統文法檢驗下，是否真能歸納成該類，在文法出題之時皆不會造成任何負面影響。換句話說，所有表面型態相同的句子，都能以相同的方法出成同樣概念的考題。

（8）I have come to bring you good news.
（9）ppps hv vbn to vb ppps jj nn

3.2 蒐集網路文章作爲試題來源

線上百科全書 Wikipedia 提供包羅萬象的資訊，其中我們選擇該網站每天更新主題的「Today's featured article」部分作爲出題題目來源。在「Today's featured article」中，有許多相關資訊的超連結（見圖四），在此我們稱這些連結爲「母連結」，進入每一個母連結，會有解釋該母連結的相關網頁（見圖四），而在這個網頁中，又包含其他連結，我們稱作「第一層子連結」，而該連結所在的網頁，便稱作「第一層子網頁」，依序重複相同動作，我們會找到「第二層子連結」和「第二層子網頁」。當找到「第二層子連結」後，即停止繼續進入下一層超連結的動作。
當取得「第一層子連結」和「第二層子連結」後，將解釋該連結的「第一層子網頁」和「第二層子網頁」的文章第一個完整句子取出，這是因爲依照百科全書的特性，會於文章首句對某名詞進行解釋，因此，這些句子就是自動出題的試題題目來源。然必須注意的是，在這些句子當中，常會出現一些無意義的句子如「For other definitions of fantasy, see fantasy (psychology). 」，或是因爲超連結的關係，造成乙主題連回甲主題，使得已經處理過與甲主題相關的句子，會被再處理一次，而致使資料庫中出現重複的句子。以上問題則透過人工的檢查解決之，而經過幾天擷取 Wikipedia 句子，我們取得 1373 句子作爲出題的試題。

圖四　Wikipedia 資料取得過程

3.3 題型與題幹的選擇

先前提到我們由 Wikipedia 網頁上搜集了許多完整且有意義的句子，這些資料就是系統的出題來源。每一個搜集來的句子都會經過逐項檢查，判斷其是否適用於某幾種題型，只要有任何符合的題型，我們的系統都會將其出成考題，呈現給系統使用者。

3.4 產生誘答選項

理想的文法考題，不只出題方向要明確，誘答選項的設計也很重要。適當的誘答選項不應該讓受試者太容易發現錯誤，最好能夠似是而非，以達到混淆受試者作答的目的。基於題型的不同，會有不同出題方法的誘答選項，也就是說，並不存在某種能適用於全部題型的誘答選項產生方式，所以每一種題型的誘答選項都是針對其特性而設計的。延用上例，如果「I have come to bring you good news.」一句的出題方向是現在完成式的用法，過去分詞和 have 的搭配使用應該是學習者容易混淆的文法重點，以此爲著眼點設計誘答選項，便可產生如例 10 的文法考題；出題方向如爲不定詞的用法，則重點應是介系詞及動詞型態的變化，由此可以產生如例 11 的文法考題。經過我們的設計，任何一種題型都可以產生三組以上的誘答選項，以符合一般考題四選一的慣例。

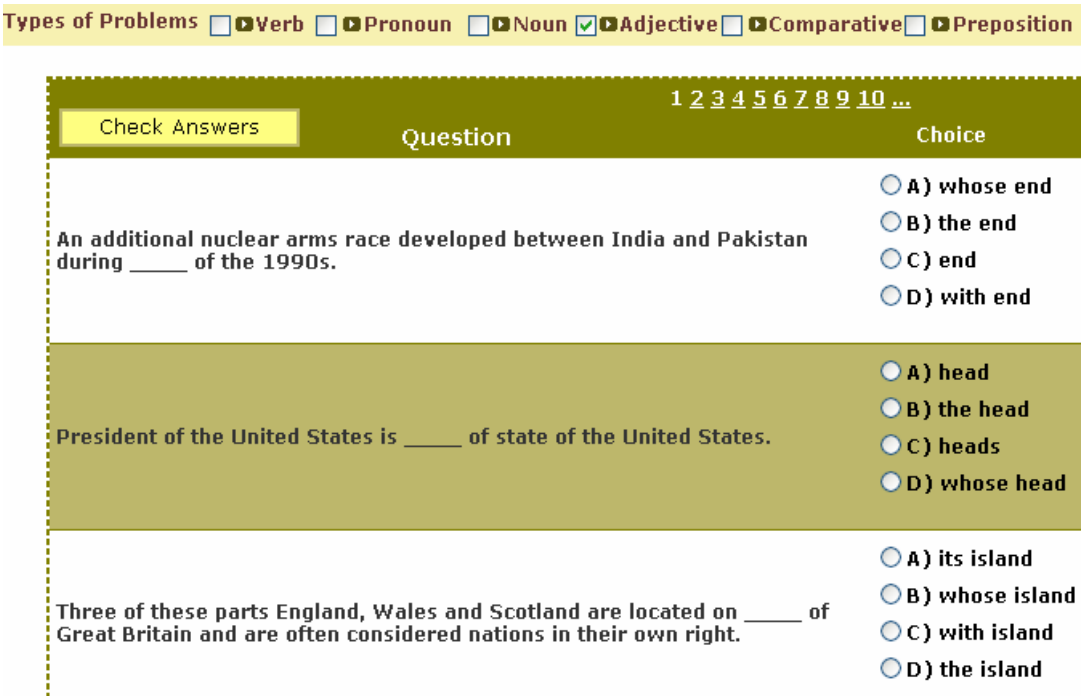
（10）I _____ to bring you good news.
(A) have came
(B) come
(C) have coming
(D) have come

（11）I have come _____ you good news.
(A) at bring
(B) to brought
(C) to bring
(D) for bring

四、FAST 系統介紹

我們的系統 FAST 目前已經上線，使用者可以由以下網址連上：http://140.114.75.15/FASTWebSite/WebForm1.aspx。

網頁的設計以簡單及易於使用爲設計原則，首頁即列出文法的九大類題型供使用者選擇（可複選），只需要簡單的勾選後按下按鈕，符合題型的文法題即刻顯示在螢幕上，如下圖五所示（以形容詞題型爲例，圖五中只顯示視窗的一小部份）：



圖五　FAST 出題介面呈現

由於缺乏受試者，本系統的難易度以及題目的可靠度目前無法進行評估，但我們對於出題策略的使用程度作了統計。我們在網路上共搜集了 497 題四選一題型之托福模擬試題，在這些題目之中，有高達 92% 的題幹也能被 FAST 出成考題，結果共生成了 1171 題文法題，而這些考題中，更有超過 25% 和原來的模擬試題考的是同樣的文法觀念。這顯示了 FAST 的確能自動產生測驗方向正確的文法考題。

另一方面，由 Wikipedia 搜集來的句子裡，共有 71%適用於自動出題。綜合以上所述，我們可以發現五十種出題策略，已經足夠應付考試出題實務上的使用。

五、結論與未來研究方向

本論文未來研究方向著重於三大目標：第一，對於題庫中的句子，我們希望能採取有效的方式，透過自動分析結構或是語義，蒐集出有意義且完整的句子當作考題，來提升系統出題的品質；第二，對於系統中文法題的結構，FAST 的所有出題策略都是獨立的，且可以和系統單獨運作；因此未來若是要增加更多的出題策略，以提高考題的難度，並不會影響到原本的系統，也不必變更已有的策略，將是件省時省力的工作；但目前 FAST 的文法題題庫是離線（offline）先建立好的，這雖然使系統網頁的呈現速度更快，卻必須犧牲掉部份的彈性，亦即資料庫無法即時更新，這一點在未來改進時，是一個著眼點。最後，我們期許 FAST 成爲一套能針對使用者提供難易度學習的系統，使得應用層面更爲寬廣，因此未來需要更多不同背景的受試人員加入（先針對老師、學生兩種），藉由作答的情形，幫助我們統計、分析系統出題的難易度，讓我們對系統的細節作設定。

本論文以自然語言處理爲背景，透過分析英文文法的結構，試圖找出其相關規則，且實作出系統 FAST，可以說是自動出題範疇裡的新嘗試。FAST 的作法非常簡單可行。在實作上，或是未來擴展方面，都能夠輕易實行。網頁介面的設計也盡量簡單化，使用者無須進行多餘的設定，只須選定出題類型，就能得到相當可觀的結果。對英文學習者來說，FAST 就是一個模擬題練習網頁；對英文教學者來說，FAST 則可以大爲節省出題的時間。

參考文獻

1. Brown, James Dean. (1997). Computers in Language Testing: Present Research and Some Future Directions. *Language Learning & Technology, Vol. 1, No. 1, pp. 44-59.* http://polyglot.cal.msu.edu/llt/vol1num1/brown/default.html
2. Coniam, David. (1997) A Preliminary Inquiry Into Using Corpus Word Frequency Data in the Automatic Generation of English Cloze Tests. *CALICO Journal, No 2-4, pp. 15- 33.*
3. Deane, K. Sheehan, Automatic item generation via frame semantics, Education Testing Service (2003): http://www.ets.org/research/dload/ncme03-deane.pdf
4. Dunkel, P. (Ed.). (1991). Computer-assisted language learning and testing: Research issues and practice. New York, NY: Newbury House.
5. Gao, Zhao-Ming. (2000) AWETS: An Automatic Web-Based English Testing System. *In Proceedings of the 8th Conference on Computers in Education/International Conference on Computer-Assisted Instruction ICCE/ICCAI, 2000, Vol. 1, pp. 28-634.*
6. Jun, Da. (2000) Online Language Testing System. http://www.bio.utexas.edu/jun/call/interactive/onlinetest.html
7. Larsen-Freeman, Diane. Grammar and its teaching: challenging the myths. (1997)
8. Liu, Chao-Lin, Wang, Chun-Hung, Gao, Zhao-Ming, and Huang, Shang-Ming. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items, In Proceedings of the Second Workshop on Building Educational Applications Using NLP, pp. 1-8, Ann Arbor, Michigan, 2005.
9. McCormack, Colin, and Jones, David. (1998). *Building a Web-Based Education System. John Wiley.*
10. Milton J. (1998) "WORDPILOT: enabling learners to navigate lexical universes." In S. Granger & J. Hung (eds.) *Proceedings of the International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* 14-16 December 1998. The Chinese University of Hong Kong, 97-98.
11. Ruslan Mitkov, Le An Ha, Computer-Aided Generation of Multiple-Choice Tests. (2003)
12. A. Oranje, Automatic item generation applied to the national assessment of educational progress: Exploring a multilevel structural equation model for categorized variables, Education Testing Service (2003): http://www.ets.org/research/dload/ncme03-andreas.pdf
13. Roever, Carsten. (2001). *Web-Based Language Testing. Language Learning & Technology, Vol.5, No.2, pp. 84-94.* http://llt.msu.edu/vol5num2/roever/default.html
14. Schank, R. C., & Cleary, C. (1995). *Engines for education.* Hillsdale, NJ: Lawrence Erlbaum Associates.
15. Pamela J. Sharpe, "How to Prepare for the TOEFL" 11th ed., Barron's Educational Series, Inc. (2004)
16. Eiichiro SUMITA, Fumiaki SUGAYA, Seiichi Yamamoto. Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In Proceedings of the Second Workshop on Building Educational Applications Using NLP, pp. 61-68, Ann Arbor, Michigan, 2005.
17. Warschauer, Mark. 1996. "Computer-assisted language learning: An introduction". In S. Fotos (ed.), *Multimedia language teaching*. Tokyo: Logos International, pp. 3-20. Also available online. Also available at http://www.gse.uci.edu/markw/call.html
18. Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds.). *Teaching and Language Corpora*. Applied Linguistics and Language Study. London and New York: Longman 1997.
19. Wilson, Eve. (1997) The Automatic Generation of CALL Exercises form General Corpora. *In Wichmann et al. (eds.) Teaching and Language Corpora, pp. 116 – 130. Longman.*
20. 王俊弘，劉昭麟，高照明。利用自然語言處理技術自動產生英文克漏詞試題之研究。（2004）
21. 王俊弘，劉昭麟，高照明。電腦輔助英文字彙出題系統之研究（Toward Computer Assisted Item Generation for English Vocabulary Tests）。（2002）
22. 維基百科 Wikipedia: http://en.wikipedia.org/wiki/Main_Page
23. 機考託福資料倉庫: http://Acezh.com

應用錯誤型態分析於英語發音輔助學習

湯士民 莊則敬 吳宗憲
國立成功大學資訊工程學系
{ming, bala, chwu}@csie.ncku.edu.tw

摘要

語言教學方法主要是由以互動理論 (interactionist theories) 為基礎的溝通式教學法 (communicative language teaching) 所主導。因此，如果要針對學生個別的問題進行糾正，需要甚多的時間，很難採用雙向互動的教學方法。要解決這樣的問題，電腦輔助語言學習系統 (Computer Assisted Language Learning System, CALL) 是個可行的方案。利用語音辨識 (Automatic Speech Recognition, ASR) 技術的電腦輔助發音訓練系統 (Computer Assisted Pronunciation Training, CAPT) 不但可以提供一個沒有壓力的環境，讓學生反覆的練習，同時也能針對學生個別的發音問題，提供回饋與糾正的功能。本論文應用語音辨識、錯誤型態分析、及三維唇型動畫等技術，建立一套適合台灣人之發音輔助教學及矯正系統。本論文的主要技術包括：(1) 利用語音辨識技術，將使用者輸入的語音訊號轉變為音素序列，以進行發音錯誤分析。(2) 針對台灣學生可能的發音錯誤類型建立發音網路，偵測發音錯誤的位置及發音錯誤的型態，並針對錯誤的發音，進而提供適當的糾正。(3) 依據訓練語句之熵值 (entropy) 與使用者的個人發音錯誤類型動態的挑選測試句。(4) 運用影像處理及 3D 動畫的合成，建立一 3D 虛擬人形動畫回饋系統，且特別針對發音時之唇型及舌位，給予使用者正確的發音動作。本論文研發之系統，未來將可提供於本國語者英語發音學習與發音矯正等範疇的實務應用。

1. 緒論

根據美國知名調查機構 IDC 統計，數位學習產業(E-Learning)的全球產值，將從 2003 年的 63 億美元，成長至 2004 年的 230 億美元，每年的複合成長率高達 54%。資策會市場情報中心統計也顯示去年台灣 E-Learning 市場有三・四五億的規模，粗估今年將達到六・二八億。近年來，由於政府對於英語學習的大力推動，使得市面上出現了琳琅滿目的相關書籍、補習班。以目前最熱門的全民英檢為例，不外乎分為聽、說、讀、寫四個部份。然而在「口說」這個部份卻較少有相關的方案可以自我評量。透過語音辨識技術的電腦輔助語言學習系統，使用者不僅可以在一個無壓力的環境下學習，更可以針對個別的發音問題，給予適當的糾正與回饋機制，讓使用者針對個人的錯誤反覆的練習，這不僅節省了人力、時間，同時也可達到較高的學習效果。因此，許多國外的學術單位或者一些商業軟體，都投入不少心力在 CALL System 的開發上。然而，這些市面上的商用軟體多數是套用現有的語音辨識引擎，例如 IBM 的 ViaVoice。而這些引擎原來都是針對母語為英語的使用者而設計的，所以如果針對母語為中文的使用者來說，其辨識率便會有所下降，而無法達到發音教學的目的[1][2]。由於目前大部份的系統針對發音的部份只是給定一個分數，然而我們希望能讓使用者可以得知其發音錯誤的型態，讓使用者知道自己到底發錯成什麼音。因此，本研究利用語音辨識的技術與錯誤型態的分析，建立一套適合台灣人的電腦輔助英語發音學習系統。在偵測發音錯誤類型的部份，首先利用本論文所找出的台灣大學生常犯的發音錯誤型態來建立辨識網路，藉由包含所有可能發音錯誤的辨識網路來找出發音錯誤的部份。且經由測試語句的挑選機制，希望能以較少量的句數歸納出使用者個人的發音錯誤型態。在回饋系統方

面，本論文則運用影像處理及 3D 動畫的合成，建立一 3D 虛擬人形動畫系統，且特別針對發音時之唇型及舌位，給予使用者正確的發音動作。

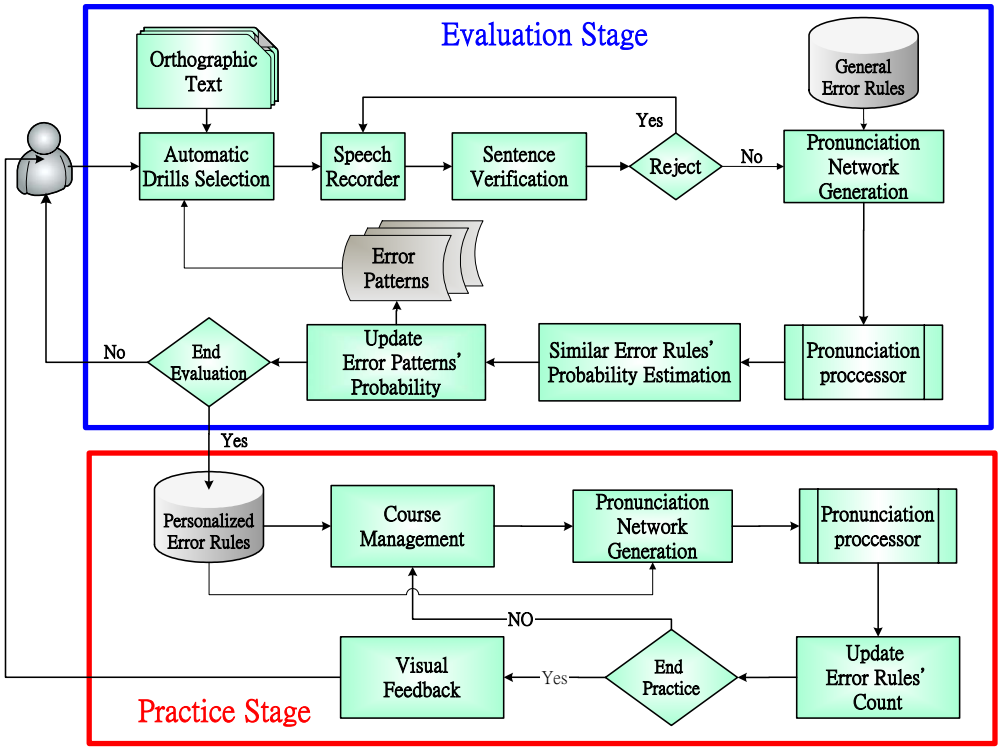
2. 相關研究

近幾年來相關的研究主要可分為發音評分與發音錯誤偵測兩部份。在發音評分的部份有 S.M. Witt 等人在 2000 年提出針對句子中的每個音素做評分，以 likelihood 為基礎的 Goodness of Pronunciation (GOP) [3]。另外，SRI EduSpeak System 則結合了 Phoneme posterior score、Duration score、及 Speech rate 等三種分數來對句子做評分[4]。Seiichi Nakagawa 等人在 2003 年則測試了 Log-likelihood、Likelihood ratio、Best log-likelihood、A posteriori probability、Phoneme recognition rate、Rate of speech 等各種評分方式來對句子評分，最後實驗發現結合 Log-likelihood、Best log-likelihood、Phoneme recognition rate、Rate of speech 與專家所評斷的結果有較高的相關性[5]。在發音錯誤類型的偵測方面，Yasushi Tsubota 等人在 2002 年利用 Pronunciation error network 來偵測日本學生發音錯誤的類型，並利用 LDA 針對發音錯誤的部份作驗證[6]。2004 年 Jong-mi Kim 等人則是根據韓國人的發音習慣建出一些可能的發音錯誤規則，辨識的時候利用這些發音規則來找出使用者發音與正確發音上的差異，並給與適當的建議[7]。

相關的 CALL application 在國外包括 SRI EduSpeak System[8]、ISLE system[9]、以及 PLASER system[10]等。EduSpeak System 主要是利用 speaker adaptation 技術結合 native 與 non-native 的語音，使得系統在辨識率方面有較好的效果。在功能方面主要就是結合 log-posterior score、duration score、及 speech rate 三個分數來對語音作評分。ISLE system 為一個針對義大利與德國人所設計的英語發音學習系統，此系統主要的功能為偵測發音錯誤的位置與發音錯誤的類型。發音錯誤偵測部份主要利用 HMM likelihood 對每個音素來做可信度分析。針對發音錯誤的音素，再利用事先定義好的錯誤規則來偵測錯誤類型。然而，此系統在錯誤類型偵測與回饋的部份效果較不理想。最後，PLASER system 則是針對母語為廣東話的中國人所設計。利用英文與廣東話的語料一起訓練聲學模型，在發音評估的部份則是計算之前所介紹過的 GOP 分數。根據評估的結果，75% 的使用者在使用此系統二至三星期後，在英文發音的正確性上均有所提升。在台灣較知名的 CALL application 有 My English Tutor (My ET)及 Train Speech。My ET 主要是針對發音、能量、音調、節奏四個部份分別給一個分數。並利用一個側面的舌位動畫與一些發音建議來提供使用者正確發音的回饋介面。Train Speech 主要的核心是利用 IBM Via Voice 的語音辨識器針對發音的部份來評分，且根據辨識的結果給使用者一些改正發音的建議。

3. 系統架構

本論文之整體架構，如圖一所示，主要分為“使用者發音錯誤類型評估”與“發音練習與視覺回饋”兩大部份。



圖一：系統流程

3.1 發音錯誤類型評估

這個部份的目的主要是要找出使用者常犯的發音錯誤類型。首先為了避免與標準語音內容差異過大，先針對輸入的語音做內容的驗證。本論文利用 log-posterior score 來對整句語音訊號做可信度分析，若分數小於門檻值則拒絕此語音的輸入。在發音錯誤類型偵測的部份，主要是透過語音辨識的方式，根據人工標記所找出來的發音錯誤規則將所有可能的發音(包含正確發音與錯誤發音)建立成對應的辨識網路，利用這樣的辨識網路來偵測發音錯誤的類型。由於我們希望能以較少的測試句來找出使用者個人的發音錯誤類型，利用計算句子的 entropy 與句子中還需納入考慮音素佔句子的比例來當做句子計分的準則。根據每一次的測試句所辨識出來的結果，我們可以計算出已測試過發音規則的發生機率，然而當測試語料量較少時，尚未出現在測試語料中之發音規則其機率則利用其它相關性較高的發音規則的機率來估計。每經過一次句子的測試，就需針對尚未被念過的句子重新計算其分數，然後挑選分數最大的句子當作下一次的測試句，直到每個音素的機率分布變化量小於我們所設定的門檻值時，即停止測試產生出個人的發音錯誤類型。

3.2 發音練習與視覺回饋

根據找出來的個人發音錯誤類型，我們挑選包含較多使用者常犯的發音錯誤的句子來讓使用者練習。在視覺回饋方面，本論文運用影像處理及 3D 動畫的合成，建立一 3D 虛擬人形動畫系統，且特別針對發音時之唇型及舌位，給予使用者正確的發音動作。運用 3D 虛擬人形動畫系統除了可增加趣味外，使用者也可以經由不同的角度來觀察唇型與舌位的變化。

4. 發音內容驗證

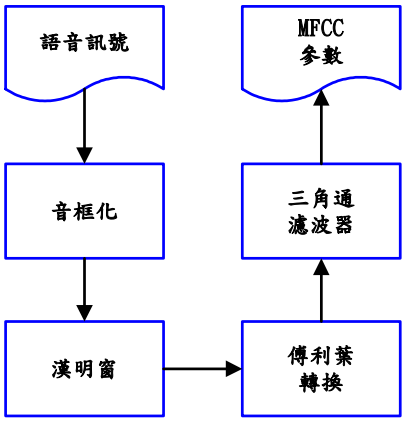
4.1 聲學模型

4.1.1 語料

針對母語為英語的聲學模型，我們使用 TIMIT 語料來訓練聲學模型。TIMIT 內容共 6300 句，由來自美國八個主要口音地區中的 438 位男性、192 位女性所錄製，每人錄製 10 句。我們以 TIMIT 建議的 4620 句做為訓練語料(語料總容量為 440 Megabytes、所有語料長度總和約為 3 小時 49 分 10 秒)來訓練母語為英語的聲學模型。由於本系統是針對母語為中文之使用者，因此我們也同時找了五位英語發音較佳的台灣人，錄製了一套台灣人口音的英語語料，來進行語者的調適。語料內容共 600 句，由 3 位男性、2 位女性大學生所錄製，每人錄製 120 句。

4.1.2 特徵參數擷取

要訓練聲學模型前必須先將訓練資料經過特徵參數的擷取。因此，對於處理語音這種高度差異的訊號時，需要找到能夠具有鑑別度特徵，這裡我們使用三十九維的梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients, MFCC)，包含有十二階的頻譜值加上一階能量值，並取一階微分和二階微分。圖二為特徵參數的擷取流程圖：



圖二：特徵參數的擷取

4.1.3 聲學模型的建立

英文的一個音節是由一或多個音標所組成，每個音標均對應一種發音。TIMIT 語料定義了 62 個聲學模型，然而由於訓練語料的不足以及台灣人發音上準確度較低的情況下，我們不考慮同一個音標在不同位置下的重音情形。因此，我們定義了 42 個聲學模型，包含 40 個 monophones、1 個 silence model 與 1 個 short pause model。我們使用 HTK[10]來訓練聲學模型，表一為我們所定義的 40 個發音模型。

表 1：聲學模型與 KK 音標對照表

模型	kk音標	模型	kk音標	模型	kk音標	模型	kk音標	模型	kk音標
AA	ɑ	ER	ɚ	B	b	K	k	T	t
AE	æ	EY	e	CH	tʃ	L	l	TH	θ
AH	ʌ	IH	ɪ	D	d	M	m	V	v
AO	ɔ	IY	i	DH	ð	N	n	W	w
AW	aʊ	OW	o	F	f	NG	ŋ	Y	j
AX	ə	OY	ɔɪ	G	g	P	p	Z	z
AY	aɪ	UH	ʊ	HH	h	R	r	ZH	ʒ
EH	ɛ	UW	u	JH	dʒ	S	s	SH	ʃ

由於直接拿 TIMIT 訓練的聲學模型來辨識臺灣人口音的英文句，其辨識率便會有所下降，因此必須針對使用者的母語經過適當的調整。我們先使用 TIMIT 語料訓練初始的聲學模型，之後使用自行錄製的臺灣人口音的英文句，利用 MLLR (Maximum Likelihood Linear Regression)[11]調整使用 TIMIT 訓練出來的聲學模型。

4.2 語音內容驗證

在偵測使用者發音錯誤類型之前，我們希望使用者的語音內容與標準語音差異不致於過大，所以必須先針對使用者的語音做一個驗證的動作。我們的驗證機制主要是利用訓練好的 HMM Model，在已知使用者發音內容的情況下，做可信度的分析。

4.2.1 驗證機制

我們參考了兩個可信度分析的方法來建立本系統的驗證機制。其一是使用 LLR(log-likelihood ratio)[12]，然而此方法需要同時訓練 native speaker 與 non-native speaker 的聲學模型，因此需要有較大量的 non-native 語料。現階段受限於收集的台灣人口音語料的不足，於是我們使用 log-posterior probability score [13]來對整句語音做評分。假設 y_t 及 q_i 分別代表輸入語句中第 t 個 frame 的語音參數及其所對應的第 i 個音素。則事後機率 $P(q_i \mid y_t)$ 的計算如下式(1)：

$$P(q_i \mid y_t) = \frac{P(y_t \mid q_i)P(q_i)}{\sum_{j=1}^{M} P(y_t \mid q_j)P(q_j)}. \tag{1}$$

假設所有 model 出現的機率均相等即 $P(q_i) = P(q_j)$，因此上式(1)可近似為下式(2)：

$$P(q_i \mid y_t) = \frac{P(y_t \mid q_i)}{\sum_{j=1}^{M} P(y_t \mid q_j)}, \tag{2}$$

第 i 個音素之 log-posterior probability score ρ_i 便是計算此音素中所有對應 frame 之 log-posterior probability 平均：

$$\rho_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log P(q_i \mid y_t), \tag{3}$$

其中 d_i 為此音素的 frame 總數。最後，整個句子的分數則為所有音素之 log-posterior probability 平均：

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\rho_i , \tag{4}$$

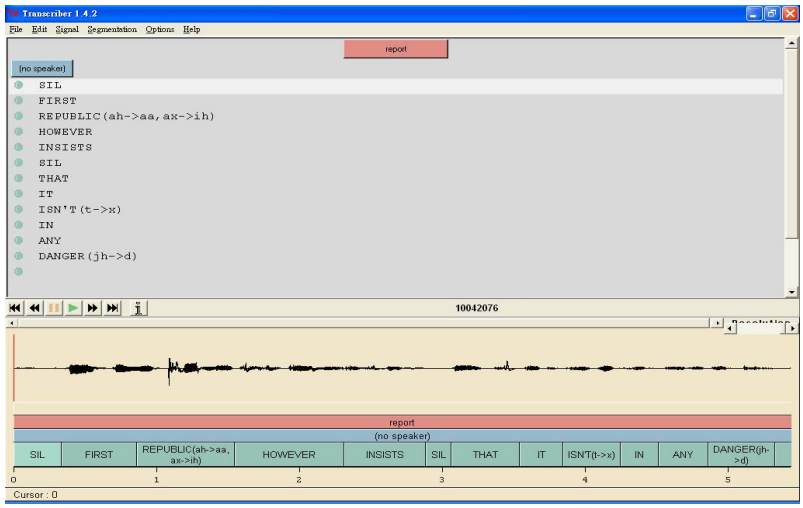
其中 N 為句子中音素的總數。計算出整個句子的分數後，將此分數與預先定好的門檻值比較，若分數小於門檻值，則拒絕此句語音的輸入，請使用者重新再輸入一次語音。相反的，若分數大於門檻值，則接受此句語音的輸入，接著進行發音錯誤類型的偵測。

5. 發音錯誤型態偵測

發音錯誤類型偵測主要是利用語音辨識的技術找出使用者發音錯誤的型態。首先，必須先定義出台灣人常犯的發音錯誤類型，在辨識時考慮此句子所有可能的發音錯誤的型態，建立其對應的辨識網路。透過此辨識網路，利用 Viterbi 演算法找出一條最佳的路徑，偵測出發音錯誤的類型。

5.1 台灣學生常犯之發音錯誤類型

我們從錄製的 2160 句台灣人口音的英文句中，盡可能的使每個音素出現的次數平均的情況下，挑選出了 1000 句英文句，其中包含 35 個男生、65 個女生，50 個非英語系學生、及 50 個英語系學生，語音內容約為 1 小時 6 分。我們將這 1000 句英文句由成大外文系 6 位受過轉寫訓練的學生做發音錯誤的標記。下圖是標記程式之介面[14]：



圖三：標記程式介面

根據標記的結果，我們整理出較常犯的錯誤類型。主要分為以下兩類：

A. 字轉音錯誤

這類型的錯誤主要由於字母拼字的關係，導致將英文字母轉成音標時發生錯誤。例如: cr**i**sis /k r aɪ s I s/，這個單字由於在字母上的拼字是 i，因此容易導致將/aɪ/這個發音念成/I/。下表列出幾個較常出現的錯誤類型：

表 2：字轉音錯誤

錯誤型態	Example
/ɑ/ -> /o/	T**o**m、J**o**hn
/z/ -> /s/	day**s**、hu**s**band
/ɔ/ -> /ɑ/	wr**o**ng、c**o**rporate
/aɪ/ -> /I/	cr**i**sis、d**i**versify
/æ/ -> /ɑ/	st**a**ff、**a**s

B. 發音錯誤

此類型的錯誤主要是由於母語的影響，導致發音的不正確。例如:full /f ʊ l/，/ʊ/這個音容易被念成/u/。這個發音錯誤主要是因為中文的母音並沒有長短之分，因此容易造成這類的錯誤。以下幾分別為母音與子音較常犯錯的類型：

表 3：母音發音錯誤型態

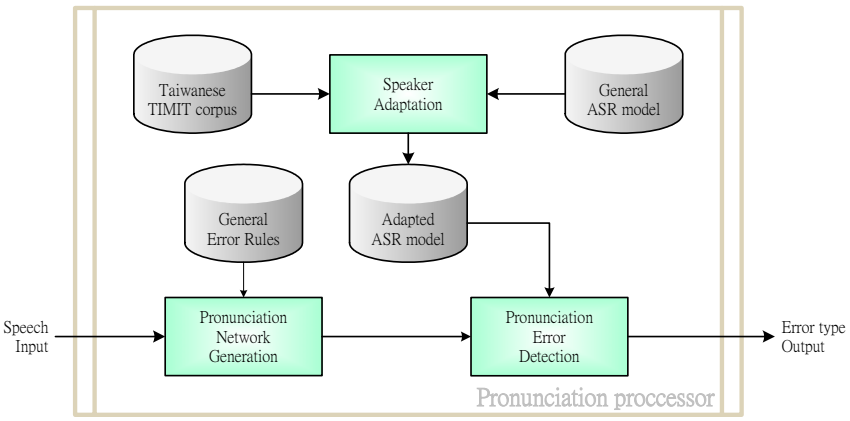
	錯誤型態	Example
短母音替代長母音	/i/ -> /ɪ/	seat、need
	/e/ -> /ɛ/	taken、made
	/u/ -> /ʊ/	fool
	/o/ -> /ɔ/	gold
長母音替代短母音	/ɪ/ -> /i/	year
	/ɛ/ -> /e/	weather、next
	/ʊ/ -> /u/	full、good
	/ɔ/ -> /o/	offer
/ɛ/替代/æ/	/æ/ -> /ɛ/	pan、matter
/ɑ/替代/ʌ/	/ʌ/ -> /ɑ/	husband、funny
非捲舌音替代捲音	/ɚ/ -> /ə/	either

表 4：子音發音錯誤型態

	錯誤型態	Example
非捲舌音替代捲舌音	/θ/ -> /s/	thank、think
	/ð/ -> /l/ or /d/	this、them
/ə/**替代節尾/r/**	/r/ -> /ə/	there、clear
/n/替代/ŋ/	/ŋ/ -> /n/	going
母音後面/r/省略	/r/ -> x	are、warm
母音後面/l/省略	/l/ -> x	almost、goal
音節節尾/n/省略	/n/ -> x	mine、one
停頓音節尾省略	/d/ -> x	stupid
	/t/ -> x	brought
	/k/ -> x	think
停頓音後增加/ə/	/d/ -> /də/	stupid
	/t/ -> /tə/	student
	/k/ -> /kə/	link

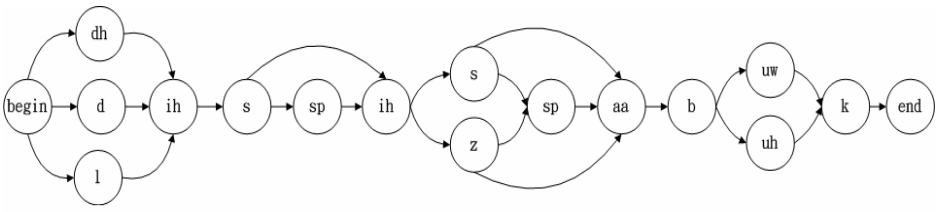
5.2 發音錯誤類型的偵測

發音錯誤的偵測主要是先針對句子建立對應的辨識網路，利用此辨識網路來辨識語音內容。其流程圖如圖四：



圖四：發音錯誤偵測流程

根據上一節介紹的發音錯誤類型，我們可以將所有可能的發音建立在辨識網路中(包括發音正確與所有可能的發音錯誤)，例如: This is a book 考慮所有發音的可能，其辨識網路如圖五：

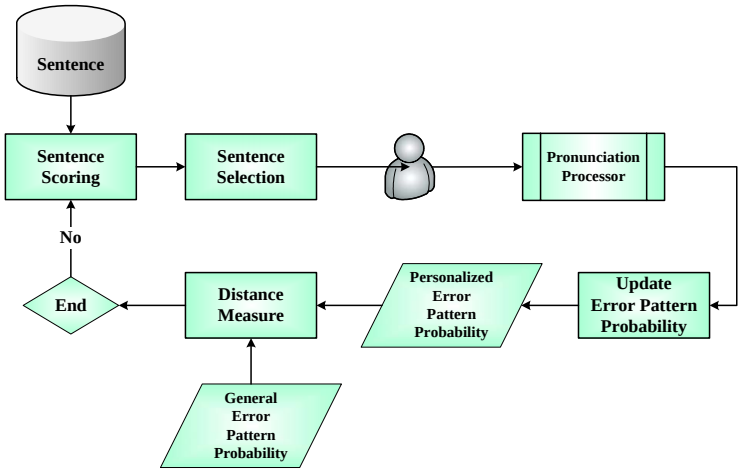


圖五：發音辨識網路

有了辨識網路後，利用 Viterbi 演算法找出一條最佳的路徑，將辨識結果與已知的語音內容比對，如此就能找出使用者該句發音錯誤的型態。

6. 最佳訓練資料之選取

最佳訓練資料選取之流程圖如圖六所示。



圖六：最佳訓練資料選取流程

在此我們希望能以最少量的測試句來找出使用者個人的發音錯誤型態。首先，針對資料庫中的測試句分別給予計分，挑選出分數最高的句子作為測試句。根據找出的錯誤型態，更新個人錯誤型態的發生機率。針對每個可能產生發音錯誤的音素，比對此音素與大量資料所統計出來的錯誤型態中機率分布的差異，倘若在連續兩測試句中所計算出的變動量已小於某個門檻值，則表示針對

這個音素使用者的發音錯誤機率已經達到一個穩定的狀態，所以在挑選下一句測試句時，便不需將此音素列入計分的考量中。依照這樣的流程，反覆的挑選測試句直到所有的音素均已不需再考量為止。

6.1 訓練語句之計分與挑選

本節首先介紹訓練語句的計分方式。對於語料庫中第 *i* 句訓練句，其 sentence score 計算方式如下式：

$$Sentece\_Score_i = ES_i \times TS_i, \tag{5}$$

其中 ES_i 表示第 *i* 句訓練句的 Entropy，計算方式如下：

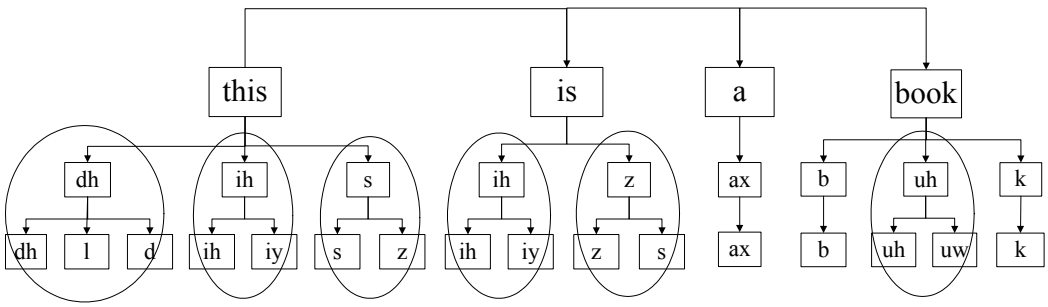
$$ES_i = -\sum_{j=1}^{m}\sum_{k=1}^{n} P_{jk}\log\left(P_{jk}\right), \tag{6}$$

m ：句子中可能發生錯誤的 phone 個數

n ：第 *j* 個 phone 中所有可能出現的情形

P_{jk} ：第 *j* 個 phone 的第 *k* 個情況的機率值

例如一訓練句 “this is a book”，其可能的發音型態如下圖所示：



圖七：句子發音型態範例

由這個例子可以得知，可能錯誤的音素個數有 6 個，因此分別對這 6 個可能錯誤的音素計算其 entropy，將此 6 個的 entropy 總和當成這整句的 ES_i 分數。我們使用 entropy 的目的在於找出最不能確定使用者會念對或會發生發音錯誤的句子。比如說以 “ih” 這個音素為例，當 “ih 念對成 ih” 與 “ih 念錯成 iy” 機率均為 0.5 的情況下，其 entropy 的值為最大，即表示我們較難以判定此音素是否較易被念錯或較易被念對。相反的當 “ih 念對成 ih” 與 “ih 念錯成 iy” 機率有一方為 1 的情況下，其 entropy 的值為最小，即表示我們可以判定此音素易被念錯或易被念對。基於上述的理由，我們考慮 entropy 越高的句子越優先讓使用者測試。

除了 ES_i 值外，第(5)式另一個變數為 TS_i。TS_i 表示計算句子中還需要納入計分考慮的音素佔句子的比例。其計算方式如下：

$$TS_i = \frac{NTC_i}{TCi}, \tag{7}$$

NTC_i：句子中尚需考慮的 phone 個數

TC_i ：句子中所有可能發生錯誤的 phone 個數

最後，根據我們人工標計發音錯誤的句子中，我們可以計算出每個音素的正確與錯誤發生機率。因此，我們將這些機率當做 general model。當使用者做測試時，我們也能根據測試結果計算出每個音素的正確與錯誤發生機率，之後利用 discrete KL distance 的方式來計算目前某個音素的機率

分布與 general model 的差距，假設累積到第 i+1 句測試句其與 general model 的差距相較於累積到第 i 句測試句其與 general model 的差距變動不大時，表示對於此音素而言，使用者發生對或錯的機率已趨於穩定，因此我們可以不需再將此音素納入句子挑選的考慮。以下將介紹對於某音素停止條件的計算方式。

若某個音素 a 有 N 種可能的唸法：$a_1 \sim a_N$，則定義 $p_{general}(a \to a_n)$ 表示在 general model 中音素 a 被唸成 a_n 的機率；$p_1^i(a \to a_n)$ 則表示在累積到第 i 句測試句時，音素 a 被唸成 a_n 的機率。則 KL distance 的計算如下：

$$KL_i = \sum_{n=1}^{N} p_1^i(a \to a_n) \log_2 \left(\frac{p_1^i(a \to a_n)}{p_{general}(a \to a_n)} \right), \tag{8}$$

$$KL_{i+1} = \sum_{n=1}^{N} p_1^{i+1}(a \to a_n) \log_2 \left(\frac{p_1^{i+1}(a \to a_n)}{p_{general}(a \to a_n)} \right), \tag{9}$$

上式中 KL_i 為累積到第 i 句與 general model 的差距，而 KL_{i+1} 則為累積到第 i+1 句與 general model 的差距。因此我們是以下式來做為停止考慮的參考標準：

$$\Delta KL = \left| KL_{i+1} - KL_i \right|, \tag{10}$$

當 ΔKL 小於某個 threshold 且此音素已被念過的次數大於五次以上時，即可停止考慮此音素。利用上述的計分方式，我們挑選出分數最高的句子當做下一次的測試語句，每經過一次測試均需根據使用者發音錯誤的結果，重新計算還未被測試句子的分數。因此，針對不同使用者之間發音錯誤類型的不同，同一測試句的分數變會有所不同。如此，便能依據個人化的發音錯誤習慣，挑選的句子變會有所差異。

6.2 發音錯誤類型機率的估計

由於使用者在測試的過程中，當某些音素在測試資料中尚未出現時，我們希望利用估計的方式來計算出其發生機率。因此，我們假設某些發音習慣會導致類似的錯誤發生。從我們人工標記的發音錯誤類型中，我們發現當某個測試者發生了/ð/念錯成/l/時，/ɚ/也容易被念錯成/ə/。由聲學的角度來看，這類的錯誤可能是由於捲舌音發的不好，導致念錯成非捲舌音。再舉例來說，由於在中文的在母音的部份沒有長短之分，所以長母音念錯成短母音的情況也容易同時出現。因此，從我們人工標記的發音錯誤的語料中，利用計算 Mutual Information 的方式找出不同發音規則間的關係(例如：/ð/→/l/表示一種發音規則)。假設 X、Y 分別代表兩個不同的發音規則，其 Mutual Information 計算方式如下：

$$I(X;Y) = \sum_X \sum_Y P(x_i, y_i) \log \frac{P(x_i, y_i)}{P(x_i) P(y_i)}, \tag{11}$$

下表列出幾個 Mutual Information 較高的發音規則：

表 5：相關性較高之發音規則

Rule X	Rule Y
/i/ -> /ɪ/	/e/ -> /ɛ/
/e/ -> /ɛ/	/o/ -> /ɔ/
/u/ -> /ʊ/	/o/ -> /ɔ/
/ð/ -> /l/	/ɚ/ -> /ə/
/o/ -> /o/	/e/ -> /e/
/e/ -> /e/	/u/ -> /u/
/l/ -> x	/r/ -> x
/k/ -> /kə/	/t/ -> /tə/
/æ/ -> /ɛ/	/l/ -> x
/ð/ -> /l/	/i/ -> /ɪ/

利用我們所設定的門檻值，我們從 80 個發音規則中(包含正確的發音規則)，找出了 53 組相關性較高的發音規則。由上述的例子中看出，大部份的相關性較高的發音規則可符合聲學方面的特性，然而因為是利用統計的方式，所以有些找出來的發音規則無法從聲學的角度來解釋。

經由上述的方式找出相關性較高的發音規則後，我們可以藉由下列的方式估計出機率值：

$$P(X,Y)=P(X\mid Y)P(Y)=P(Y\mid X)P(X), \tag{12}$$

$$P(X)=\frac{P(X\mid Y)}{P(Y\mid X)}P(Y) \quad , \quad P(Y)=\frac{P(Y\mid X)}{P(X\mid Y)}P(X) \quad , \tag{13}$$

其中 $P(X\mid Y)$ 與 $P(Y\mid X)$ 我們事先從經過人工標記過的大量語料中訓練出來，因此當使用者在測試語料中只出現 X 或 Y 其中之一，就可藉由上述的方式估算出另一方發生的機率。

假設與 X 相關的較高的發音規則有：$R_1, R_2, R_3 \ldots R_n$，因此 X 出現機率的計算方式如下：

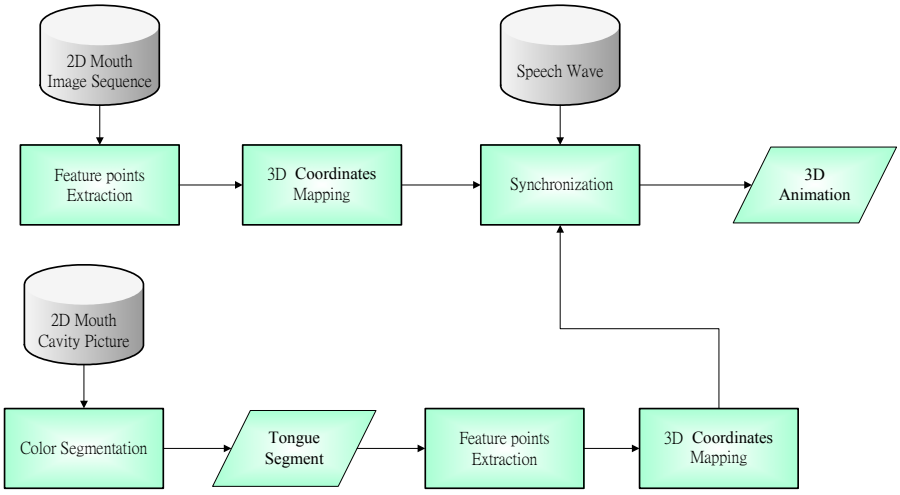
$$P(X)\approx\sum_{i=1}^{n}W_i\times\frac{\bar{P}(X\mid R_i)}{\bar{P}(R_i\mid X)}\times P(R_i), \tag{14}$$

$$W_i=\frac{\bar{P}(X,R_i)}{\sum_{i=1}^{n}\bar{P}(X,R_i)}, \tag{15}$$

上述的權重值也可事先從經過人工標記過的大量語料中訓練出來，因此，我們假設尚未出現在測試語料中的發音規則間的相關性與大量資料所訓練出來的 joint probability 是不變的，所以我們可以藉由這樣的方式估算出尚未出現的發音規則機率。

7. 視覺回饋

我們運用影像處理及 3D 動畫的合成，建立一 3D 虛擬人形動畫系統，且特別針對發音時之唇型及舌位，給予使用者正確的發音動作。圖八是 3D 動畫合成之流程：

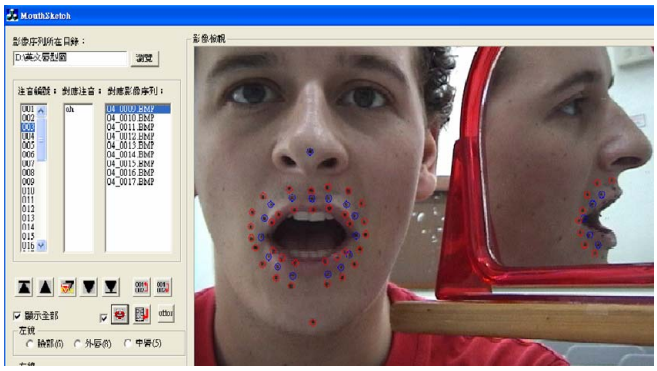


圖八：三維動畫合成

7.1 三維唇型動畫

7.1.1 唇型特徵點擷取

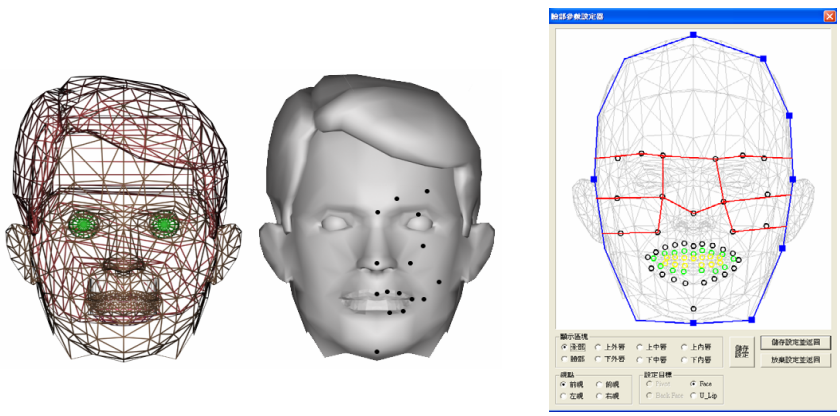
首先經由一組事先拍攝的唇型變化影片來擷取出 62 個唇形特徵點變化參數。因此，根據我們所定義的 40 個聲學模型，分別拍攝其發音之唇型影帶。接著利用 Optical Flow[15]動態偵測的方式自動偵測唇形週圍幾個特徵點的變化。圖九為唇型特徵點偵測之結果：



圖九：唇型特徵點偵測

7.1.2 三維座標轉換

擷取出唇形特徵點在三個座標軸中的位移之後，我們必須先在 3D 模型中，定義出此 62 個特徵點的位置，其界面如圖十所示。



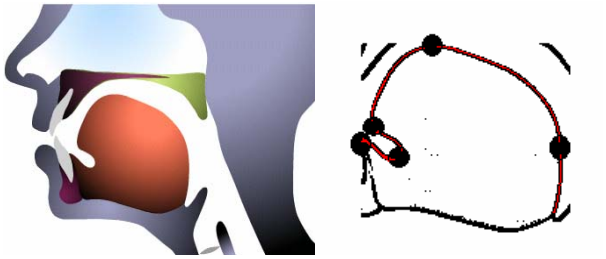
圖十：三維唇型控制點定義

其餘網格中的點則由鄰近的控制點來控制，其位移量為鄰近控制點位移量乘以個別的權重之總和，且控制點的權重與控制點到網格點距離的平方成反比。

7.2 三維舌位動畫

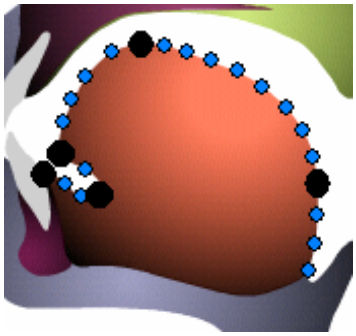
7.2.1 舌頭特徵點擷取

由於直接在舌頭上貼 Sensor 來偵測發音時舌位的變化是不容易的，因此我們實作一非侵入式 3D 舌位測量方法。首先我們由網路上的開放資源中蒐集了每個發音的口腔 2D 圖(如圖十一) [16]，由於要找出舌頭的部份，因此先將顏色由 RGB 轉換為 HSI 後，便可很容易的將舌頭的部份給切割出來。

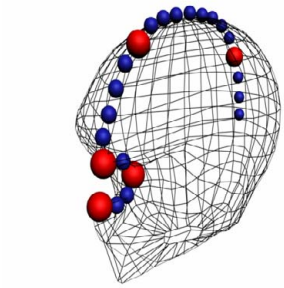


圖十一：發音口腔圖

由口腔的剖面圖我們發現通常轉折的地方(舌尖、舌根)變化較大，因此我們將這幾個轉折的地方定義為舌頭主要的特徵點(如圖十一中的黑色點)。根據不同的發音，我們必需記錄每個特徵點在不同時間下的位移量，所以利用影像處理的技術，自動偵測出舌頭的轉折點。首先將舌位圖經由 sobel operator 做邊界偵測，然後利用八鄰域的方式做邊界追蹤以擷取出特徵點的位置。為了讓動畫可以更精細，除了 5 個主要轉折點外我們還擷取了其他的特徵點(如圖十二)。



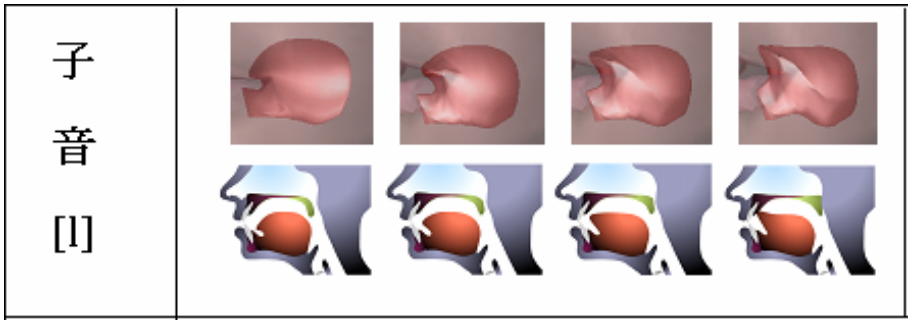
圖十二：舌頭特徵點偵側結果



圖十三：三維舌頭控制點定義

7.2.2 三維座標轉換

有了每個發音時的舌位特徵點位移變化量，接下來需將 2D 舌頭特徵點座標 map 到 3D 空間，並利用特徵點自動計算出 3D 模型網格點位移量。利用這些控制點座標與各控制點對相鄰網格點的權重(與網格點的距離成反比)，便可計算出 3D 模型中每個時間所有網格點的位移量。圖十四為一實作之舌頭 3D 動畫：



圖十四：舌位動畫與發音口腔圖比較

8. 實驗結果與討論

8.1 語音訊號切割實驗

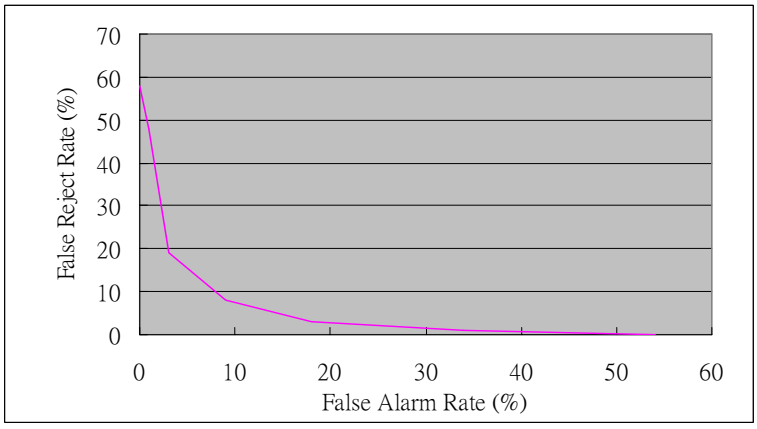
我們從所錄製的臺灣人口音之語音訊號中，挑選其中發音正確的 300 句語音來做訊號切割實驗，語音總長約 21 分 36 秒。平均每句訓練句包含 6.87 個 Word、35.26 個 phoneme。由於此 300 句有經過人工標記的動作，因此在我們可直接將利用 Forced Alignment 切割出來的時間點和人工標音出來的結果作比較。由於人工標音的部份只有標記到字，所以我們假設若切割出來字的時間區段和該字在人工標音下的時間區段前、後各相差在 0.1 秒以內，則稱此字的切割結果為正確。在此我們比較兩個不同的聲學模型：一個為使用 Native Speaker 所訓練出來的模型，另一個為使用台灣口音之英文語料經由語者調適所產生的模型。表 6 為英文語音訊號切割的正確率：

表 6：語音訊號切割正確率

模型 / 正確率	Model without Adaptation	Model with Adaptation
Word 時間正確率	84.63 %	87.93 %

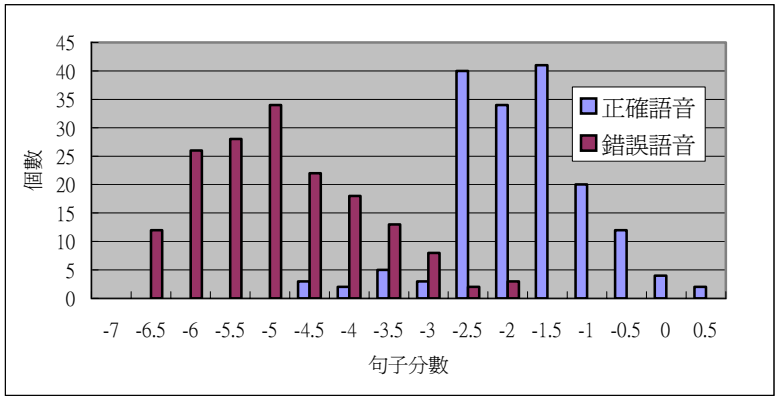
8.2 語音內容驗證實驗

將輸入的語音訊號做可信度的分析後，若分數高於門檻值則表示此輸入語音是可靠的，反之若小於門檻值，則拒絕此語音進入系統。因此，為了找到較佳的門檻值，我們取 166 句語音當作標準語音內容，語音的總長度約為 12 分 42 秒。此外取另外 166 跟標準語音內容不同的語音當作錯誤語音，錯誤語音總長度約為 10 分 38 秒。我們利用正確語料被拒絕(False Rejection, Type Ⅰ Error)及錯誤語料被接受(False Alarm, Type Ⅱ Error)的 ROC 關係圖(Receiver Operator Characteristic)來找出最佳的門檻值：



圖十五：False Rejection Rate 與 False Alarm Rate 之 ROC

我們利用正確拒絕加錯誤接受的和最小來找出門檻值。因此，當門檻值為-3.2 時有最佳的結果，其正確接受率為 93.4%，正確拒絕率為 6.6%，錯誤拒絕率為 95.2%，錯誤接受率為 4.8%。經由以上的實驗訂出門檻值之後，我們使用另外的 166 句正確語音與 166 句錯誤語音作測試，正確接受率為 92.2%、正確拒絕率為 7.8%、錯誤拒絕率為 96.9%、錯誤接受率則為 3.1%。圖十六為正確語音與錯誤語音經由計算可信度後的分數分布圖：



圖十六：語音內容驗證結果分布圖

8.3 發音錯誤偵測實驗

我們從所錄製的臺灣人口音之語音訊號中，挑選其中的 300 句語音來做發音錯誤偵測實驗，語音總長約 22 分 33 秒。將利用發音網路所辨識出來的結果與人工標記的答案比較來計算正確率，結果如下表所示：

表 7：發音錯誤偵測實驗結果

模型 / 正確率	Model without Adaptation	Model with Adaptation
Phoneme 正確率	72.59 %	78.18 %

8.4 最佳訓練語句挑選評估

目前我們所使用的測試句共 166 句，我們找了 5 位測試者每人均唸完所有的 166 句。以下針對三個不同的挑選句子方法做評估：

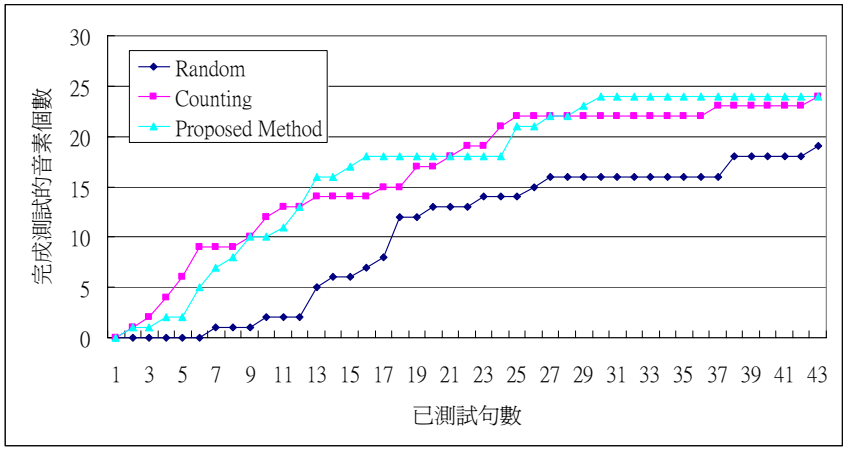
Random：隨機挑選測試語句

Counting：只考慮句子中尚需納入挑選的音素總數

Proposed Method：本論文所提出的句子計分方式

8.4.1 句子總數的評估

所需測試的音素總數共 24 個，根據我們所定義的音素完成測試的條件下，比較此三種方式所需測試的句子總數。

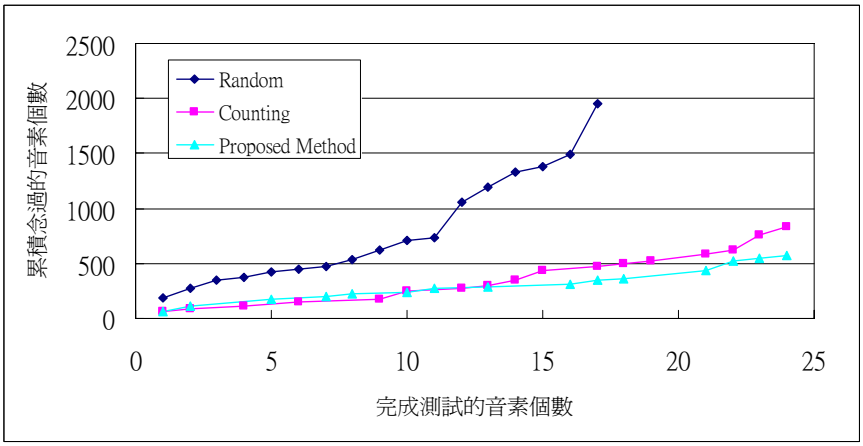


圖十七：句數評估實驗結果

由實驗結果可知，我們所提出的方法明顯比用 Random 的方式所需的句數較少。除此之外，與 Counting 比較之下雖然在完成 13 個音素測試前 Counting 所需句子數較少，然而比較所有完音素均完成測試時所需的句子數，我們所提出的方法需要 30 句，相較 Counting 所需的 43 句，所需的總句數較少。由以上的實驗結果顯示，在完成所有的音素測試下，我們所提出的方法有較佳的結果。

8.4.2 累積唸過音素總數的評估

所需測試的音素總數共 24 個，根據我們所定義的音素完成測試的條件下，比較此三種挑選句子的方式下，所累積唸過的音素總數。



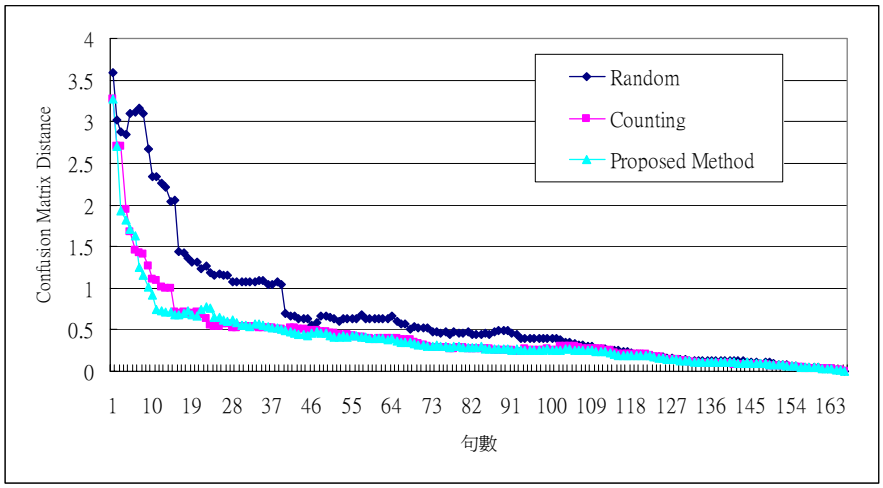
圖十八：音素總數實驗結果

由實驗結果可得知，我們所提出的方法明顯比 Random 所累積唸過的音素個數較少。在完成 10 個音素測試後，Counting 所累積唸過的音素個數為 251 個，我們所提出的方法所累積唸過的音素個數只有 239 個。在此之後，我們所提出的方法均較 Counting 所需唸過的音素總數少。當所有的音素均完成測試的情況下，Random 所累積唸過的音素個數為 1956 個，Counting 所累積唸過的音素個數為 829 個，我們所提出的方法為 570 個。由以上的實驗結果顯示，我們所提出的方法句子的平均長度相較其他兩個方法短。

8.4.3 錯誤型態機率的評估

我們由使用者所唸的 166 個句子中，可以計算出所有發音規則的出現機率。因此，我們可以建立一個 40 乘 40 的 Confusion Matrix。在此我們比較三種不同的挑選句子的方式，觀察其 Confusion

Matrix 在每經過一句測試資料後，我們將 Confusion Matrix 看成一個維度為 1600 的 vector，利用簡單的 Euclidean Distance 的方式計算與 166 句所計算出來的 Confusion Matrix 距離。如圖十九所示：



圖十九：使用三種不同方式時 Confusion Matrix 收斂結果

由圖中我們可以看出，用隨機的方式挑選句子 Confusion Matrix 的機率值收斂的最慢。運用我們所提出的挑句子的方式，可以發現少量的句子下，Confusion Matrix 機率值收斂的較快速，可以較快逼近接近真正的機率值。

9. 結論與未來展望

本論文提出一個以分析語者發音錯誤型態來輔助英語學習之方法。根據我們所錄製的台灣人口音之英文句，我們統計出了台灣大學生在英語發音上較常見的發音錯誤類型。我們利用包含可能的發音錯誤類型所建立的發音網路來偵測使用者發音錯誤的型態，且利用統計方法依據訓練語句之熵值(entropy)與使用者的發音錯誤類型動態的挑選測試句。由實驗中我們可以發現所提之方法可以有效降低訓練語句之數量，提高學習者之學習成效，充份顯示本論文所提之方法在實用上具有一定之效果。除此之外，在回饋系統的部份，我們實作了一個 3D 虛擬人物動畫，透過這樣的 3D 動畫能夠讓使用者以多個不同的角度來觀察發音時唇型與舌位動作，更可清楚的呈現發音的完整過程。在未來研究方向方面，可以從以下幾個地方來著手：(1) 自動新增錯誤型態：目前由於我們的錯誤類型是事先定義好的，因此無法動態偵測出不在定義中的發音錯誤。因此，若能根據系統不斷的使用的過程中，自動新增出一些個人化的發音錯誤類型，如此便能更有效的改正使用者發音錯誤。(2) 英文發音錯誤類型與中文發音的關係：由本論文所找出來的發音錯誤類型中，不難發現之所以會導致發音錯誤，其實與使用者本身的母語有一定的關係存在。因此，若能分析出中文與英文在子母音上的異同，便可更有效的從中文的發音習慣上來給予使用者較好的發音建議。(3) 錯誤發音的糾正：目前的系統在這個部份的一直沒有較好的成效。若能從聲學的角度或母語發音上的習慣來糾正錯誤發音，藉此建立一套更好的發音錯誤糾正的機制。

參考文獻

[1]. T. M. J. Munro and M. Carbonaro, “Does Popular Speech Recognition Software Work with ESL Speech?”, TESQL Quarterly 34, pp.592-603, 2000

[2]. D. Coniam, “Voice Recognition Software Accuracy with Second Language Speakers of English”, System 27, pp.49-64, 1999

[3]. Witt,S.M. and Young, S.J. “Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning”, Speech Communication 30, 95-108. 2000

[4]. Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., and Butzberger, J. “The SRI EduSpeak(TM) System: Recognition and Pronunciation Scoring for Language Learning", Proceedings of InSTILL 2000, Dundee, Scotland, 123-128. , 2000

[5]. Seiichi Nakagawa, Kazumasa Mori, Naoki Nakamura “A Statistical Method of Evaluation Pronunciation Proficiency for English Words Spoken by Japanese”, Eurospeech 2003

[6]. Yasushi Tsubota, Tatsuya Kawahara, Masatake Dantsuji “CALL System for Japanese Students of English Using Pronunciation Error Prediction and Formant Structure Estimation” InSTILL 2002

[7]. Jong-mi Kim, Chao Wang, Mitchell Peabody, Stephanie Seneff “An Interactive English Pronunciation Dictionary for Korean Learners” ICSLP 2004

[8]. Menzel, W., Herron, D., Bonaventura, P., and Morton, R. (2000). “Automatic detection and correction of non-native English pronunciations”, Proceedings of InSTILL 2000, Dundee, Scotland,

[9]. Mak, B., Siu, M.H., Ng, M., Tam, Y.C., Chan, Y.C., Chan, K.W., Leung, K.Y., Ho, S., Chong, F.H., Wong, J., Lo, J. (2003). “PLASER: Pronunciation Learning via Automatic Speech Recognition”, Proceedings of HLT-NAACL 2003, Edmonton, Canada, 23-29

[10]. Steve Young, *The HTK Book version 3*, Microsoft Corporation, 2000

[11]. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models C. J. Leggetter and P. C. Woodland, Computer Speech and Language (1995) 9, 171–185

[12]. Franco, H., Neumeyer, L., Ramos, M., and Bratt, H. (1999). “Automatic Detection of Phone-Level Mispronunciation for Language Learning”, Proceedings Eurospeech '99, Budapest, Hungary, 851-854.

[13]. H. Franco, L. Neumeyer, and Y. Kim, “Automatic Pronunciation Scoring for Language Instruction”, Proc. ICASSP , pp.1471-1474, 1997

[14]. Transcriber: Development and use of a tool for assisting speech corpora production Claude Barras , Edouard Geoffrois , Zhibiao Wu , Mark Liberman , speech communication 2001

[15]. Horn, B.K.P and Schunck, B.G., “Determining Optical Flow”, Artificial Intelligence, vol.17, nos.1-3, pp.185-203 (1981-8).

[16]. http://www.uiowa.edu/~acadtech/phonetics/english/frameset.html

使用韻律階層及大量詞彙的中文文轉音系統

A Mandarin Text-to-Speech System Using Prosodic Hierarchy and a Large Number of Words

余明興、張唐瑜、許燦煌、蔡育和
國立中興大學資訊科學所
msyu@dragon.nchu.edu.tw, s9256047@cs.nchu.edu.tw, s9256040@cs.nchu.edu.tw, s9256013@cs.nchu.edu.tw

摘要

本論文實作了一個中文的文轉音系統(Test-to-Speech)系統，它使用大量的詞彙來做爲合成單元(Synthesis units)，並且配上適當的韻律階層。韻律階層可以使語意更加清晰，也可以幫助選取適當的合成單元。因此本篇論文主要包含兩個重點：韻律階層的求取和以大量詞彙作爲合成單元的架構，在韻律階層的求取上，我們實驗了利用剖析器爲基礎的方法以及著名的統計式方法-CART(Classification And Regression Trees)來進行求取。我們使用大量詞彙來當成我們的合成單元，可以免去許多語音處理不易做好的連音處理。我們也利用韻律預估模組所得到的參數，進行音量和音長的調整。最後我們完成一套包含 12224 個二字詞以及 2690 個三字詞的中文文轉音系統，並開放於線上試用。

關鍵字：Text-to-Speech, Parser, Prosodic Hierarchy.

1.緒論

1.1.中文文轉音系統

近年來文轉音系統在實作上，最常見到的爲波形拼接法（waveform- concatenation）。這種作法主要是利用預先錄製好的聲音，稱之爲合成單元（synthesis units），存放在語音資料庫中，要用時再將其取出拼接，來合成所要讀出的語句。這些合成單元要能包含所有可能的發音，這些預錄的單元可能是音素(phoneme)、雙音素(di-phone)、音節(syllable) …等。

傳統的 TTS 包含三個模組：（一）文句分析(Text analysis)：這部份包含斷詞以及一些語言知識上的標記，例如詞類（Part-of-Speech, POS）等，另外也會處理字轉音。（二）韻律預估(Prosody prediction)：預估合成音的音長（Duration）、音量（Energy）、音高(Pitch)等聲學參數。（三）語音合成(Speech generation)：利用已經預估好的韻律參數來進行韻律的調整。常見的方法，有 PSOLA（Pitch-Synchronous Overlap and Add）[9]…等。傳統的架構發音不自然，主要的問題[8]爲（一）連音無法獲得充分解決，（二）韻律調整範圍過大。因爲早期記憶體的限制，錄製大量語音來做爲合成單元非常困難，所以合成語音品質遲遲無法突破。

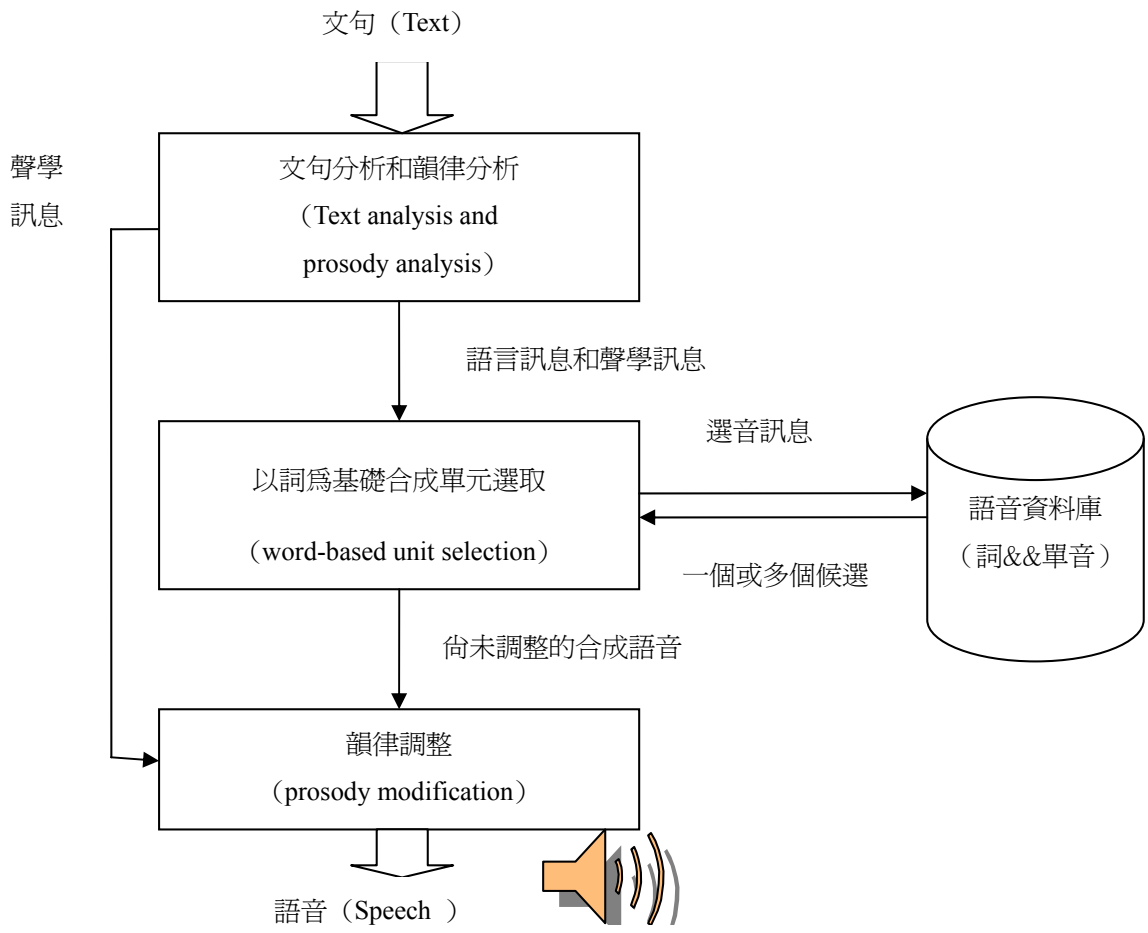
近年來，文轉音(Text-to-Speech)系統漸漸以 Corpus-based 的架構來實作[6][9]，並以波形拼接來做爲合成方法。該方法的優點在於可以得到不錯的語音品質，相較於傳統以音節爲合成單元的架構，更容易被聽者接受。這種架構的特徵在於：（一）會錄製大量語料，（二）盡可能不做訊號處理，（三）選擇適當的合成單元。軟體界的巨人－微軟，所推出的「木蘭」雙語(中文以及英語)TTS

系統[6]，便是使用 Corpus-based 的方式所組成的。錄音的語料爲錄製句子(sentence)。在連續音語料上採用一個階層式的韻律模型。最後再利用 Decision Tree[6][9]的方法來找尋適當的非固定長度合成單元來進行拼接，過程中，完全不做訊號處理。

本論文所提出的系統，使用大量的詞彙來做爲合成單元，並配合適當的語音處理。我們認爲從詞(主要是二字詞和三字詞)中抽取出來的合成單元，在音程上較爲完整，聽的較清楚。而且當我們用許多的詞來做爲合成單元時，大部分的連音現象都已包含在合成單元中，而連音是語音合成的各種處理中非常難以做好的部份。語音合成中的音量調整只要不超出位元數的最大音量限制，並不會影響聲音的品質，所以可以做較大程度的調整。語音合成中的音長調整，只要範圍不大，可以用切音加上淡出的方式來處理，對語音品質的影響也很小。

整個系統的流程圖如圖一所示。文句分析和韻律分析模組負責提供語言訊息和聲學訊息，聲學訊息包含音長與音量，語言訊息主要爲韻律階層，這是決定停頓長度和選取合成單元的根據。接下來經由單元選取模組來選音拼接，最後的韻律調整模組則是利用韻律分析預估出的聲學參數來進行最後的微調，例如做淡入（fading-in）和淡出（fading-out）等。

在文句分析的工作方面，我們使用了 rule-based 的斷詞方法[10]，用 bigram 機率模型來標記詞類，還有做字轉音…等。韻律階層的求取我們實驗了使用 CART 以及使用剖析器兩種方法。在本節的其它篇幅，我們會介紹中文剖析器、韻律階層和我們所使用的大量詞彙。

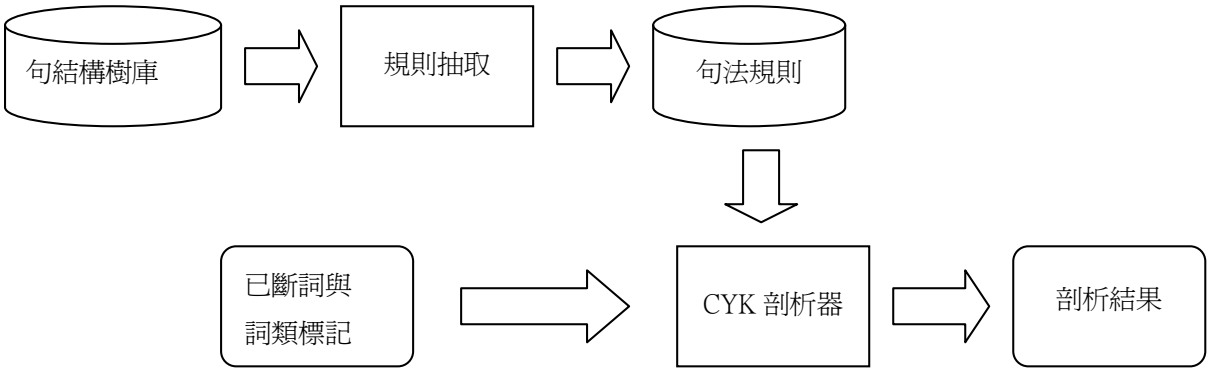


圖一 － 系統流程圖

1.2. 中文剖析器

句法剖析[1]在自然語言處理的過程中有許多用途，如：問答系統、機器翻譯、關鍵字擷取等等。在本文中我們完成一個剖析器，其用途是用來求取韻律階層，利用此韻律階層讓TTS(Text-to-Speech)系統的語音可以聽起來更加好聽。

在過去研究中，利用樹庫(treebank)訓練出來的 probabilistic context-free grammar(以下簡稱PCFG)，拿來對句子做剖析是很常用的技術。在英文部分，因爲有大量的英文樹庫資料，所以利用此英文樹庫所訓練出來的 PCFG 來剖析英文句子，依目前的資料顯示正確率約可至九成[5]。中文剖析器方便目前由中研院所完成的中文剖析器其正確率約在六成[15]。本文中的剖析器也是利用 PCFG 來剖析句子，所訓練出來的 PCFG 是由中研院的中文句結構樹庫中所擷取出來的。，我們是使用 Bottom-up 的 Cocke-Younger-Kasami（CYK）演算法[1][7]來實做我們的中文剖析器。此剖析器的系統架構如圖二所示。



圖二 - 剖析器系統架構圖

1.3.韻律階層

我們的 TTS 系統使用到韻律階層架構，它主要的用途是用來預測停頓的位置和類型。韻律階層由上到下爲句子(Sentence，簡寫爲 S)、韻律片語(Prosodic phrase，簡稱 PP)、韻律詞(Prosodic word，簡稱 PW)、和詞(word)。韻律階層可以描繪成一個樹狀的結構，稱爲韻律階層樹。一顆典型的韻律階層樹如圖三所示。這個架構下的階層共有 4 個，下面爲各階層的定義：

1． 詞(word)

定義：詞爲有意義的基本單位。

例：台北。

2． 韻律詞（Prosodic Word，簡稱爲 PW）

定義：組成單元爲一個或多個詞。

例：台北去玩。

3． 韻律片語（Prosodic Phrase，簡稱爲 PP）

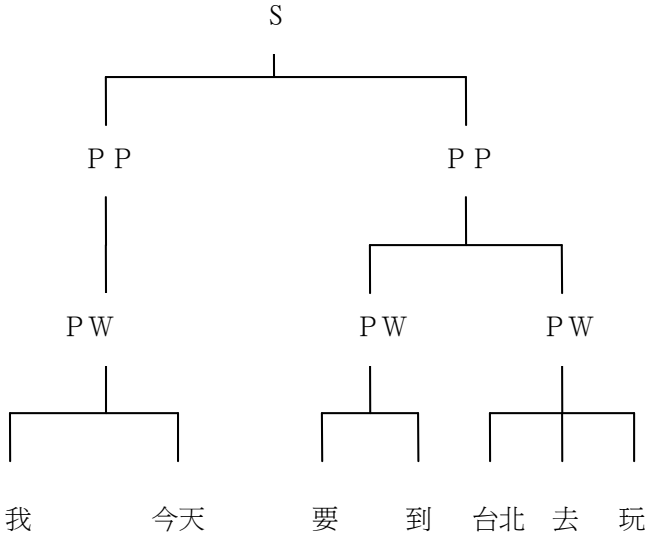
定義：組成單元爲一個或多個韻律詞。

例：要到台北去玩。

4． 句子（Sentence，簡稱爲 S ）

定義：以五大標點符號（包含：“，” “。” “；” “！” “？”）爲區隔，組成單元爲一個或多個韻律片語。

例：我今天要到台北去玩。



圖三 - 韻律階層樹

韻律詞裡面不停頓，爲一個連續發音的單位，韻律詞跟韻律詞之間爲小停頓（minor break），韻律片語跟韻律片語之間爲中停頓（major break)，另外還有一種大停頓是發生在標點符號出現的時候。停頓對於語音的可辨度和自然度有相當程度的影響，良好的停頓可以讓人易於理解句中涵義。

1.4.大量詞彙

我們有鑒於錄製句子，可能無法錄到一些罕見詞，以及考慮到詞跟詞之間的連音現象較微弱[6] ，所以嘗試使用大量詞彙(word-based)當成我們的合成單元。換句話說，如果可以錄製大量詞彙來做爲合成單元，這些合成單元會包含相當豐富的連音訊息。表一所示就是 word-based 對於 corpus-based 主要的優缺點比較。主要的缺處在於句子韻律上的音高問題。人在講話時，詞的發音會因爲在句中位置的不同，而有不同的音高變化。而錄詞的方式在選音的時候，合成單元聲調高低的變化範圍較小。除此之外，以音節平均音長來說，當我們錄製詞或單音的時候，合成單元會比錄製句子時來的長，錄音者必須注意這個問題。

表一 – word-based 相對於 corpus-based 的優缺表

優點	缺點
1． 包含較多連音訊息。 2． 合成單元音程較完整。	1． 較無法錄製到帶有句子韻律的音。 2． 錄音時需注意音長。

現階段在系統的實作上，總共錄製約 12224 個二字詞和 2690 個三字詞，這些詞是出現頻率較高的詞。另外，還有一個單音庫，這個單音庫包含所有中文可能出現的音。在不管聲調的情況下，中文約有 409 個音，而中文有五個聲調，所以這個單音庫整整錄製了 409*5 種音，雖然說實際上中文的發音只會有約 1300 種。

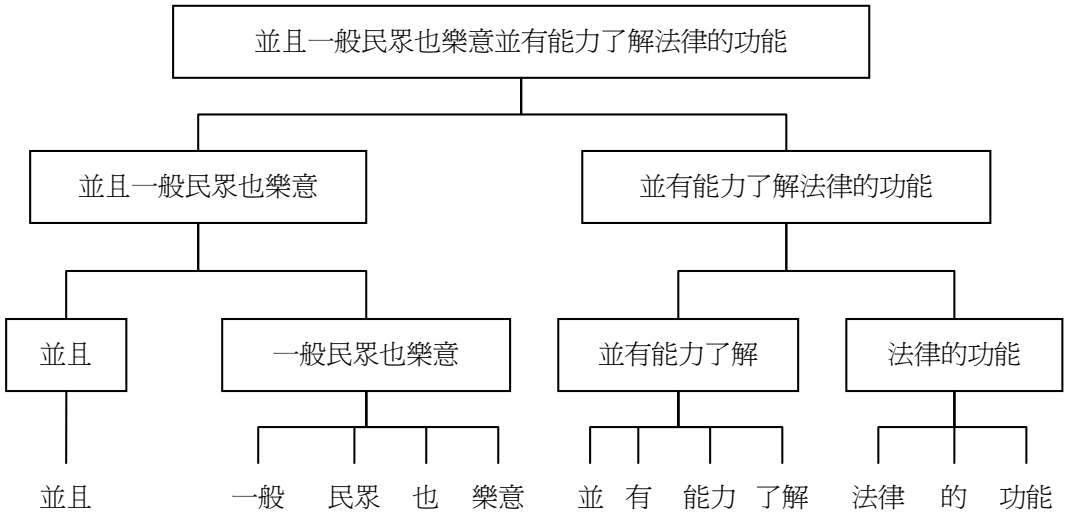
2.韻律階層的求取

在本節中，我們先描述如何得到帶有停頓標記的實驗語料。然後說明求取停頓標記的兩種方式：使用 CART 和使用剖析器的方式。

2.1.標記實驗語料

我們求取韻律階層的實驗語料，是從中研院的中文句結構樹庫 2.1 裡的句結構樹，拿掉文法剖析的結果，還原成原本的句子。再由實驗室的成員，以自己的感覺下去標記的。標記的方法是實際唸過一次，並從自己唸的方式來對句子標記韻律片語和韻律詞的中斷。總共有九份語料，除了第九份只有 1,923 句外，其於八份每份有 6,250 句。由實驗室的成員一人標記一份語料。在檢查標記完的結果時，若某句的標記有將中斷標在原本結構樹的詞內的話，則視爲無效的標記，並刪除該句。經過這項檢查後，全部有效的標記句子有 51,525 句。以下列出一個帶有標記的句子和它所對應的韻律階層樹（圖四），其中＊表示韻律片語的邊界，空白表示韻律詞的邊界。

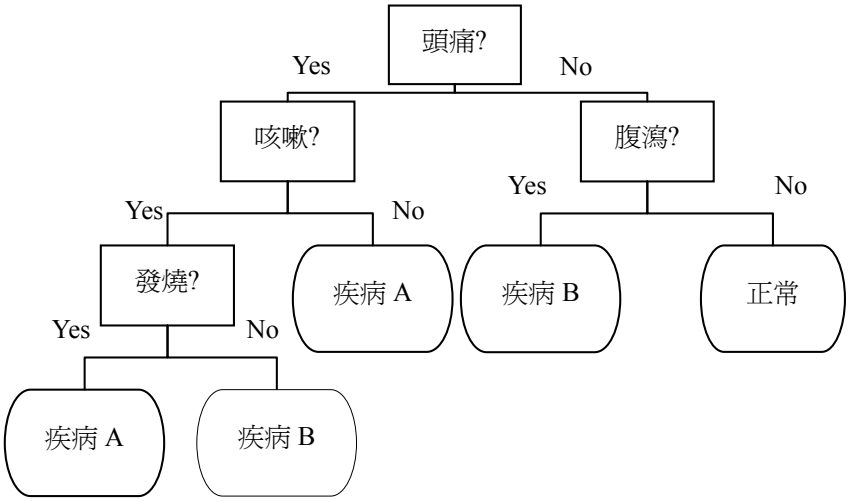
EX:並且　一般民眾也樂意＊並有能力了解　法律的功能



圖四 - 韻律階層樹

2.2. CART(Classification And Regression Trees)

CART 是一種二元(binary)分割的方法。分割條件的選擇是根據資料的分類數及其屬性來決定，並依據 Gini 規則[2]來決定分割的條件。每經過一次分割後，資料會被分成兩群，如圖五所示。經由不斷的切割資料，達到分類的目的。

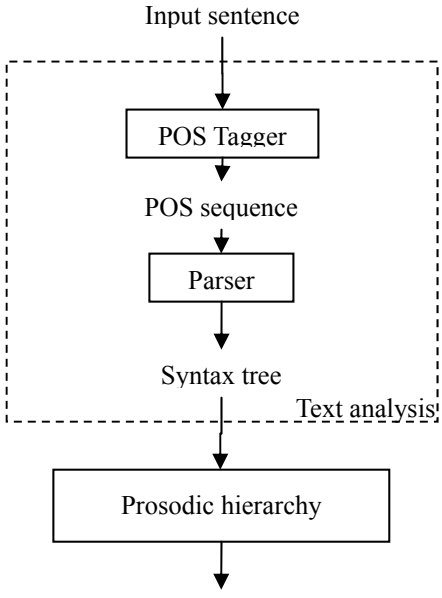


圖五 CART 分類樹

以圖五爲例子，假設我們要分類的是得到 A 疾病和 B 疾病以及正常的人。首先我們先確定是否有頭痛的現象，若是沒有，則走到腹瀉（即檢查腹瀉的節點），若是有，則到咳嗽（即檢查咳嗽的節點）。接下來再繼續問問題，此時腹瀉這個節點的集合已經可以分類出來了，一個是正常，而另一個是有 B 疾病。其它的類推，便是以此種方法來分類。

我們將詞和詞之間的停頓分成下列三類：無停頓(no break)、小停頓(minor break)、以及中停頓(major break)，所以我們可以使用 CART 這樣的方法來分類。而 CART 的特點是，它是自動來產生這樣的決策樹，也就是我們只需準備語料和所用到的特性，便可以使用 CART 來對資料作分類的工作。

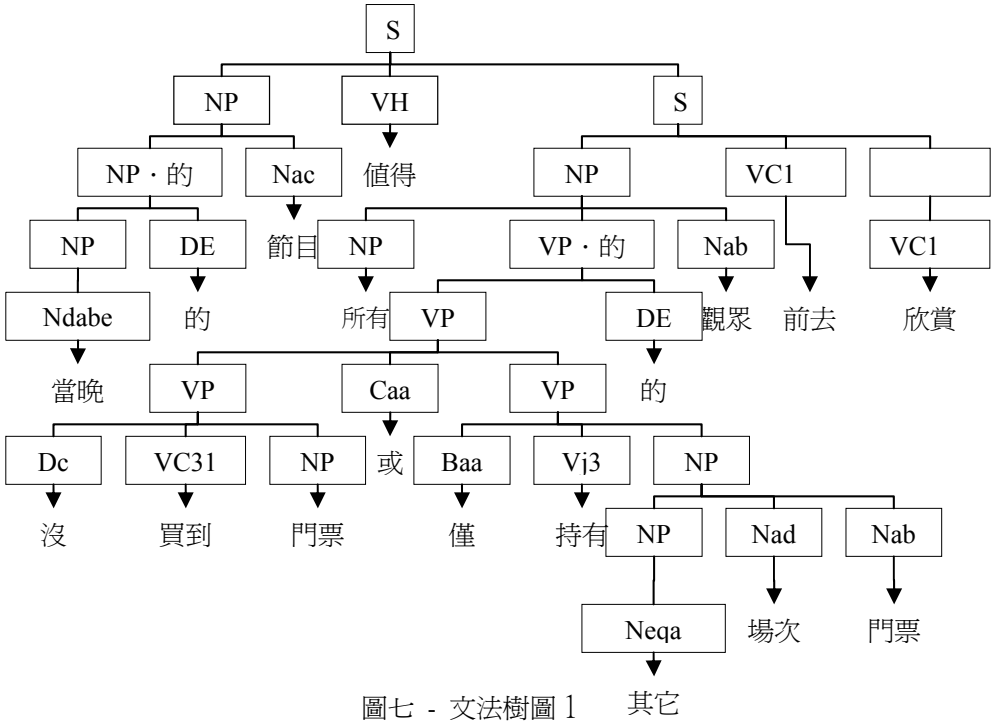
2.3.剖析器爲基礎的方法



圖六 - 模組關係圖

我們的另一個求取停頓型態的實驗，是以文法剖析爲基底的方式[13]。先求得整句的文法樹之後，再從該文法樹來得到整段句子的韻律階層(prosodic hierarchy)。這個模組在整個系統中是包含在文句分析(Text analysis)這個大模組下。大略的架構如圖六，句子進來先經過斷詞器斷詞並標記詞性，將輸出的結果輸入給文法剖析器，接著得到該句子的文法樹，再從文法樹中得到韻律階層。

接下來我們用實例來說明這種方法。假設輸入的句子爲：當晚的節目值得所有沒買到門票或僅持有其它場次門票的觀眾前去欣賞，此句的文法樹結構如圖七所示。我們會依照由底層往上層合併的方式來求取韻律詞和韻律片語，在合併時我們會給韻律詞（片語）字數的限制。我們依序說明字數的限制和合併的方式。



圖七 - 文法樹圖 1

2.3.1.韻律詞及韻律片語的字數

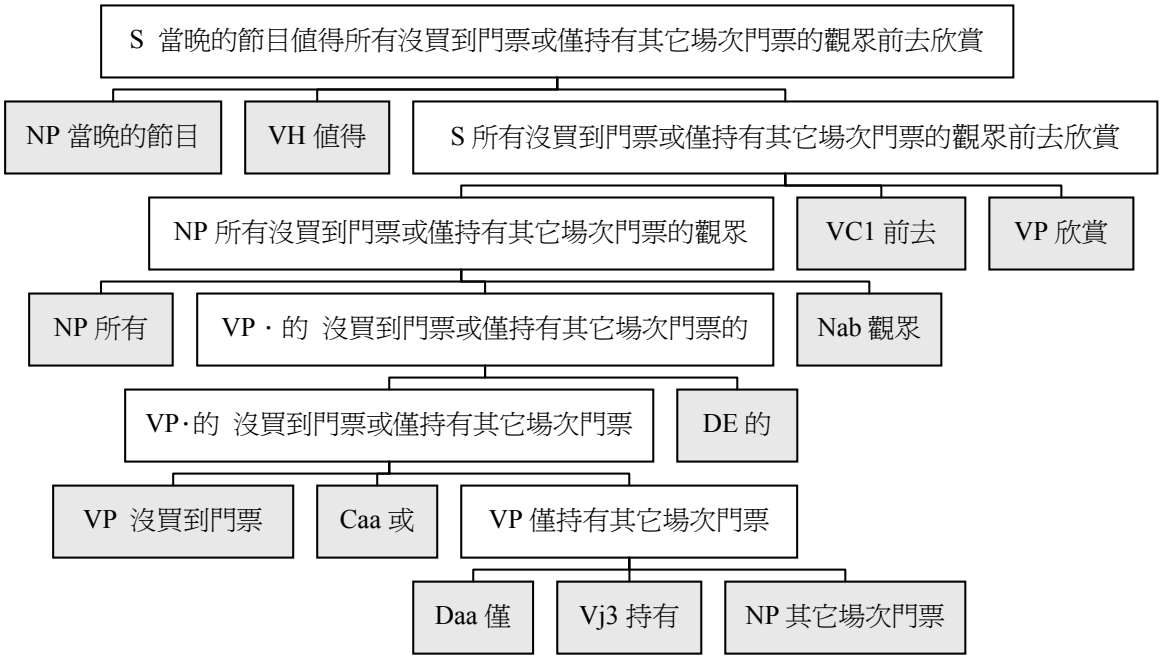
圖七表現出這個句子的文法結構。接著我們便從這樣的結構中找出可能的韻律詞(prosodic word)和韻律片語(prosodic phrase)。我們認爲韻律詞和韻律片語所含的字數會隨著句子的長度(句中字數)而改變。也就是說，在較長的句子中，韻律詞(片語)的字數也會比較多。經過觀察和實驗，我們給予韻律詞(片語)一個最大字數的限制，這個限制會隨著句長而變動。表二顯示我們使用的句長與最大韻律詞(片語)的字數關係。

表二　句長與最大韻律詞(片語)字數關係表

句長 / 最大字數	1-3	4-8	9-12	13-18	19-22	23-30	31-39	40-41	>42
PW 最大字數	句長	6	6	6	6	6	7	7	8
PP 最大字數	句長	句長	9	12	13	14	15	16	16

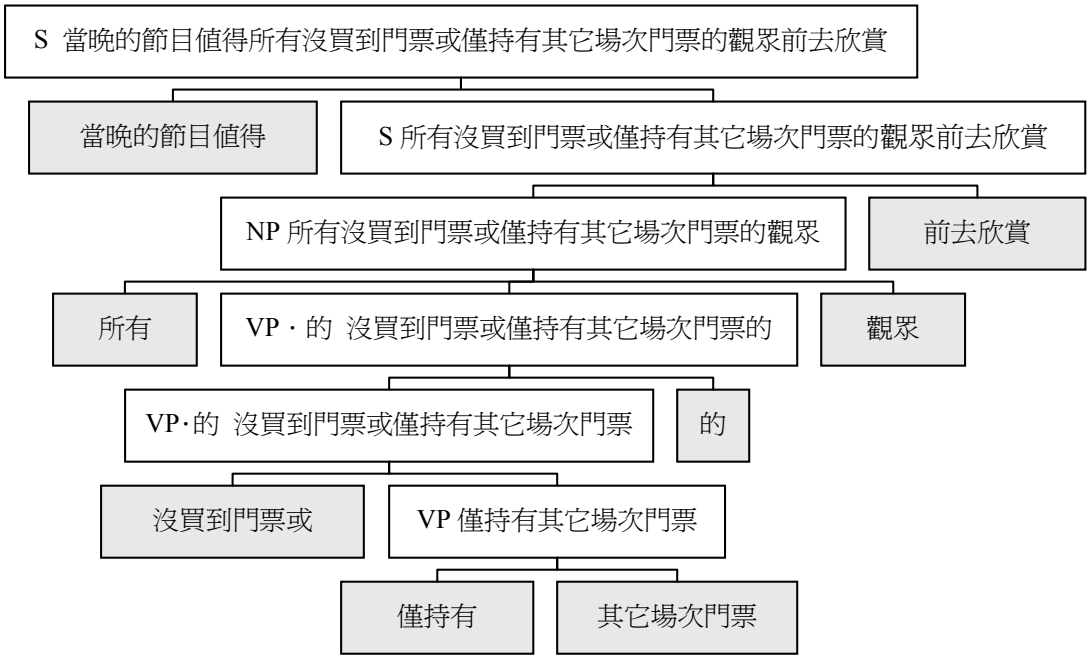
2.3.2. 由底層往上層合併

定好韻律詞和韻律片語的最大字數後，接下來是用由最底層往上層(bottom-up)合併的方式來找出韻律詞和韻律片語。我們以三個步驟來實作這個方法，首先是將小樹縮成韻律詞，接著再同層節點合併成韻律詞，最後由底向上合併。這個方法直接依照長度合併韻律詞和韻律片語。第一個步驟會掃描文法樹，將文法樹中較小的樹縮減成一個節點。經過第一個步驟，我們將小樹的部分縮成韻律詞，縮減的結果如圖八所示。



圖八 - 小樹縮減結果

接下來是第二步驟，同層葉節點的合併。由圖八中可以很清楚的看出同層的單元。在此步驟中，合併的限制是不能超過韻律詞的最大字數。在此步驟我們會將 NP(當晚的節目)和 VH(值得)合併成一個單位，VC1(前去)和 VP(欣賞)合併成一個單位。以此類推，同層合併的結果，如圖九所示。最底層有 Daa(僅)、Vj3(持有)、NP(其它場次門票)，我們優先合併字數較短的節點，所以會合併 Daa(僅)和 Vj3(持有)。

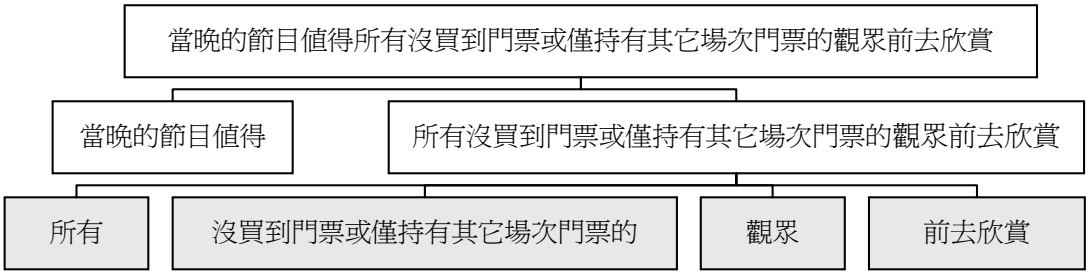


圖九 - 同層合併結果

最後我們由底向上合併韻律詞(片語)，從最底層開始(第六層)，首先處理的就是“僅持有”和“其它場次門票”，兩個節點的字數和爲 9，雖然超過最大韻律詞的字數，但未超過最大韻律片語的字數，所以合併成一新的韻律片語。最底層的部分合併完畢。

接著往上昇一層處理，要處理的單元爲：“沒買到門票或”、“僅持有其它場次門票”。由於兩個單元的字數合爲 15，剛好到最大的韻律片語字數，所以可以再合併成韻律片語。接著往上昇一層，處理“沒買到門票或僅持有其它場次門票”、“的”。雖然兩個單元的字數和會 16，但是由於“的”是詞綴[16]，而在我們的方法中，詞綴合併不受字數限制，因此合併成“沒買到門票或僅持有其它場次門票的”。

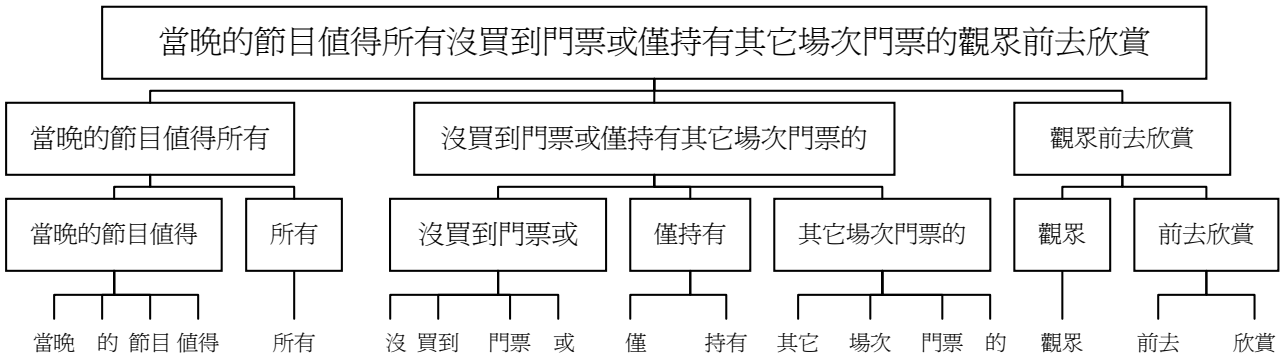
再往上昇一層，處理“所有”、“沒買到門票或僅持有其它場次門票的”、“觀眾”。在這次的處理中，由於字數的限制，因此不會產生任何合併的動作。所以將這三個單元往上昇一層處理，如圖十所示。



圖十 - 上昇處理圖示

接下來處理的單元變成了：“所有”、“沒買到門票或僅持有其它場次門票的”、“觀眾”、“前去欣賞”。而在這個部分，我們可以合併的單元爲“觀眾”、“前去欣賞”，合併的形態爲韻律詞。合併後再上昇，得到四個單元：“當晚的節目值得”、“所有”、“沒買到門票或僅持

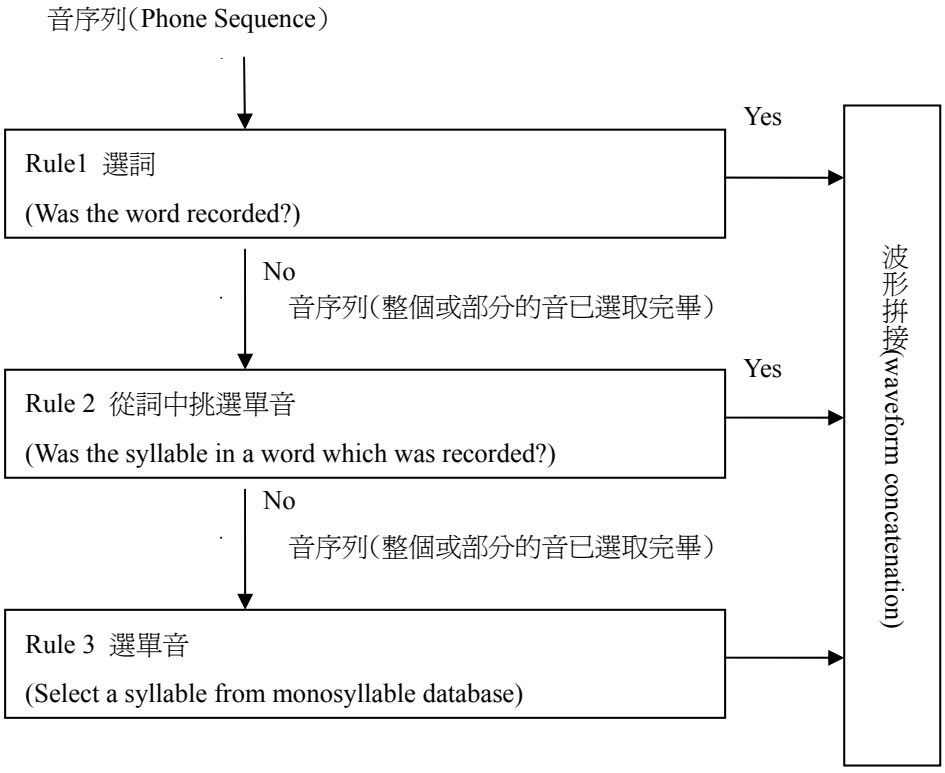
有其它場次門票的”、“觀眾前去欣賞”。而這四個單元可以合併的單元爲“當晚的節目值得”、“所有”，合併的形態爲韻律片語。最後可以得到如圖十一的韻律階層。



圖十一 - 韻律階層圖(Bottom-up)

3.合成單元選取

在選擇適當的合成單元方面，主要是使用 rule-based 的方法來進行挑音，規則爲「有同音詞（二字以上）找同音詞，沒有則從詞（二字以上）中找尋單音，沒有再從單音庫裡面尋找單音拼接」，選音的流程如圖十二所示。



圖十二 - 選音流程圖

在流程圖中，通常 rule1 和 rule3 都只有一個候選，在 rule2 上，可能有許多的候選音，我們使用了一個代價公式來幫助我們選音。這個代價公式的設計，是依據（1）詞中位置（Position in word）和（2）前音與後音的聲調（Tone）和（3）子音/母音結構（C/V 結構）來做音節單元的

挑選。這個代價公式所呈現的實際上的意義爲「先選詞中位置相同的，再考慮前後音的子音、母音和聲調的代價」。先介紹詞中位置的定義，詞中位置包含：詞首、詞中、詞尾，例如：後悔莫及（“後”爲詞首、“悔”和“莫”爲詞中、“及”爲詞尾）。我們以一個範例解釋從詞中選單音的概念。表三爲選音時的ㄉㄤ的候選。

表三 – 選音範例

候選音	候選所在詞	前音是否在詞內	前音	後音是否在詞內	後音
ㄉㄤ	**當**天	否	ㄑㄩㄢˊ	是	ㄊㄧㄢ
ㄉㄤ	便**當**	是	ㄅㄧㄢˋ	否	ㄧㄡˇ
ㄉㄤ	便**當**店	是	ㄅㄧㄢˋ	是	ㄉㄧㄢˋ
ㄉㄤ	**當**選	否	ㄕˋ	是	ㄒㄩㄢˇ

假設我們合成的句子的斷詞結果爲「當晚　我們　要　到　台北」，而(ㄉㄤ　ㄨㄢˇ)這個同音詞並沒有錄過，接下來便要從詞中去選擇ㄉㄤ。因爲目標音爲詞首所以只要考慮詞中位置和後音的代價。“當天”和“當選”的詞中位置和目標位置都是在詞首，所以優先選之。在這兩個ㄉㄤ的候選中，“當選”的ㄉㄤ因爲“選”的聲調和“晚”的聲調是相同的，因此計算出來的代價爲最小。所以最後會選擇“當選”的ㄉㄤ來做拼接。（詳細請參考[11]）

4.韻律調整

通常中文 TTS 要調整的韻律包含音調、停頓、音長以及音量。我們以合成單元選取的方式來得到適當音調的合成單元，而不做音調的調整。在音長和音量的調整上，我們用[13]的統計方式韻律預估模型來預估各音節的音長和音量，然後盡量調整成所要求的音長和音量。

在韻律調整模組上，主要處理工作包含停頓值的給予、適當的重疊（overlapping）、音長調整和音量調整。停頓的部份，我們採取給予固定停頓值的作法，我們根據聽測結果來決定適當的停頓值，如表四所示。

表四 – 停頓時間表

停頓類型	停頓長度(ms)
Minor break	50
Major break	250
逗號（，）	400
句號（。）	625
問號（？）	625
驚嘆號（！）	625
分號（；）	500
頓號（、）	250
冒號（：）	300

在我們的韻律階層架構中，Prosodic Word 內是不會發生停頓的。但是在實際利用合成單元來拼接的時候，某些合成單元間直接拼接會造成連接處停頓的感覺，這跟韻律階層的概念有所出入，所以我們在合成單元間給予短暫的重疊（Overlapping）。重疊的長度爲我們參考[8]所制定出來的，如表五所示（ㄐ代表爲無聲母）。

表五 - 重疊長度比例表

後音節子音的類型	重疊的部分佔目前音長的比值
ㄅㄆㄉㄊㄍㄎ	0
ㄐㄓㄕ	0.05
ㄑㄔㄘ	0.1
ㄈㄏㄒㄕㄙ	0.15
ㄇㄋㄌㄖㄐ	0.2

在音長的調整上，我們並不使用 PSOLA 來進行調整，而是使用 cut off（切音，切掉單音後段）的作法，這個方法簡單而且能確保音的清晰。主要的原則爲「如果目標音比合成單元短則進行切音，否則不進行調整」。在切音之後，我們會做一個淡出（fading-out）的處理，使其聽起來較爲自然。

在音量上，只要不要調整到爆掉，皆不會發生如調整音長時的失真情形。淡入（fading-in）和淡出（fading-out）主要用來降低音量上的突兀，假設我們要調整的部分總共有 n 個點，每點振幅值爲$< x_1, x_2, x_3, ..., x_n >$，淡出的公式如(式 1)所示。同樣的假設下，淡入的公式如(式 2)所示。我們認爲有三個地方是需要 fading-in 和 fading-out，分別是（1）合成單元與後音有較強的連音現象（fading-out），（2）合成單元與前音有較強的連音現象（fading-in），以及（3）用切音縮短過音長（fading-out）。

$$x_{i(fading\_out)} = x_i * \frac{n-i+1}{n+1} \quad , \quad i = 1 \sim n \quad , \quad \cdots\cdots\cdots\text{(式 1)}$$

$$x_{i(fading\_in)} = x_i * \frac{i}{n+1} \quad , \quad i = 1 \sim n \quad 。 \quad \cdots\cdots\cdots\text{(式 2)}$$

5.實驗結果

5.1. 剖析器實驗數據

測試與訓練語料我們用中研院句結構樹庫 Version 2.1 來作測試與句法規則抽取，在中研院句結構樹庫中總共有 54,902 個中文句結構樹。但是在資料庫中有一些標示“%”符號的句子，它代表意義是句子不完整導致剖析器無法剖析，或是語意錯誤不合文法，我們必須事先將它剔除。所以剩下 51939 句來作爲訓練語料，從中抽取句法規則和統計機率；測試語料也是利用訓練語料來做內部測試（inside test），另外對於詞類標記我們也有做縮減，原因在於可以提升覆蓋率[15]，讓無法剖析出的句子減少。所以我們利用縮減詞類的語料來製作剖析器，同時也拿來作測試，詞

類簡化對應如附錄一。

針對剖析器的評估，我們有對樹結構(tree structure)、標記(Label)、括號(Bracket)作正確率的評估，其中對於評估標記與括號好壞的評估模型是 PARSEVAL[4]。我們採用如下的評估項目與公式[15]：

- 結構樹正確率 SP(Structure Precision)

$$SP=\frac{\#\text{correct parsing tree of testing data}}{\#\text{treebank parsing tree of testing data}}$$

- 詞組標記正確率 LP（Labeled Precision）

$$LP=\frac{\#\text{label correct constituents in parser's parse of testing data}}{\#\text{label constituents in parser's parse of testing data}}$$

- 詞組標記召回率 LR（Labeled Recall）

$$LR=\frac{\#\text{label correct constituents in parser's parse of testing data}}{\#\text{label constituents in treebank's parse of testing data}}$$

- 詞組標記效能評估 LF（Labeled F-measure）

$$LF=\frac{LP*LR*2}{LP+LR}$$

- 括號精確率 BP（Bracketed Precision）

$$BP=\frac{\#\text{bracket correct constituents in parser's parse of testing data}}{\#\text{bracket constituents in parser's parse of testing data}}$$

- 括號召回率 BR（Bracketed Recall）

$$BR=\frac{\#\text{bracket correct constituents in parser's parse of testing data}}{\#\text{bracket constituents in treebank's parse of testing data}}$$

- 括號效能評估 BF（Bracketed F-measure）

$$BF=\frac{BP*BR*2}{BP+BR}$$

實驗數據如表六所示。

表六 剖析器實驗結果(單位：%)

SP	LP	LR	LF	BP	BR	BF
38.78	61.96	64.31	63.11	70.04	72.80	71.39

雖然樹結構正確率不高，不過實驗數據中我們最主要的括號正確率與召回率分別爲 70.04%與 72.80%。這兩個數字很重要，因爲在我們求取韻律階層時用到的是語法樹(parse tree)的括號結構，詞性並未使用到。

5.2. 韻律階層求取實驗數據

接下來的部分是韻律階層的求取。我們會將分類的結果和原始人工標記的結果作比對，產生一混淆矩陣，如表七所示。接著用此混淆矩陣來計算三種停頓型態預估的正確率和召回率。

表七 – confusion matrix

True labels	Predicted labels		
	B_0	B_1	B_2
B_0	C_{00}	C_{01}	C_{02}
B_1	C_{10}	C_{11}	C_{12}
B_2	C_{20}	C_{21}	C_{22}

在表七中，$B_i\,(i = 0,1,2)$ 表示詞邊界(詞間)的停頓型態，B_0 表示該邊界在韻律詞內，不加停頓(no break)；而 B_1 表示該邊界爲韻律詞間，應加入小停頓(minor break)；B_2 表示該邊界爲韻律片語間，加入中停頓(major break)。其中對角線 C_{ii} (i =1 ,2 ,3)表示所預測的結果和標準答案一致的次數，而 C_{ij} (i ,j= 1 ,2 ,3 ; $i \neq j$)表示真實答案的標記爲 B_i 而程式標記成 B_j 的次數。

某一類別的召回率(Recall)的計算方式爲：

$$\mathrm{Rec_i} = C_{ii} \Big/ \sum_{j=0}^{2} C_{ij} \quad (i = 0, 1, 2)$$

以表七的混淆矩陣而言，我們可得到 B_0 (no break)的召回率

$$\mathrm{Rec_0} = C_{00} \Big/ (C_{00} + C_{01} + C_{02})$$

某一類別的精確率(Precision)的計算方式爲：

$$\mathrm{Pre_i} = C_{ii} \Big/ \sum_{j=0}^{2} C_{ji} \quad (i = 0, 1, 2)$$

以表七的混淆矩陣而言，我們所得到的 B_0 (no break)的正確率

$$\mathrm{Pre_0} = C_{00} \Big/ (C_{00} + C_{10} + C_{20}) \text{ 。}$$

全部類別的正確率(Accuracy)的計算方式爲：

$$\mathrm{Acc} = \sum_{i=0}^{2} C_{ii} \Big/ \sum_{i=0}^{2} \sum_{j=0}^{2} C_{ij}$$

以表七的混淆矩陣而言，我們所得到的正確率 Acc 爲

$$(C_{00} + C_{11} + C_{22}) / (C_{00} + C_{01} + C_{02} + C_{10} + C_{11} + C_{12} + C_{20} + C_{21} + C_{22})$$

接著是實驗的數據，表八是將所有的人工標記的語料合成一份，接下來將語料分成五等分，

四等分當成訓練語料來訓練 CART 的分類樹，而一等分爲測試語料，用來計算正確率。表八爲 CART 得到的結果。表九是以剖析樹的方式，對語料標記韻律階層。由於這個方法是規則式的，所以不用分訓練語料和測試語料。表九是九份語料合在一起標記所得到的結果。而表八和表九中的 Acc1 表示總正確率，Acc2 表示將 B1 和 B2 視爲同一類標記所得到的總正確率。

表八 - CART 結果

True labels	Predicted labels		
	B0	B1	B2
B0	30,434	3,198	126
B1	5,758	6,810	372
B2	635	1,381	514
Acc1：0.767	Pre0：0.826	Pre1：0.598	Pre2：0.508
Acc2：0.791	Rec0：0.902	Rec1：0.526	Rec2：0.203

表九 - Bottom-Up 結果

True labels	Predicted labels		
	B0	B1	B2
B0	156090	7384	4502
B1	54390	5471	5062
B2	7854	1828	2911
Acc1：0.669	Pre0：0.715	Pre1：0.372	Pre2：0.233
Acc2：0.698	Rec0：0.929	Rec1：0.084	Rec2：0.231

5.3. 文轉音系統實驗數據

在我們的實驗中，我們會分別進行自然度(naturalness)測試、偏好測試（preference testing）和可辨度(intelligibility)測試。我們利用 MOS（Mean Opinion Score）來評量我們合成語音的自然度。實驗中，測試方法爲播放語音，請一些人來當測試者，爲這些語音來評分數。分數分成五個等級：5 分：非常好（excellent）、4 分：好（good）、3 分：普通（fair）、2 分：差（poor）、1 分：極差（unsatisfactory）。偏好測試測試方法爲連續播放兩個合成語音，請聽者選取較佳的一個。偏好測試與自然度測試使用相同的文句，可辨度測試以句子爲單位，測試方法爲請聽者寫下合成語音的音或文字。

第一次的測試者爲 8 名本實驗室的成員，計有研究生 6 人和教師 2 人。在 MOS 測試和偏好測試中，我們是以段落（paragraphs）爲單位，測試總共有 20 段，每段長度介於 15 至 25 個字之間。而測試段落的來源爲新聞語料，每個段落選自不同主題，有政治、體育、影劇…等，偏好測試語自然度測試使用相同的段落。實驗目的方面，我們較感興趣於 Prosodic Word 間的小停頓是否要停頓？所以進行了自然度以及偏好測試，實驗結果如表十和表十一。兩者的結果相反，我們認爲聽者不易區別 Prosodic Word 間的小停頓。在可辨度測試中，得到 97.2%的正確率。

表十　自然度測試數據

測試者編號	有給停頓值(平均 MOS)	標準差	不給停頓值(平均 MOS)	標準差
M01	4.05	0.497	4	0.474
M02	3.15	0.963	3.15	0.792
M03	3.3	0.714	3.3	0.640
M04	4.385	0.504	4.67	0.181
M05	3.55	0.668	3.65	0.852
M06	3.9	0.538	4	0.632
M07	4.2	0.748	4	0.707
M08	2.95	0.804	2.95	0.804
平均	3.68		3.715	

表十一　偏好測試數據

測試者編號	有給停頓值（%）	不給停頓值（%）
M01	40	60
M02	50	50
M03	55	45
M04	50	50
M05	70	30
M06	70	30
M07	50	50
M08	55	45
平均	55	45

另一個測試是請本校的八位研究生作 MOS 測試，測試的句子共有四十句，測試的語音以 CART 和文法樹方式兩種方法所得到的韻律階層實際合成的語音各二十句。除了句子外並合成六篇文章，用 CART 和文法樹的韻律階層各合成三篇。測試前先定義 MOS 給分的準測[3]。我們定義：

5 分：難以分辨是合成語音還是自然語音。

4.5 分：清楚可辨度佳，在半小時內聽不累。

4 分：可以很清楚的了解在語音的意思，且沒有特別的斷詞錯誤。會有一或二個音節發音不清。

3 分：大部分能聽懂合成語音的意思，有明顯的錯誤。聽者無法連續聽十分鐘。

2 分：聽者無法聽出一些關鍵字，且此種的合成語音聽起來像是直接由音節連在一起。

1 分：聽起來像是機器人講話的聲音，並且聽不出語音所表達的意思。

測試結果如表十二所示。表十二表示在 MOS 測試中，八位聽者所給的平均分數。上表的數字表示在 MOS 的測試中，語者給分的平均值。用文法樹所產生的韻律階層分數比用 CART 所產生的韻律階層稍高，但相差不多。雖然文法樹在正確率的數據輸給 CART，但是實際合成語音的表現比 CART 稍好。

表十二　MOS 測試結果

		聽者 1	聽者 2	聽者 3	聽者 4	聽者 5	聽者 6	聽者 7	聽者 8	平均
句子	CART	3.45	3.65	3.55	3.35	3.5	3.9	4.2	4.1	3.71
	文法樹	3.48	3.9	3.45	3.95	3.45	4.18	3.88	4.25	3.82
文章	CART	4	4	4.66	4.66	4.66	4.66	3.66	5	4.42
	文法樹	4	4	4.33	5	5	4.83	4	5	4.52

6.結論與未來研究

在文轉音系統方面，目前已經得到一套發音清晰（可辨度測試 97.2%）的中文文轉音系統（線上系統網址：http://140.120.15.239/onlineTTS/cgitest.html）。未來還有需要處理的工作有（一）構詞部分的加強。例如：等看看。（二）連音變調需要的語意分析。例如：老李買好酒。（三）破音字的判別。例如：得 （ㄉㄜˊ）or（ㄉㄜ•）。（四）停頓型態的預測的正確率。（五）自動切音的工作。（六）錄製大量語料與罕見詞。在韻律階層預測上，實驗上韻律片語比韻律詞更爲重要，所以未來的工作朝向更準確更合理的韻律片語預測。

參考文獻

[1]Aho, A. V. and Ullman, J. D., "The Theory of Parsing, Translation, and Compiling ",1972, Vol. 1, Prentice-Hall, Englewood Cliffs, NJ.

[2]Breiman L, Friedman J. H., Olshen R. A., et al, "Classification and Regression Trees", Wadsworth, Inc, 1984.

[3]Bao H., Wang A., Lu S., "A Study of Evaluation Method for Synthetic Mandarin Speech", Proceedings of ISCSLP 2002, PP:383-386, Taipei, Taiwan.

[4]Charniak, E., "Treebank Grammars", In Proceedings of the Thirteenth National Conference on Artificial Intelligence, pp. 1031-1036. AAAI Press/MIT Press, 1996.

[5]Collins, M. J., "Head-Driven Statistical Models for Natural Language Parsing. ", Ph.D. Thesis, University of Pennsylvania, Philadelphia, 1999.

[6] Chu M., Peng H., Yang H. Y. and Chang E., " Selecting Non-Uniform Units from A Very Large Corpus for Concatenative Speech Synthesizer ", Proceedings of ICASSP 2001, IEEE, Volume 2, pp.785 - 788, Salt Lake City.

[7]Ney, H. "Dynamic Programming Parsing for Context-Free Recognition", IEEE Transactions on Signal Processing 1991, 39(2), 336-340.

[8] Hwang S. H. and Yei C. Y., "The Synthesis Unit Generation Algorithm for Mandarin TTS", Proceedings of ICASSP 2002, IEEE, Volume 1, pp. 457 - 460, Orlando, Florida.

[9]周福強, "以語料庫爲基礎之新一代中文文句翻語音合成技術", 國立臺灣大學電機工程學研究所博士論文, 1998 年。

[10]唐大任, "中文斷詞器之研究", 國立交通大學電信工程學所碩士論文, 2001 年。

[11]張唐瑜, "以大量詞彙作爲合成單元的中文文轉音系統", 國立中興大學資訊科學所碩士論文, 2005 年。

[12]許燦煌, "機率式中文剖析器之設計與實作", 國立中興大學資訊科學所碩士論文, 2005 年。

[13]潘能煌, "中文文轉音系統的韻律預估及其改進", 國立中興大學應用數學所博士論文, 2004 年。

[14]蔡育和, "中文文轉音系統中韻律階層的求取", 國立中興大學資訊科學所碩士論文, 2005 年。

[15]謝佑明, 楊敦淇, 陳克健, "語法規律的抽取及普遍化與精確化的研究", Proceedings of ROCLING XVI, 2004, pp.141-150。

[16]中央標準局委辦「中文資料分類處理分詞規範」計畫公聽會, 1998。

附錄一

附錄 詞類縮減對應表

中研院詞類標記	本剖析器所用詞類	說明
A	A	非謂形容詞
Caa	C	對等連接詞
Cab	C	連接詞
Cba	C	連接詞
Cbb	C	關聯連接詞
D	D	副詞
Da	D	數量副詞

DE	DE	的, 之, 得, 地
Dfa	D	動詞前程度副詞
Dfb	D	動詞後程度副詞
Di	D	時態標記
Dk	D	句副詞
FW	N	外文標記
I	I	感嘆詞
Na	N	普通名詞
Nb	N	專有名稱
Nc	N	地方詞
Ncd	Ncd	位置詞
Nd	N	時間詞
Nep	Ne	指代定詞
Neqa	Ne	數量定詞
Neqb	Ne	後置數量定詞
Nes	Ne	特指定詞
Neu	Ne	數詞定詞
Nf	N	量詞
Ng	Ng	後置詞
Nh	N	代名詞
P	P	介詞
SHI	V	是
T	T	語助詞
VA	V	動作不及物動詞
VAC	V	動作使動動詞
VB	V	動作類及物動詞
VC	V	動作及物動詞
VCL	V	動作接地方賓語動詞
VD	V	雙賓動詞
VE	V	動作句賓動詞
VF	V	動作謂賓動詞
VG	V	分類動詞
VH	V	狀態不及物動詞
VHC	V	狀態使動動詞
VI	V	狀態類及物動詞
VJ	V	狀態及物動詞
VK	V	狀態句賓動詞

VL	V	狀態謂賓動詞
V_2	V	有

Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification

Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, Yuan-Hao Chang
Department of Computer Science and Engineering, Tatung University, Taipei
tlpao@ttu.edu.tw, d8906005@mail.ttu.edu.tw, g9206026@ms2.ttu.edu.tw

Abstract. In this paper, we proposed a weighted discrete K-nearest neighbor (weighted D-KNN) classification algorithm for detecting and evaluating emotion from Mandarin speech. In the experiments of the emotion recognition, Mandarin emotional speech database used contains five basic emotions, including anger, happiness, sadness, boredom and neutral, and the extracted acoustic features are Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). The results reveal that the highest recognition rate is 79.55% obtained with weighted D-KNN optimized based on Fibonacci series. Besides, we design an emotion radar chart which can present the intensity of each emotion in our emotion evaluation system. Based on our emotion evaluation system, we implement a computer-assisted speech training system for training the hearing-impaired people to speak more naturally.

1 Introduction

Recognizing emotions from speech has gained increased attention recently, because automatic emotion recognition can help people to develop and design many applications about human-machine communication. In emotion recognition, collecting corpus and selecting the suitable features and classification algorithms are the two most difficult problems.

Language is the most basic and main tool for the human to communicate thoughts, convey messages and express aspiration. For the hearing-normal people, the process of speak learning is very natural. But for the hearing-impaired people, it becomes almost impossible due to no auditory input. Fortunately, the hearing-impaired people are not profoundly deaf and remain some level of hearing. Using these residual hearing and other perception, the hearing-impaired people can still communicate with other people. In real life, we can often see that hearing-impaired people converse with others by sign language, lip reading or writing. In fact, sign language has low popularity among general people. Lip-reading is just a reference because it has some limitations in Mandarin vowels. Writing is not a convenient way. So in many language training, to teach the hearing-impaired people to speak is the ultimate goal.

For this reason, we want to design a computer-assisted emotional speech training system. By using this system, it can assist the hearing-impaired people to learn not only to speak correctly but also to speak naturally, just like the hearing-normal people. Besides, we also use the visual feedback in our system. Just like in many singing training system, we can see the singing score on the screen after singer has sung. Speech therapist can have no trouble to teach hearing-impaired people to speak with emotions when they communicate with people. This mechanism can let the hearing-impaired people better understand their

speaking state and make the whole system more complete.

The emotional state of a speaker can be identified from the facial expression [1] [2] [3], speech [4] [5] [6], body language, perhaps brainwaves, and other biological features of the speaker. A combination of these features may be the way to achieve high accuracy of recognition. But they all are not unconditionally prerequisites necessary for extraction of an emotion.

In this paper, a system is proposed to classify and evaluate the emotions, including anger, happiness, sadness, boredom and neutral, from Mandarin speech. Several early research works in this area are reviewed as follows.

ASSESS [4] is a system that makes use of a few landmarks – peaks and troughs in the profiles of fundamental frequency, intensity and boundaries of pauses and fricative bursts in identifying four archetypal emotions. Using discriminant analysis to separate samples that belong to different categories, a classification rate of 55% was achieved.

In [5], over 1000 utterances emotional speeches, incorporating happiness, sadness, anger and fear from different speakers were classified by human subjects and by computer. Human subjects were asked to recognize the emotion from utterances of one speaker in random order. It was found that human's classification error rate was 18%. For automatic classification by computer, pitch information was extracted from the utterances. Several pattern recognition techniques were used and a miss-classification rate of 20.5% was achieved.

Nicholson et al. [6] analysed the speech of radio actors involving eight different emotions. The emotions chosen were joy, teasing, fear, sadness, disgust, anger, surprise and neutral. In the study, which was limited to emotion recognition of phonetically balanced words, both prosodic features and phonetic features were investigated. Prosodic features used were speech power and fundamental frequency while phonetic features adopted were Linear Prediction Coefficients (LPC) and the Delta LPC parameters. A neural network was used as the classifier. The best accuracy achieved in classification of the eight emotions was 50%.

Machine recognition of emotions using audiovisual information was conducted by Chan [7]. Six basic emotions, happiness, sadness, anger, dislike, surprise and fear, were classified using audio and video model separately. The recognition rate for audio alone is about 75% and video alone is about 70%. For audio processing, statistics of pitch, energy and the derivatives are extracted. Nearest mean criterion approach was adopted for classification. Joint audiovisual information of facial expression and emotive speech were used. The correct recognition rate is 97%.

For the system proposed in this paper, 20 MFCC components and 16 LPCC components were selected as the features to identify the emotional state of the speaker. Subsequently, a weighted D-KNN modified from K-Nearest Neighbor decision rule is adopted as classifier.

2 System Architecture

Figure 1 shows the block diagram of the proposed emotion recognition and evaluation system. The process of emotion recognition is the same as most previous studies. Differently, the evaluation of

emotional speech is a research that only a few people focus on. Therefore, we will emphasize the evaluation of emotional speech in our research. Of course the recognition of emotional speech is also the core of our research we were interested in.

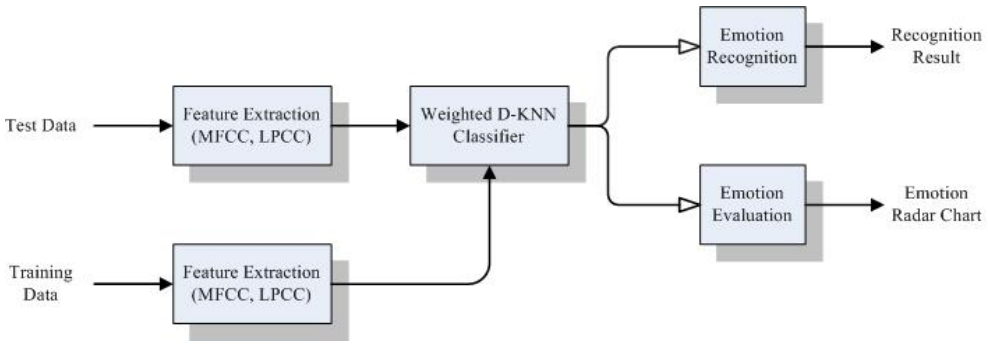


Figure 1: System architecture block diagram

2.1 Emotional Speech Database

We invite 18 males and 16 females to express five emotions, anger, happiness, sadness, boredom and neutral, in their speech. A prompting text with 20 different sentences is designed. The sentences are meaningful so speakers could easily simulate them with emotions. Finally, we obtained 3,400 emotional speech sentences.

After the three-pass listening test procedure, 839 sentences are remained and evaluated by 10 people whom did not have their speech data in the 839 sentences to take part for the final listening test [8]. Table 1 shows the human performance confusion matrix. The rows and the columns represent simulated and evaluated categories, respectively. We can see that the most easily recognizable category is anger and the poorest is happiness. And we can find that human sometimes are confusing in differentiating anger from happiness, and boredom from neutral.

Table 1: Confusion matrix of human performance (%)

	Angry	**Happy**	**Sad**	**Bored**	**Neutral**	**Others**
Angry	89.56	4.29	0.88	0.77	3.52	0.99
Happy	6.67	73.22	3.28	2.36	13.56	0.92
Sad	2.94	1.00	82.76	9.29	3.29	0.71
Bored	1.26	0.44	8.62	75.16	13.65	0.88
Neutral	1.69	0.91	1.56	12.27	83.51	0.06

Table 2: Datasets size

Data set	**D80**	**D90**	**D100**
Size (number of sentences)	570	473	283

For further analysis, we only need the speech data that can be recognized by most people. So we divide speech data into different dataset by their recognition accuracy. We will refer to these data sets as D80, D90, D100, which stand for recognition accuracy of at least 80%, 90%, and 100%, respectively, as listed in Table 2.

In this research, the D80 dataset containing 570 utterances was used. Table 3 shows the distribution of sentences among the five emotion categories for the data set.

Table 3: Distribution of 570 sentences

Emotion Category	Number of Sentence
Angry	151
Happy	96
Sad	124
Bored	83
Neutral	116

2.2 Feature Extraction

A critical problem of all recognition systems is the selection of the feature set to use. In our previous experiment, we investigated the following feature set. Formants (F1, F2 and F3), Linear Predictive Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), first derivative of MFCC (dMFCC), second derivative of MFCC (ddMFCC), Log Frequency Power Coefficients (LFPC), Perceptual Linear Prediction (PLP) and RelAtive SpecTrAl PLP (Rasta-PLP). Due to the highly redundant information, a forward feature selection (FFS) or backward feature selection (BFS) should be carried out to extract only the most representative features.

In FFS, LPCC is the most representative feature while in BFS, it is the MFCC. Finally, we combine MFCC and LPCC as the feature set and is used in our emotion recognition system.

2.3 Classifiers

The simplest classification algorithm is K-Nearest Neighbor (KNN) which is based on the assumption that samples residing closer in the instance space have same class values. Thus, while classifying an unclassified sample, the effects of the k nearest neighbors of the sample were considered. It yields accurate results in most of the cases. However, the classification seems unfair only determine with a point. The k-nearest neighbor classification takes k nearest samples of the testing sample to make a decision.

When a new sample data *x* arrives, KNN finds the k neighbors nearest to the unlabeled data from the training space based on some distance measure. In our case, the Euclidean distance is used. Now let the *k* prototypes nearest to *x* be $N_k(x)$ and *c(z)* be the class label of *z*. Then the subset of nearest neighbors within class $j \in \{ 1 , \ldots , \text{number of classes } l \}$ is

$$N_k^j(x) = \{ y \in N_k(x) : c(y) = j \} \quad (1)$$

Finally, the classification result $j^* \in \{1,...,l\}$ is defined as a majority vote:

$$j^* = \arg\max_{j=1,...,l} \left| N_k^j(x) \right| \tag{2}$$

Modified-KNN is a technique based on the KNN [8]. When a new sample data *x* arrives, M-KNN finds the k neighbors nearest to the unlabeled data in each class from the training space and calculates their distances d^i. Now let the *k* prototypes nearest to *x* be $N_{k,i}(x)$ which is defined as

$$N_{k,i}(x) = \arg\ \min_{j=1,...,l}\ d_j^i\ , \quad i = 1,...,k \tag{3}$$

The following steps are similar to KNN method in making a decision from majority vote.

In this paper, we propose a weighted D-KNN which is a combination of weighting scheme and M-KNN to improve the performance of M-KNN. The purpose of weighting is to find a vector of real-valued weights that would optimize classification accuracy of some classification or recognition system by assigning low weights to less relevant features and higher weights to features that are more important.

2.4 Emotion Evaluation

Emotion evaluation is used to evaluate emotion expression of a sentence. In this paper, the evaluation method we used is based on weighted D-KNN classification. When we take a test data to evaluate, we can obtain five values by the M-KNN classifier. The five values are the distance to the sets of emotion categories respectively. Then each emotional evaluation value can be obtained from each distance set.

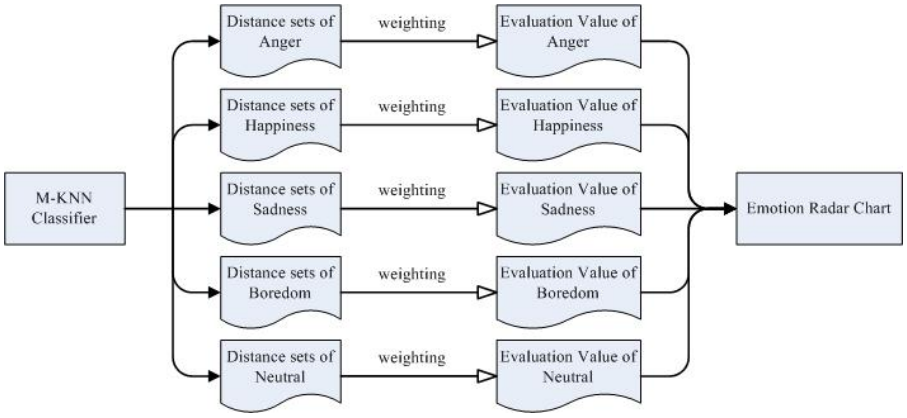


Figure 3: Block diagram of emotion evaluation

Figure 3 shows a block diagram of emotion evaluation. After the calculation of weighted D-KNN, we will obtain five evaluation values from five emotion categories. Moreover, each evaluation value of emotion can be plotted in Emotion Radar Chart that is described in detail in section 3.3.

3 Experimental Results

The weighted D-KNN classification is used in our experiment. All experiments used the MATLAB

software and all results are based on the Leave-One-Out cross-validation which is a method to estimate the predictive accuracy of the classifier [9]. The extracted acoustic features were MFCC and LPCC.

3.1 Experimental Results of Weighted D-KNN Classifier

In the beginning of the experiment, we try to assign different weighting series to the calculation. These series are often used in lots of previous studies that were not limited in the field of signal processing. In addition, the constraint $w_1 \geq w_2 \geq \cdots \geq w_k$ were enforced in the weighting series lookup process. Three different series, from 10 to 1, the power of 2 and Fibonacci series, were chosen as our weighting series.

In KNN based classification, larger weighting values in the series are more important. So, in the case of Fibonacci series, such assumption is not groundless in our experiments. In our assumption, we want to assign the series that the importance of a certain value in the series is equal to the importance of the sum of two values behind that value.

Table 4: Comparison of weighted D-KNN using different weighting schemes

Weighting Scheme	Accuracy (%)
$w_i = k - i + 1$ **(k→1)**	75.39
$w_i = 2^{k-i}$ **(The power of 2)**	78.86
$w_i = w_{i+1} + w_{i+2}$, $w_k = w_{k-1} = 1$ **(Fibonacci series)**	79.31

The experimental results of the weighted D-KNN with different weighting series are summarizes in Table 4. Their corresponding confusion matrices are given in Table 5 to Table 7. The results show that different weighting scheme have different ability and property. The best accuracy is obtained with Fibonacci series scheme, and the best recognition rate is 79.31%.

Table 5: Confusion matrix of weighted D-KNN (*k*=10, weighting: 10→1)

Accuracy (%)	Angry	Happy	Sad	Bored	Neutral
Angry	88.74	3.97	2.65	0	4.64
Happy	22.92	54.17	6.25	0	16.67
Sad	4.03	1.61	79.03	6.45	8.87
Bored	0	0	9.64	84.34	6.02
Neutral	0.86	4.31	12.93	11.21	70.69

Table 6: Confusion matrix of weighted D-KNN (k=10, weighting: the power of 2)

Accuracy (%)	Angry	Happy	Sad	Bored	Neutral
Angry	90.07	5.29	1.99	0	2.65
Happy	19.79	61.46	4.17	0	14.58
Sad	3.23	2.42	82.26	2.42	9.68
Bored	0	2.41	6.02	85.54	6.02
Neutral	0.86	6.03	7.76	10.35	75.00

Table 7: Confusion matrix of weighted D-KNN (k=10, weighting: Fibonacci series)

Accuracy (%)	Angry	Happy	Sad	Bored	Neutral
Angry	90.73	4.64	1.32	0	3.31
Happy	18.75	62.50	3.13	0	15.63
Sad	4.03	2.42	82.26	2.42	8.87
Bored	0	1.20	8.43	84.33	6.02
Neutral	0.86	5.17	6.90	10.35	76.72

3.2 Weighting Optimization

Furthermore, we try to optimize the weighting series based on weighting value we used in last subsection. In addition, weighting series are also follow the constraint $w_1 \geq w_2 \geq \cdots \geq w_k$. In the experiment, we modified one weighting value and kept others fixed. The modification was done from right to left or from left to right, in the process of searching the optimum weighting values.

Table 8: Recognition accuracy of different optimum weighting series

Weighting Scheme	Accuracy (%)	
	From Left to Right	From Right to Left
k→1	78.44	77.13
The power of 2	79.07	79.31
Fibonacci series	79.52	79.55

In the next experiment, we try to optimize the weighting series that were used in section 3.1 in accordance with the directions from left to right and from right to left respectively. Table 8 shows the recognition accuracy of each optimized series, and we can find that weighted D-KNN classifier with optimum weighting series yields better results than without optimum weighting series: 3.05% improvement for weighting scheme in k→1, 0.45% improvement for the power of 2, and 0.24% improvement for Fibonacci series. The best recognition accuracy of 79.55% is obtained with weighted D-KNN optimized based on Fibonacci series. The corresponding confusion matrix is given in Table 9.

Table 9: Confusion matrix (optimized weighting from right to left with Fibonacci series)

Accuracy (%)	Angry	Happy	Sad	Bored	Neutral
Angry	90.73	4.64	1.32	0	3.31
Happy	18.75	62.50	3.13	0	15.63
Sad	4.032	2.42	82.26	2.42	8.87
Bored	0	1.20	7.23	85.54	6.02
Neutral	0.86	5.17	6.90	10.35	76.72

3.3 Emotion Radar Chart

An emotion radar chart is a multi-axes plot. Each of the axes stands for one emotion category. In our system, emotion radar chart just look like a regular pentagon as shown in Fig. 4. Figure 4 is an Emotion Radar Chart plotted using the data from Table 10 and 11. We can find that this input data is closed to angry emotion, and anger intensity of the speech is greater than the other emotions.

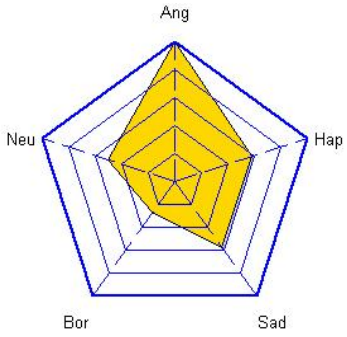


Figure 4: Emotion Radar Chart of test data with angry emotion

Table 10 shows the fifty distance values, 10 neighbors from each emotion class nearest to the input test data which is an angry speech. For example, first row shows the first 10 distances from input test data to training data of angry emotion. Here we call the value of the first row the distance set of Anger, and detailed description and operation is described in section 2.4. We can see clearly that the minimum distance in each round is almost the distance from input test data to the training data of angry emotion. Table 11 shows the calculation result of each distance set obtained by weighted D-KNN classification.

Table 10: Distance measured by M-KNN with k =10

Round	1	2	3	4	5	6	7	8	9	10
Angry	8.17	9.62	9.64	10.23	11.44	11.53	11.62	12.58	12.66	12.67
Happy	11.26	11.72	13.16	13.80	14.65	11.53	11.62	12.58	12.66	12.67
Sad	11.34	12.21	12.83	13.06	13.21	15.24	15.91	16.14	16.17	16.64
Bored	16.40	19.04	19.06	19.20	19.29	19.67	19.85	20.02	20.17	20.26
Neutral	11.96	12.40	14.55	15.12	15.57	15.72	15.74	15.87	15.95	16.09

Table 11: Evaluation value obtained by weighted D-KNN (Normalized with the maximum)

Emotion	Anger	Happiness	Sadness	Boredom	Neutral
Evaluation Value	1.0000	0.6032	0.5768	0.2699	0.5048

3.4 System Interface

Figure 5 is the user interface of our system. First, the source of the test speech has to be chosen from Source block. Test speech can get from disk or from recording. Second, after choosing the source, the Evaluation button in the Evaluation block can be pressed to plot the emotion radar chart on the lower graph. Finally, the Message frame will show the current state or error message, and Result block shows the recognition result of emotion of the test speech.

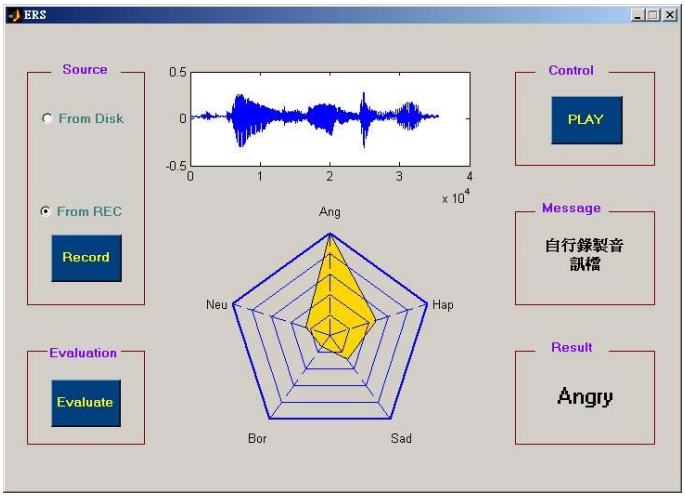


Figure 5: System interface (Evaluate test data from recording)

4 Conclusions

In this paper, we optimized the weights in weighted D-KNN to improve the recognition rate in our emotion recognition system. That is, we tried to modify slightly the weights in weighted D-KNN, and the accuracy of emotion recognition increased. The highest recognition rate of 79.55% is obtained with weighted D-KNN optimized based on Fibonacci series.

We also propose an emotion recognition and evaluation system. We regard the system as a computer-assisted emotional speech training system. For hearing-impaired people, it could provide an easier way to learn how to speak with emotion more naturally or help speech therapist to guide hearing-impaired people to express correct emotion in speech.

In the future, it is necessary to collect more acted or spontaneous speech sentences. Furthermore, it might be useful to measure the confidence of the decision after performing classification. Based on confidence threshold, classification result might be classified as reliable or not. Moreover, we also want to make the emotion evaluation more effectively, and a more user friendly interface of system for

hearing-impaired people needs to be designed. Besides, how to optimize the weights in weighted D-KNN to improve the recognition rate in emotion recognition system is still a challenge work.

5 Acknowledge

A part of this research is sponsored by NSC 93-2213-E-036-023.

References

[1] P. Ekman, *Darwin and Facial Expressions*, Academic, New York, 1973.
[2] M. Davis and H.College, *Recognition of Facial Expression*, Arno Press, New York, 1975.
[3] K. Scherer and P. Ekman, *Approaches to Emotion*, Lawrence Erlbaum Associates, Mahwah, NJ, 1984.
[4] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, S. Stroeve, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark," *ISCA Workshop on Speech and Emotion*, Belfast, 2000.
[5] F. Dellaert, T. Polzin and A. Waibel, "Recognizing Emotion in Speech," *Fourth International Conference on Spoken Language Processing*, Vol. 3, 1996, pp. 1970-1973.
[6] J. Nicholson, K. Takahashi, R. Nakatsu, "Emotion Recognition in Speech Using Neural Networks," *6th International Conference on Neural Information Processing*, ICONIP '99, Vol. 2, 1999, pp. 495-501.
[7] L. S. Chan, H. Tao, T.S. Huang, T. Miyasato, and R. Nakatsu, "Emotion Recognition from Audiovisual Information," *IEEE Second Workshop on Multimedia Signal Processing*, 1998, pp. 83-88.
[8] Tsang-Long Pao, Yu-Te Chen, Jhih-Jheng Lu and Jun-Heng Yeh, "The Construction and Testing of a Mandarin Emotional Speech Database," *Proceeding of ROCLING XVI*, Sep. 2004, pp. 355-363.
[9] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh and Jhih-Jheng Lu, "Detecting Emotions in Mandarin Speech," *Proceeding of ROCLING XVI*, Sep. 2004, pp. 365-373.

閩南語語句基週軌跡產生：兩種模型之混合與比較

Min-Nan Sentence Pitch-contour Generation: Mixing and Comparison of Two Kinds of Models

古鴻炎　　黃維
Hung-Yan Gu　and　Wei Huang

國立台灣科技大學資訊工程系
Dept. CSIE, National Taiwan University of Science and Technology
e-mail: guhy@mail.ntust.edu.tw　http://guhy.csie.ntust.edu.tw

摘要

本文對兩種基週軌跡的產生模型，SPC-HMM 和 ANN，作改進與擴充然後用以建造閩南語的基週軌跡產生模型，希望以閩南語作為工作語言，而其它的語言(如國語、客家語)的基週軌跡，能夠以調適過的閩南語所訓練出模型，來直接產生。除了研究閩南語的 SPC-HMM 和 ANN 模型的建造，我們也對此二種模型作了內外部測試之比較，及嘗試作模型混合的研究，希望能藉以提升效能。對於所建造的閩南語基週軌跡 SPC-HMM、ANN、和混合模型，我們也作了聽測評估，初步的結果顯示自然度方面，混合模型的分數比 ANN 模型的好，而 SPC-HMM 模型的則居末，所以混合模型在小型訓練語料下，是一個不錯的選擇。

關鍵詞： 語音合成，基週軌跡，類神經網路，隱藏式馬可夫模型

1. 前言

語音合成的研究，最近有不少人採取 corpus based 的研究方向[1,2,3,4]。大家也許會認爲，只要 copus 夠大，就不需要作音調(基週軌跡)調整之信號處理，以保持合成單元的信號清晰度與自然度，亦即不作基週軌跡模型建造和參數值產生的動作。可是實務上，要錄音及製作夠大的 corpus，所需的人力與經費，並不是一般的研究者能夠負擔得起，即使以程式自動作標音(labeling)和斷音(segmenting)，也仍然需要作人工的校正。然而當 corpus 不夠大時，會遇到的一個和基週軌跡有關的問題是，合成出的句子的語調，缺乏人類說話時的自然下傾(declining)的表現[5]，這會讓聽者較難掌握說話的節奏，而另一個可能發生的情況是，在合成單元的邊界，音調未能平順銜接。此外，還存在的一個明顯問題是，合成語句的說話速度會忽快忽慢地不一致，原因應是來自不同原始語句的合成單元，發音的速度不匹配。基於前述的問題，corpus based 的語音合成作法，除非是 corpus 夠大且錄音者能夠一直維持說話的平穩，不然還是有需要建造模型來

產生基週軌跡，用以檢驗所合成的語句，其語調是否有下傾的表現，相鄰音節間的基週軌跡是否有不平順銜接的地方。

其實本文研究閩南語語句基週軌跡之產生，是從另一個方向來思考台灣地區的語音合成研究的問題，就是強勢語言的文句翻語音(text-to-speech, TTS)研究，匯聚了大部分的研究資源(人力、經費)，而當要對另一弱勢語言作 TTS 之研究時，仍然需要再投入大量的資源，這樣的情況對於 corpus based 之研究方向應是很明顯的。因此我們開始思考，如何讓一種語言的語音合成之研究成果，能夠很經濟地轉移給另一個語言使用，如此弱勢語言面臨的存續問題，就可獲得至少一些些的舒解。對於同樣是以音節爲組成單位的聲調語言來說，例如國語(北京話)、閩南話、客家話、廣東話...等，我們可先選擇其中一個作爲工作語言(working language)，來研究它的韻律參數產生模型，然後透過模型調適，以工作語言訓練的模型，來產生出標的語言(target language)的韻律參數。這樣的想法，先前我們已曾經以基週軌跡參數爲例，選擇閩南語作爲工作語言，來建造它的 SPC-HMM 模型[6]，然後調適此模型以產生出國語和客家語語句的基週軌跡，初步實驗結果顯示，我們的想法是可行的[7]。

台灣的三種主要語言：國語、閩南語、客家語，我們所以會選擇閩南語作爲工作語言，是因爲它的聲調數量(7 個)和基本音節數量(785 個)都是最多的，並且它的聲調類型可含蓋客家語的(四縣腔 6 個、海陸腔 7 個)，及國語裡的前四個，而國語裡的輕聲也可以閩南語的低入聲來近似。雖然先前我們曾研究建造閩南語基週軌跡的 SPC-HMM 模型，不過有不少研究成果指出類神經網路(ANN)模型具有不錯的效能[8, 9]，因此我們覺得對於工作語言的基週軌跡模型，再嘗試以 ANN 模型或以 SPC-HMM 和 ANN 之混合來提升效能，是值得去研究、探討的，如此對於標的語言的效能也將會獲得改進。所以，本論文先在相同訓練語料的情況下，比較個別的 SPC-HMM 和 ANN 模型的基週軌跡預測誤差，再嘗試以模型混合之作法來降低預測誤差。由於訓練語料的大小、來源不一樣，所以本文得到的基週軌跡效能，並未和他人的研究成果作比較。

2. 訓練語句錄音與基週軌跡之前處理

基週軌跡模型的訓練與測試語料，都是由本文第二作者在實驗室以麥克風發音直接錄音到電腦中，取樣率爲 22,050Hz，樣本值寬度 16bits，總共錄了 643 句(3,696 音節)閩南語句作爲訓練之用，65 句(437 音節)閩南語句作爲外部測試之

用，各句分別存成一個音檔。發音用的文句主要是取自閩南語歌曲的歌詞，部分則是自行造句，聲調大部分是採變過調的以方便唸讀錄音，在錄音之前我們已對文句作過分析篩選，以確保訓練語句裡前後三音節的聲調組合，能夠含蓋所有可能的聲調組合，實際情形是各種組合最少都出現三次，而出現次數在 3 至 5 次之間的有 194 種組合，在 6 至 10 次之間的有 126 種組合。

錄音後，接著進行基週頂點標記和音節邊界標記的動作，先以程式作自動偵測，再由人工更正錯誤的標記。由於錄得的各音節的時間長度不相同，因此我們先作時間正規化之處理，讓各個音節的基週軌跡都以相同的 16 維度(dimensions)的頻率向量 $<f_0, f_1, \ldots, f_{15}>$來表示，$f_k$ 表示在音節有聲部分時間比例 k/15 地方的基頻頻率值取對數，該頻率值可依基週頂點標記資料去內差得到[6]。

由於訓練及測試語句的錄音是分散在很多天裡，錄音者很難一直維持在同一種精神狀況與心情下來錄音，而使得錄到的語句有的音調較高有的音調較低，若不作音高正規化之處理，則模型產生出的句子基週軌跡會有高低起伏的不平順銜接的現象。因此作完時間正規化處理之後，接著還需進行音高正規化的處理。一種簡單且有效的正規化方法是[6]: (a)求各個音節的平均音高，亦即求取頻率向量的 16 個維度數值的平均；(b)求第 k 個語句的平均音高 HS_k，亦即求取該語句的組成音節的音高的平均值；(c)求總體語句的平均音高 HT，亦即將各個語句的音高加總再除以語句數量；(d)依據 $HD_k = HS_k - HT$，將第 k 個語句的各個組成音節的音高減去 HD_k，亦即頻率向量的各個維度都要減去 HD_k。

3. 句子基週軌跡 HMM 模型

由於整句話的行進(語調)對於音節基週軌跡的影響是不易精確描述的，因此先前我們對國語作了這樣的研究[10][6]，即以 HMM 裡的隱藏狀態來描述，音節基週軌跡在一句話行進中的不同時間位置所受的不同影響，稱爲 SPC-HMM(sentence-pitch-contour HMM)。其觀念是以 HMM 的三個隱藏狀態來對應一個句子(或呼吸群)內的”句首”、”句中”與”句尾”等三個隱含的韻律狀態，如圖 1 所示。至於 HMM 的觀測(observation)值，我們採用離散式的觀測符號，且令一個語句的各個音節分別產生出一個觀測符號。由於考慮時刻 t 時的音節基週軌跡至少會受到前、後及本音節聲調的影響，所以我們定義觀測符號爲，前一個音節聲調、本音節聲調、下一個音節聲調及本音節基週軌跡量化碼等四個因素

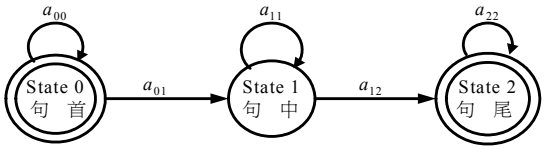


圖 1　韻律狀態轉移圖

之組合，即令時刻 *t* 時的觀測符號爲:

$$O_t = G_{t-1} \times G_t \times G_{t+1} \times V_t \tag{1}$$

G_t 表示句子中第 *t* 個音節的聲調, $0 \le G_t \le 6$

V_t 表示句子中第 *t* 個音節的基週軌跡量化碼, $0 \le V_t \le 7$

由於公式(1)中用到的V_t，是音節基週軌跡之向量量化碼，因此我們必須先訓練閩南語七個聲調各自的碼書(code book)，這裡採用了 K-means Clustering 演算法 [11]，向量量化之距離量測是均方根(root mean square)距離，其公式爲

$$d(X,Y) = \sqrt{\frac{1}{16}\sum_{k=1}^{16}(x_k - y_k)^2} \tag{2}$$

至於向量量化碼書大小的選擇，如果碼字越多的話，則量化誤差會比較小，但是對於後面的 HMM 的訓練則需要有更多的訓練語句來訓練，才能使 HMM 內部測試之誤差改進。所以，在有限的訓練語句之下，我們選擇將各個聲調的音節基週軌跡都量化成 8 類。

HMM 模型的訓練語句當然是愈多愈好，如果訓練語句不足，將使某一種音節聲調之組合沒有出現過，此時可用降階機制和資料分享來解決語料不足的問題。降階機制的第一層表示沒有降階，就是使用公式(1)來產生出觀測符號；第二層是，只考慮前一個音節的聲調、本音節的聲調和本音節的基週軌跡量化碼的組合。然後各層各自訓練，以得到該層所對應的 a_{ij}、$b_j(k)$等參數。解決訓練語料不足的方法，除了降階機制外，資料分享是另一方法，其主要觀念是將一個觀測符號的出現機率分享給距離量測上最近的另外幾個觀測符號，來提升未出現或出現機率較少的觀測符號的機率值。我們的作法是，一個音節組合出的觀測符號，它的出現機率分享給距離最近的另外兩個，相同聲調組合但是由不同基週軌跡量化碼字所形成的觀測符號，分享的比率爲 0.99, 0.005, 0.005。

以 SPC-HMM 模型來作一個語句的基週軌跡的合成，輸入的是聲調序列而沒有基週軌跡量化碼字的資料(這和訓練階段的情況不同)，因此當對於第 t 個音節的聲調，要以公式(1)來組合出觀測符號時，會有八種觀測符號可供選擇，此時必須將訓練階段的最佳路徑搜尋之維特比演算法作修正，將原本由時間軸和狀態軸構成的搜尋平面，擴充成時間軸、狀態軸和軌跡量化碼字軸之三度搜尋空間，再配合 HMM 模型的 a_{ij}、$b_j(k)$等參數，以找出最佳的三度搜尋空間裡的路徑，再作回朔(back tracking)，以確定各音節所選到基週軌跡量化碼字。實作上，可依上游文字分析所得的呼吸群、詞邊界訊息來直接設定各音節所停留的狀態，如此可產生出更自然的句子基週軌跡。

4. 句子基週軌跡 ANN 模型

本文依據前人研究 ANN 來產生國語語句基週軌跡的經驗[12]，作進一步的研究改進，以用來產生閩南語語句的基週軌跡。這裡所採用的是遞迴式類神經網路，其結構如圖 2 所示，分爲四層:輸入層、隱藏層、隱藏遞迴層、輸出層。輸入層有 28 個輸入單元，來輸入語境參數；輸出層有 16 個單元，以表示一個音節之基週軌跡(即 16 維度的頻率向量)；隱藏層的單元數，依實驗結果設定爲 30。內部鏈結的權重值需經由學習來決定數值，這裡使用的是遞迴學習演算法[13]。

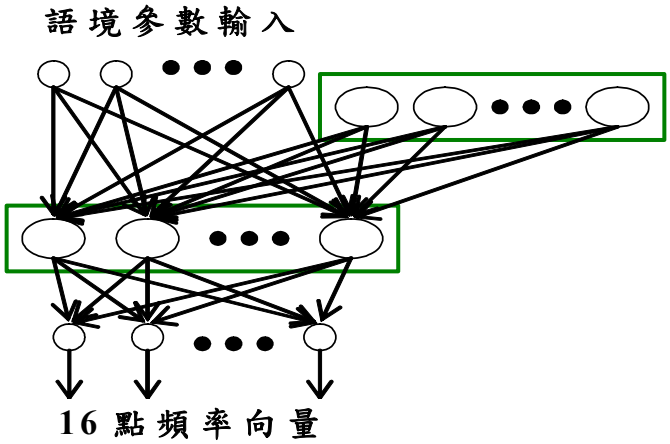


圖 2　遞迴式類神經網路之結構

關於 ANN 模型的輸入，本文使用以音節爲單位的語境參數：聲母、韻母、聲調、時間比例等參數。由於語音是時序性的資訊傳遞，所以除了要考慮本音節的聲調種類、聲母類型、韻母類型，也要考慮到前一個音節的聲調種類和韻母類型，以及後一個音節的聲調種類和聲母類型，此外使用了一個時間比例參數，以

代表本音節在整句話中之時間進度，此參數相當於傳入韻律狀態的資訊。上述的語境參數共使用了 28 個 bits，詳細情形如表 1 所示，表中從左起依序爲輸入層

表 1 ANN 輸入層的資料項目及 bit 數

項目	上個音節聲調	上個音節韻母類別	本音節聲調	本音節聲母	本音節韻母	$\frac{t}{N+1}$	下個音節聲調	下個音節聲母類別
bits	**3**	**4**	**3**	**5**	**6**	**1**	**3**	**3**

項目和其所佔的 bit 數。其中前一個音節的韻母類別佔 4 bits，因爲它被**粗**分類成 12 類，下一個音節的聲母類別佔 3 bits，因爲有 6 個聲母之**粗**分類類別，本音節聲母佔 5 bits，因爲有 18 個聲母之細分類，本音節韻母佔 6 bits，因爲有 60 個韻母之細分類。本音節及前後兩音節的聲調三個項目都各佔 3 bits，因爲閩南語有 7 個聲調。表中另一項目 $t/(N+1)$ 用以指示本音節在句子裡的位置，N 代表句子的總字數，這個項目用一個浮點數值表示，不像其他項目是用二進位數值表示。

聲母、韻母在細分類(原始分類)之外，這裡還另外作粗分類的原因是，聲母和韻母的種類太多，使得本音節和前後音節之聲母、韻母組合的數量過多，這意味我們必須準備目前人力難以達成的龐大訓練語料，否則不能讓所有聲母、韻母組合有足夠的出現次數，而必然會影響到 ANN 模型的效能[14]。所以我們採取前後音節的聲母、韻母先作粗分類再作組合的方式，以減少組合的數量，讓每個組合在訓練語料中有足夠的出現次數。粗分類僅用於前後音節之聲母、韻母，本音節之聲母、韻母仍採用細分類。關於閩南語聲母的粗分類，這裡依據前人研究國語粗分類的觀念[12, 14]，分類成如表 2 所示的方式，國語和閩南語都有的聲

表 2 閩南語聲母之粗分類(通用拼音符號)

類別	聲母
1	零聲母, m, n, l, r, ng, q, v
2	h, s
3	b, d, g
4	z
5	p, t, k
6	c

母放在相同的類別上，而閩南語才有的/ng, q, v/，因爲都是有聲的子音，所以將它們放在第一類上。關於閩南語韻母的粗分類，詳細的分類方式如表 3 所示，雖然基本上參考了國語韻母粗分類的觀念，但我們也依據發音口形作了一些修

表 3 閩南語韻母之粗分類(通用拼音符號)

類別	韻母
1	空韻母
2	-a, -ia, -ua
3	-o, -io
4	-er
5	-e, -ue
6	-ai, -uai, -au, -iau
7	-i, -u, -ui, -iu
8	-ing, -eng, -in, -un, -en
9	-ang, -iang, -uang, -ong, -iong, -an, -uan
10	-am, -iam, -im, -om
11	-ah, -eh, -ih, -oh, -uh, -auh, -erh, -iah, -ierh, -ioh, -uah, -ueh
12	-ap, -iap, -ip, -op, -at, -et, -it, -uat, -ut, -ak, -iak, -ik, -iok, -ok

改，例如把/-au, -iau/改放於第 6 類，把/-ing, -eng/改放於第 8 類，把/-an, -uan/改放於第 9 類。另外，第 11、12 類是國語所無的閩南語入聲韻母，第 10 類是國語所沒有的其它閩南語韻母。

5. 模型混合之方法

這裡的模型混合不是要把兩個模型合併成一個模型，再用以產生基週軌跡，而是將兩個模型各自產生的輸出作組合，以得到一個新的輸出。包括國語的漢語方言的電腦語音合成上，過去研究者提出的 HMM 和 ANN 模型，各有其優缺點，作模型混合就是希望能夠取長補短以提升合成語音之品質，例如兩種模型分別在基週軌跡的穩定性和活潑性上佔優勢，則混合此兩種模型有可能產生出兼具穩定性和活潑性的語音。

關於混合的方法，本文研究後只找到兩種比較有實際效果的方式，在此稱爲: 簡單加權方式和 16 維度標準差加權方式。簡單加權方式就是將 HMM 和 ANN 模型所產生的基週軌跡值，直接依照各種加權比例去組合出新的基週軌跡值；16 維度標準差加權方式，則是針對兩個模型產生的基週軌跡之 16 個維度來分別作加權，音節的各個維度分別當成一個集合來求出標準差，因此共有 16 個標準差，再利用兩個模型各自的 16 個標準差，作爲權重比例。簡單加權方式的公式和 16 維度標準差加權方式的公式分別如下所列:

$$F^S = W_H \cdot F^H + (1 - W_H) \cdot F^A \tag{3}$$

$$f_j^M = \frac{W_H \cdot \frac{\sigma_j^A}{\sigma_j^H} \cdot f_j^H + (1 - W_H) \cdot f_j^A}{W_H \cdot \frac{\sigma_j^A}{\sigma_j^H} + (1 - W_H)} \quad , \quad j=1, 2, \cdots, 16 \qquad (4)$$

其中 F^S 表示簡單加權方式混合後的基週軌跡頻率向量，W_H 表示 HMM 模型的權重，F^H 與 F^A 分別表示 HMM 和 ANN 模型產生出的頻率向量；f_j^M 表示 16 維度標準差加權方式混合後第 j 維度的頻率值，f_j^H 與 f_j^A 分別是 F^H 與 F^A 的第 j 維度的頻率值，σ_j^A 表示 ANN 模型產生的頻率向量第 j 維度之標準差，σ_j^H 表示 HMM 模型產生的頻率向量第 j 維度之標準差。公式 4 裡，我們依 σ_j^A 與 σ_j^H 的比率來個別調整各個維度中 ANN 與 HMM 兩模型的加權值 W_H 與 $1-W_H$。

爲了分析混合後模型的效能，在此我們先以模型之訓練語句作爲輸入，來計算兩模型輸出的頻率向量之間的相關係數，結果發現 16 個維度各別的相關係數 R_j 都在 0.95 左右(0.944 至 0.972)，對於如此高的相關性，我們大約已經知道，混合後模型的預測能力，應不會有很大幅度的改變。我們量測相關係數的公式如下:

$$R_j = \frac{\frac{1}{N}\sum_{k=1}^{N}(f_j^H(k) - g_j^H) \bullet (f_j^A(k) - g_j^A)}{\sqrt{\frac{1}{N}\sum_{k=1}^{N}(f_j^H(k) - g_j^H)^2} \bullet \sqrt{\frac{1}{N}\sum_{k=1}^{N}(f_j^A(k) - g_j^A)^2}} \quad , \; j = 1,2,...,16 \qquad (5)$$

其中 N 表示訓練語句裡的總音節個數，$f_j^{H(k)}$ 與 $f_j^{A(k)}$ 分別表示第 k 個音節的 f_j^H 與 f_j^A，而 g_j^H 與 g_j^A 分別表示 HMM 和 ANN 模型產生出的頻率向量的第 j 維度的平均值，即

$$g_j^H = \frac{1}{N}\sum_{k=1}^{N} f_j^H(k) \quad , \quad g_j^A = \frac{1}{N}\sum_{k=1}^{N} f_j^A(k) \quad , \quad j = 1,2,...,16 \qquad (6)$$

6. 模型效能之比較

依據公式(3)與(4)，接著我們進行模型混合之實驗，在內部測試(inside test)時，使用的測試語句就是模型訓練所用的 643 個語句，而在外部測試(outside test)時，則使用另外的 65 個未參加模型訓練的語句。一個音節的原始基週軌跡和模型預測該音節所輸出的基週軌跡，兩者之間以公式(2)來量測誤差距離。當採簡單加權方式時，我們得到如表 4 所示的誤差數值；而當採 16 維度標準差加權方式時，我們得到如表 5 所示的誤差數值。

表 4 簡單加權式模型混合之預測誤差

權重 W_H	Inside test			Out test		
	AVG	STD	MAX	AVG	STD	MAX
-0.1	0.0396	0.0193	0.1579	0.0545	0.0384	0.4299
0	0.0386	0.0188	0.1563	0.0530	0.0386	0.4324
0.1	0.0380	0.0185	0.1601	0.0528	0.0394	0.4350
0.2	0.0377	0.0185	0.1683	0.0539	0.0407	0.4376
0.25	0.0377	0.0186	0.1725	0.0549	0.0415	0.4390
0.3	0.0377	0.0188	0.1767	0.0561	0.0425	0.4403
0.4	0.0380	0.0195	0.1852	0.0592	0.0450	0.4431
0.6	0.0395	0.0217	0.2023	0.0674	0.0517	0.4488
0.8	0.0421	0.0248	0.2197	0.0776	0.0600	0.4548
1.0	0.0456	0.0282	0.2373	0.0891	0.0694	0.4826

表 5 十六維度標準差加權式模型混合之預測誤差

權重 W_H	Inside test			Out test		
	AVG	STD	MAX	AVG	STD	MAX
-0.1	0.0396	0.0193	0.1579	0.0543	0.0384	0.4301
0	0.0386	0.0188	0.1563	0.0530	0.0386	0.4324
0.1	0.0380	0.0185	0.1600	0.0528	0.0394	0.4348
0.2	0.0377	0.0184	0.1682	0.0537	0.0405	0.4373
0.25	0.0377	0.0186	0.1724	0.0546	0.0412	0.4386
0.3	0.0377	0.0188	0.1766	0.0557	0.0422	0.4399
0.4	0.0380	0.0194	0.1850	0.0586	0.0445	0.4426
0.6	0.0395	0.0217	0.2021	0.0666	0.0510	0.4483
0.8	0.0420	0.0247	0.2196	0.0770	0.0595	0.4544
1.0	0.0456	0.0282	0.2373	0.0891	0.0694	0.4826

在表 4 和 5 裡，AVG 表示所有參加測試的音節的平均預測誤差，STD 表示預測誤差的標準差，而 MAX 表示所有音節的預測誤差的最大值。首先比較 W_H 權重值爲 0 和爲 1 的兩列，W_H=0 代表只使用 ANN 模型，W_H=1 代表只使用 SPC-HMM 模型，由表 4 和 5 可看出，ANN 模型的基週軌跡預測誤差在 AVG 項分別是 0.0386 (內部測試)與 0.0530 (外部測試)，都比 SPC-HMM 模型的 0.0456 與 0.0891 來得小許多，0.0386 相當於 120Hz 音高時線性的 4.72Hz 的差異，而

0.0530 則相當於線性 6.53Hz 的差異；另外在 STD 和 MAX 項，ANN 模型的誤差也都是比 SPC-HMM 模型的好很多，所以個別的模型來說，ANN 模型的確比 SPC-HMM 模型具有比較準確的預測能力。

如果再考慮各種不同的 W_H 權重值來了解混合模型的效能，則由表 4 和表 5 可看出兩種加權方式都呈現相同的趨勢，以項目 AVG 來看，在內部測試方面，最好的權重值都是令 W_H = 0.25，而在外部測試方面，最好的權重值則是令 W_H = 0.1，以獲得較小的預測誤差平均值。在內部測試時，使用混合模型可以讓平均預測誤差，從 ANN 模型的 0.0386 降至 0.0377，下降幅度約爲 2.3%；在外部測試時，則可從 ANN 模型的 0.0530 降至 0.0528，下降幅度只有 0.4%。所以這裡所研究的混合方式，並不能夠大幅度提升效能，其原因之一應是，兩種模型所產生出的基週軌跡之間，已經具有非常強的相關性。

7. 聽測評估

由於本文的研究主題是基週軌跡之產生，所以在此初步的聽測評估裡，其它的韻律參數(如音長、音量)的產生，及信號波形的合成，都是直接沿用以前的成果[12,15,16]。音長、音量等韻律參數的產生，是採用簡單的規則式作法，而信號波形合成裡，每個合成單元(音節)只用一個固定的平調發音，合成方法則是 TIPW[16]，它可說是 PSOLA 的改進作法。基週軌跡的產生，分別使用了 ANN 模型、SPC-HMM 模型、及混合模型，混合方式採簡單加權式，權重值則設爲 W_H = 0.25，因爲它是內部測試裡 AVG 項最好的權值，在此不採外部測試的權值，因爲外部測試的句子數量不夠多，可信度我們仍存懷疑。對於模型輸出的 16 維度頻率向量，我們再於所關心的時間點附近，找出相鄰的 4 個維度的頻率值，作 Lagrange 內差，就可求得該時間點上的週期長度。至於文句分析的處理，我們可說是幾乎沒有作，因爲我們直接把如表 6 所示的拼音文句，輸入給所建造的語音合成系統，不過這些文句未參加模型之訓練。

參與聽測的測試者爲 15 人，其中 10 人爲各實驗室的研究生，年齡在 20~30 歲，另外 5 人是年齡在 30~50 歲的鄰居，我們以可辨度和自然度來衡量所合成出來的語音品質，可辨度是指測試者聽得懂合成語音的程度，自然度是指合成語音像人類語音的程度，這裡分成五個評分段供試聽者作爲評分標準：9.0 ~ 10(非常像人類語音)、8.0 ~ 8.9(很像人類語音)、 7.0 ~ 7.9(接近人類語音)、6.0 ~ 6.9(及

表6 聽測用的閩南語文句

	文句	通用拼音(聲調採傳統編號)
1	我有滿腹的心聲	qua-1 wu-3 <mua-1 bak-4> e-7 <sim-7 siann-1>.
2	有話想要對你講	<wu-3 we-7> <siunn-3 veh-8> <dui-2 li-1> gong-2.
3	無講誰人會知影	vo-7 gong-2 <siann-1 lang-5> e-3 <zai-7 yann-2>.
4	有啥麼代誌	wu-3 <siann-1 mi-1> <dai-3 zi-3>.
5	我攏總聽無	qua-2 <long-1 zong-1> <tiann-7 vo-5>.
6	請汝講卡大聲	<ciann-1 li-2> gong-1 kah-8 <dua-3 siann-1>.

格)、5.9 以下(劣等)。自然度聽測時，隨機播放三種基週軌跡模型合成出的語音檔，然後再選擇以自然度最高的語音檔作可辨度評估，試聽者每聽完一句就將該句覆述一遍，以便記錄聽錯的音節，聽完全部語句後就可計算可辨度，可辨度定義爲聽對的音節數佔總音節數的比率，最後依據 15 人的評分來算出平均值。結果我們得到如表 7 所示的數值，由此數值可知，ANN 模型的自然度 7.2 分比

表 7　合成語句的聽測評分

	混合模型	ANN模型	SPC-HMM模型
自然度	**7.4**	**7.2**	**6.9**
可辨度	**96.0%**		

SPC-HMM 模型的 6.9 分好一些，而混合模型的自然度 7.4 分又比 ANN 模型的好一些，這剛好和表 4 和 5 裡的預測誤差數值，呈現一致的走勢。另外在可辨度方面，96%可以被聽對，我們覺得是不錯的，因爲通常短的語句比較難聽得懂。

不過，SPC-HMM 模型的基週軌跡預測誤差，比另二者大許多，而混合模型的預測誤差僅比 ANN 模型的改進一些些，但是在聽測上，SPC-HMM 模型的 6.9 分也不會比 ANN 模型的 7.2 分差很多，而混合模型的 7.4 分，則也明顯的有改進。這樣的現象，我們認爲是因爲，SPC-HMM 模型和 ANN 模型的輸出之間，基本上具有非常強的相關性，而 SPC-HMM 模型的預測誤差，具有比 ANN 模型大許多的直流成分(電學觀念)，這應是肇因於向量量化處理，實際上我們計算基週軌跡預測誤差在各維度上的平均值得知，SPC-HMM 模型的誤差平均各維度都約在 0.013 左右，而 ANN 模型的誤差平均，各維度則大多在絕對值 0.002 以下。另外，ANN 模型偶而會有偏移量大許多的誤差值出現(比較不穩定)，而 SPC-HMM 模型則比較無此種情況(較穩定)，所以混合模型在聽測上，會表現得有一些改進。

8. 結論

本文將過去研究國語基週軌跡產生的 SPC-HMM 模型和 ANN 模型作更改與擴充，以用來訓練及建立閩南語的基週軌跡模型，考慮了更多的閩南語聲調，及聲母、韻母的粗分類問題。此外，我們也比較了兩種模型的效能，及嘗試作模型的混合，希望能夠藉以提升效能，所嘗試的兩種混合方式是，簡單加權方式、和十六維度標準差加權方式。經由內、外部的測試實驗後，結果顯示混合模型與 ANN 模型的效能都比 SPC-HMM 模型的好很多，並且混合模型的又比 ANN 模型的好一些。

此外，我們也把模型產生出的基週軌跡拿去合成出語音信號，再作主觀的聽測評估，結果在自然度方面，混合模型得到 7.4 分，比 ANN 模型的 7.2 分好一些，而 SPC-HMM 模型也可得到 6.9 分，雖然說 SPC-HMM 模型的基週軌跡預測誤差，在內、外部測試裡都比另二者差很多。在可辨度方面，結果顯示 96% 的字可被聽對。由於本文的研究裡，訓練的語句共只有 3,696 個音節，算是小型語料的情況，所以在小型語料的情況下，混合模型可說是建造基週軌跡模型的不錯選擇。

9. 致謝

感謝國科會計畫的支援，計畫編號 NSC 92-2213-E-011-078.

參考文獻

[1] Sagisaka, Y., *et al.*, "ATR v-talk Speech Synthesis System", ICSLP'92, Canada, pp. 483-486, 1992.

[2] Chou, Fu-chiang, Corpus-based Technologies for Chinese text-to-Speech Synthesis, Ph. D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, 1999.

[3] Min Chu, *et al.*, "Microsoft Mulan - a bilingual TTS system", ICASSP '03, Vol. 1, pp. I264-I267, 2003.

[4] 張唐瑜，以大量詞彙作爲合成單元的中文文轉音系統，碩士論文，國立中興大學資科所，2005。

[5] O`Shaughnessy, D., Speech Communication: Human and Machine, 2’nd ed., IEEE Press, 2000.

[6] Gu, H. Y. and C. C. Yang, "A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", ISCSLP’2000, Beijing, pp. 125-128, 2000.

[7] Gu, H. Y. and H. C. Tsai, "A Pitch-Contour Model Adaptation Method for Integrated Synthesis of Mandarin, Min-Nan, and Hakka Speech", 9’th IEEE Int. Workshop on Cellular Neutral Networks and their Applications (Hsinchu, Taiwan), pp. 190-193, 2005.

[8] Chen, S. H., 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, 1998.

[9] Lin, C. T., R. C. Wu, J. Y. Chang, and S. F. Liang, “A Novel Prosodic-Information Synthesizer Based on Recurrent Fuzzy Neural Network for the Chinese TTS System”, IEEE trans. Systems, Man, and Cybernetics, Vol. 34, No. 1, pp. 309-324, Feb. 2004.

[10] 楊仲捷，基於 VQ/HMM 之國語語音合成基週軌跡產生之研究，碩士論文，國立台灣科技大學電機所，1999。

[11] Rabiner, L. and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[12] 曹亦岑，使用小型語料類神經網路之國語語音合成韻律參數產生，碩士論文，國立台灣科技大學電機所，2003。

[13] Lee, S. J., K. C. Kim, H. Y. Jung, and W. Cho, "Application of Fully Recurrent Neural Networks for Speech Recognition", ICASSP’91, pp. 77-80, 1991.

[14] 郭威志，使用語者辨認做前處理之國語 TTS 系統發展，碩士論文，國立交通大學電信系，2000。

[15] 李雪貞，客語語音合成之初步研究，碩士論文，國立台灣科技大學資工所，2001。

[16] Gu, H. Y. and W. L. Shiu, "A Mandarin-syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control", Proceedings of the National Science Council, Republic of China, Part A: Physical Science and Engineering, Vol.22, No.3, pp.385-395,1998.

Statistical Analysis of Two Polarity Detection Schemes in Speech Watermarking

Bin Yan[1], Zhe-Ming Lu[1], Jeng-Shyang Pan[2], Sheng-He Sun[1]

[1]Department of Automatic Test and Control, Harbin Institute of Technology,
P. O. Box 339, 150001 Harbin, P. R. China.
yanbinhit@hotmail.com
zhemingl@yahoo.com
[2]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences,
415 Chien-Kung Road, Kaohsiung 807, Taiwan, R.O.C.
jspan@cc.kuas.edu.tw

Abstract

Polarity inversion based speech watermarking scheme hide data in speech by modification of the speech polarity. This paper build a statistical model of the polarity detection problem, based on this model, the original polarity detection scheme and the optimal detection scheme are analyzed and compared. The theoretical analysis results are validated by Monte Carlo simulation, the optimal polarity detector shows significant performance gain compared with the original polarity detection algorithm.

1 Introduction

Polarity Inversion (PI) based watermarking scheme utilizes the fact that the human auditory system (HAS) is insensitive to the polarity of the speech signal [1]. Secure data can be hidden in speech signal by inverting the polarity of certain portion of the signal. PI watermarking can be classified as phase coding scheme [2, 3], the phase of the speech is changed by 180 degrees for PI watermarking. PI is very robust against noise addition and filtering operations because the polarity of the voiced frame won't change under these manipulations. The drawback of PI watermarking is that it is not secure, but it is very useful for content annotation and in-band signalling applications [4]. The problem to be solved by this paper is to evaluate the performance of the polarity detection algorithms. This is done by first building a statistical model for the speech residual signal, based on this model, the performance of the original polarity detection algorithm is analyzed, this result is compared to the performance of the optimal polarity detector. Finally, we perform Monte Carlo simulation to validate the theoretical analysis.

2 Detection Performance of the Original Method

The polarity detection scheme proposed by [1] can be summarized as a two-step procedure for each syllable, first, the polarity of the maximum peak in the LPC residual signal of each voiced frame is estimated, second, the polarity of each syllable is determined by majority vote. In this section we will analyze the error probability of this detection scheme. When the AR model order is properly chosen, the residual signal of the voiced frame can be modeled as impulse train in Additive White Gaussian Noise (AWGN), so we can build the following model under each hypothesis:

$$\mathcal{H}_0 \quad : \quad s[n] = -\sum_{k=0}^{K-1} A\delta[n-kP] + w[n] \tag{1}$$

$$\mathcal{H}_1 \quad : \quad s[n] = + \sum_{k=0}^{K-1} A\delta[n - kP] + w[n] \ , \tag{2}$$

where P is the pitch period, A is the amplitude of the impulse train, K is the number of pitch period in each frame, $w[n]$ is Gaussian noise with mean zero and variance σ^2. For ease of analysis, we made the following assumptions: the sample values in the location of impulses are not affected by the AWGN, the validity of this assumption will be verified by Monte Carlo simulation. Under this assumption, the probability of detection error in each frame is

$$P_{\mathrm{EF}} = \frac{1}{2} \Pr\left\{\max(\mathbf{s}) > A | \mathcal{H}_0\right\} + \frac{1}{2} \Pr\left\{\min(\mathbf{s}) < -A | \mathcal{H}_1\right\} \ ,$$

where $\mathbf{s} = [s[0], \cdots, s[N-1]]^T$, it is assumed that $P(\mathcal{H}_0) = P(\mathcal{H}_1) = 1/2$. The error probability under $\mathcal{H}_1$ is calculated as

$$\begin{aligned} & \Pr\left\{\min(\mathbf{s}) < -A | \mathcal{H}_1\right\} = \Pr\left\{\min(\mathbf{w}) < -A | \mathcal{H}_1\right\} \\ = \ & 1 - \left[1 - Q\left(\frac{A}{\sigma}\right)\right]^N \ , \end{aligned}$$

where $Q(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right) \mathrm{d}t$. By symmetry of the problem and the noise distribution, the error probabilities under each assumption are equal, so we have

$$P_{\mathrm{EF}} = 1 - \left[1 - Q\left(\frac{A}{\sigma}\right)\right]^N \ .$$

Suppose that the voiced portion of one syllable has M frames, since the estimation error probability of each frame is P_{EF}, then the error probability of final decision by majority vote is

$$P_{\mathrm{E}} = \sum_{m=\lceil M/2 \rceil}^{M} \binom{M}{m} P_{\mathrm{EF}}^m \left(1 - P_{\mathrm{EF}}\right)^{M-m} \ , \tag{3}$$

where $\lceil x \rceil$ rounds x to the nearest integers towards $+\infty$.

3 Detection Performance of the Optimal Detector

In this section, we consider a more systematic approach for detecting speech polarity using signal detection framework. Fig. 1 shows the signal generation model for voiced speech, If the information bit to hide is 0, the speech signal is modeled as the output of an all-pole model excited by the summation of $u_0[n]$ and $w[n]$, otherwise, the excitation signal is the summation of $u_1[n]$ and $w[n]$. Let $h[n]$ be the impulse response of the all-pole system, then the detection problem is to distinguish between the following two hypotheses

$$\begin{aligned} \mathcal{H}_0 : x[n] &= -\sum_{l=1}^{P_{\mathrm{AR}}} a_l x[n-l] - \sum_{k=0}^{K-1} A\delta[n - kP] + w[n] \\ &= u_0[n] * h[n] + w[n] * h[n] \\ &= \hat{u}_0[n] + \hat{w}[n] \\ \mathcal{H}_1 : x[n] &= -\sum_{l=1}^{P_{\mathrm{AR}}} a_l x[n-l] + \sum_{k=0}^{K-1} A\delta[n - kP] + w[n] \\ &= u_1[n] * h[n] + w[n] * h[n] \\ &= \hat{u}_1[n] + \hat{w}[n] \end{aligned}$$

The parameters $\{a_l\}_{l=1}^{P_{\mathrm{AR}}}, P_{\mathrm{AR}}, P, K$ are assumed known or can be estimated from the speech signals [5]. Since $\hat{w}[n]$ is the output of an all-pole model excited by IID WGN, so $\hat{w}[n]$ is

stationary WGN with mean zero and covariance matrix $\mathbf{C}$. To minimize the probability of decoding error, the optimal detection statistic for distinguishing between $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_0$ can be found to be [6][7]

$$T(\mathbf{x}) = \mathbf{x}^T \mathbf{C}^{-1} (\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_0) \quad , \tag{4}$$

which minimizes the probability of detection error. When N is large and the noise is wide sense stationary(WSS), the test statistic is approximated by

$$T(\mathbf{x}) \simeq \int_{-1/2}^{1/2} \frac{X(f) \left[\hat{U}_1(f) - \hat{U}_0(f)\right]^*}{P_{\hat{w}\hat{w}}(f)} \mathrm{d}f \quad ,$$

where $X(f), \hat{U}_1(f), \hat{U}_0(f)$ are the DTFT of $x[n], \hat{u}_1[n], \hat{u}_0[n]$ respectively. $P_{\hat{w}\hat{w}}(f)$ is the power spectrum density(PSD) of the noise $\hat{w}[n]$, i.e.,

$$P_{\hat{w}\hat{w}}(f) = \frac{\sigma^2}{\left|1 + \sum_{l=1}^{P_{\mathrm{AR}}} a_l \exp(-j2\pi f l)\right|^2} \quad .$$

The test statistic can be further simplified by invoking the Parseval's theorem, so we have

$$\begin{aligned} T(\mathbf{x}) &\simeq \int_{-1/2}^{1/2} \frac{X(f) \left[\hat{U}_1(f) - \hat{U}_0(f)\right]^*}{\sigma^2} \left|1 + \sum_{l=1}^{P_{\mathrm{AR}}} a_l \exp(-j2\pi f l)\right|^2 \mathrm{d}f \\ &= \frac{2}{\sigma^2} \sum_{n=P_{\mathrm{AR}}}^{N-1} \left(x_{\mathrm{w}}[n] \sum_{k=0}^{K-1} A\delta[n - kP] \right) \\ &= \frac{2}{\sigma^2} \sum_{k=0}^{K-1} A x_{\mathrm{w}}[kP] \quad , \end{aligned}$$

where x_{w} is the whitened $x[n]$ by inverse filtering [5].

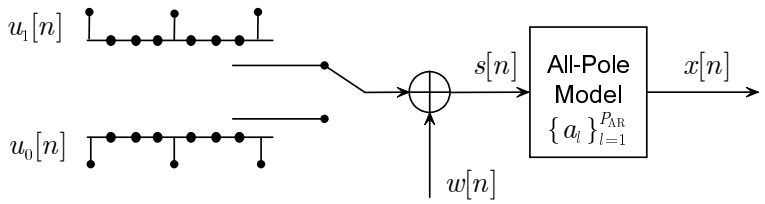


Figure 1: Polarity inversion watermarking model

The detection threshold γ is found to be

$$\gamma = \frac{1}{2} \left(\hat{\mathbf{u}}_1^T \mathbf{C}^{-1} \hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_0^T \mathbf{C}^{-1} \hat{\mathbf{u}}_0 \right) \quad .$$

It can be shown that

$$\hat{\mathbf{u}}_1^T \mathbf{C}^{-1} \hat{\mathbf{u}}_1 = \hat{\mathbf{u}}_0^T \mathbf{C}^{-1} \hat{\mathbf{u}}_0 = \frac{A^2 K}{\sigma^2} \quad ,$$

which implies that the detection threshold is $\gamma = 0$. In summary, the optimal detector decide $\mathcal{H}_1$ if

$$T(\mathbf{x}) = \frac{2}{\sigma^2} \sum_{k=0}^{K-1} A x_{\mathrm{w}}[kP] > 0 \quad . \tag{5}$$

The detector performance in terms of probability of error can be proved to be

$$P_{\mathrm{E}} = Q\left[\frac{1}{2} \sqrt{(\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_0)^T \mathbf{C}^{-1} (\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_0)} \right] \tag{6}$$

$$= Q\left(\sqrt{\frac{A^2 K}{\sigma^2}} \right) \quad . \tag{7}$$

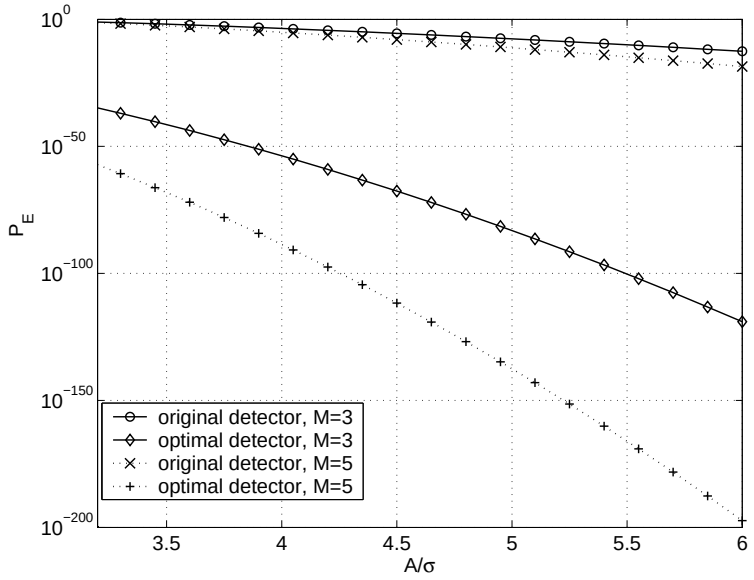


Figure 2: Comparison of theoretical P_{E}

Table 1: Parameters of Monte Carlo simulation to validate the assumptions

Parameter	*Value*
f_s	8kHz
P	60
N	300
A	from 3 to 5
σ	1

Performance Comparison of the Two Methods

In order to compare the theoretical results of (3) and (7), we set $K \times P = N \times M$. The pitch period P is chosen as 60. The results are shown in Fig. 2, the theoretical P_{E} of the optimal detector outperforms the original detector by tens of order of magnitude. When the number of frames in the voiced segment increases, more information-carrying samples are available, the P_{E} of both detectors decrease, this is shown in the figure for $M = 3$ and $M = 5$.

4 Monte Carlo Simulation

In this section, we perform the Monte Carlo Simulation to validate the theoretic analysis.

Validation of the Assumption in Section 2

In section 2, we have made the following assumptions to simplify the analysis: the sample values in the location of impulses are not affected by the AWGN, it is also assumed that the amplitude of the impulse is larger than the maximum absolute value of the AWGN. Here we will use Monte Carlo simulation to evaluate the effects of these assumptions on the final results. The parameters used in the simulation are shown in Table 1. The Monte Carlo simulation results are shown in Fig. 3 for $A = 3, 4, 5$, the comparison between analytical results and simulation results reveals that the analytical P_{EF} is about ten times larger than the simulation results, however, the analytical results with the assumptions provide an upper bound for the true situations, the assumptions tends to be more realistic for larger A/σ. Furthermore, in the above comparison

between the performance of the original and the optimal detector, we see that the P_{E} of the optimal detector is tens of order of magnitude smaller than the non-optimal case, so the analytical results are valid for comparison purpose.

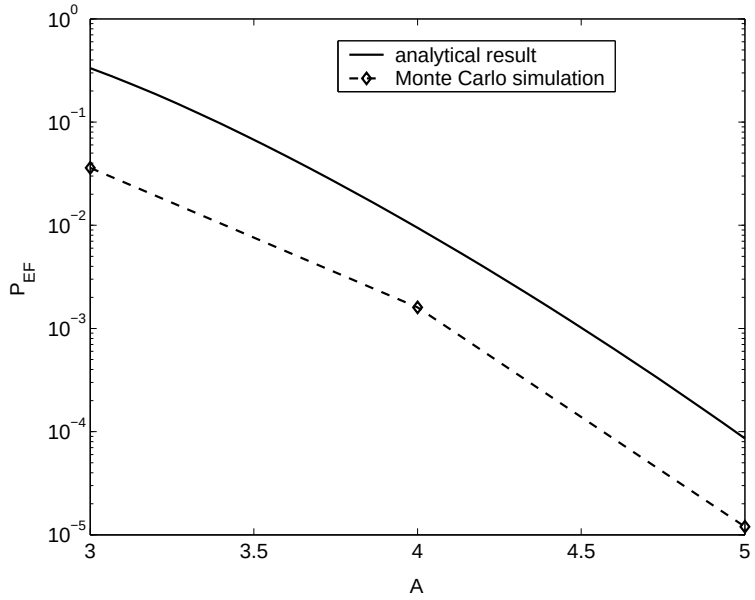


Figure 3: Comparison between analytical results and Monte Carlo simulation results of P_{EF}

Validation of the Impulse Train + WGN Residual Model

In order to validate the "Impulse Train + WGN" model for the residual signal, we perform the following statistical experiments on the real speech signals: first, we compute the residual signal of the voiced frames, then, the impulses are treated as outliers [8], they were eliminated from the residual signal, the histogram of the the remaining residual signal are calculated and compared to the empirical Gaussian PDF with parameters estimated from the data by maximum likelihood (ML) estimation. The outlier detection scheme was based on method proposed in [9], which calculate $M_i = 2\,(x_i - x_{50\%})/(x_{84\%} - x_{16\%})$ for each data sample x_i, where $x_{50\%}$ is the median of data sequence $\{x_i\}_{i=1}^{N}$, $x_{84\%}$ and $x_{16\%}$ are the 84% and 16% percentile respectively, x_i is classified as outlier when $|M_i| > 3$. Fig. 4 shows the experimental results when applying the outlier detection and elimination algorithm on residual signals, due to the non-stationary nature of the speech signal, the outlier detection and elimination algorithm were performed frame by frame . After the removal of outliers, the histogram of the residual signal is fitted by Gaussian distribution. The result is shown in Fig. 5, which shows good matching between the histogram and the Gaussian distribution PDF. The sample mean and standard deviation is estimated to be -0.0016 and 0.0277 respectively. The amplitude of the impulses are found to be between 0.1 to 0.15, the quantity A/σ is between 3 and 6. The pitch period can be found manually to be 40 and 41. Using the parameters estimated above, the synthesized speech residual signal is shown in Fig. 6. We will use this model to validate the theoretical P_{E} of optimal detector by Monte Carlo simulation.

Validation of P_{E} of the Optimal Detector

To validate the results in (7), we perform the Monte Carlo simulation to estimate $\hat{P}_{\mathrm{E}}$. The detector (5) is applied on data sequences generated by the "impulse train + WGN " model, the number of detection errors is counted, the estimated $\hat{P}_{\mathrm{E}}$ can be calculated as

$$\hat{P}_{\mathrm{E}} = \frac{\#\text{ of detection errors}}{\#\text{ of Monte Carlo simulations}} \quad .$$

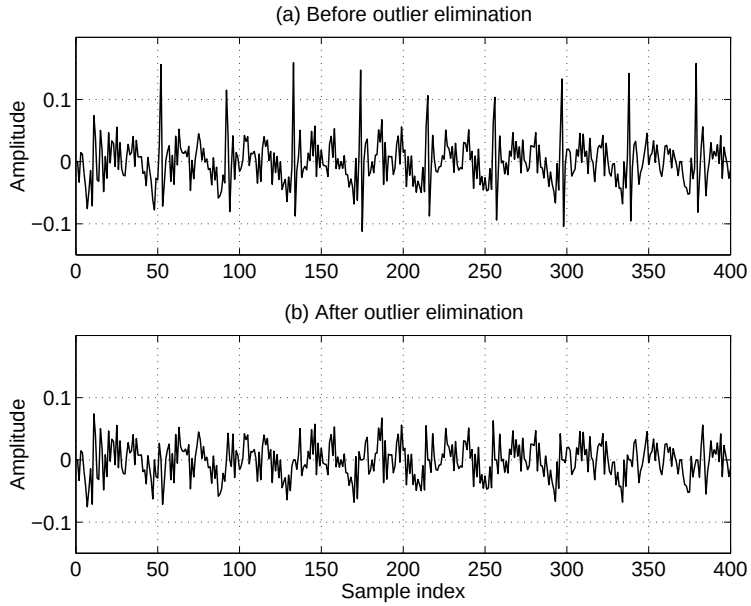


Figure 4: Speech residual signal before and after outlier elimination

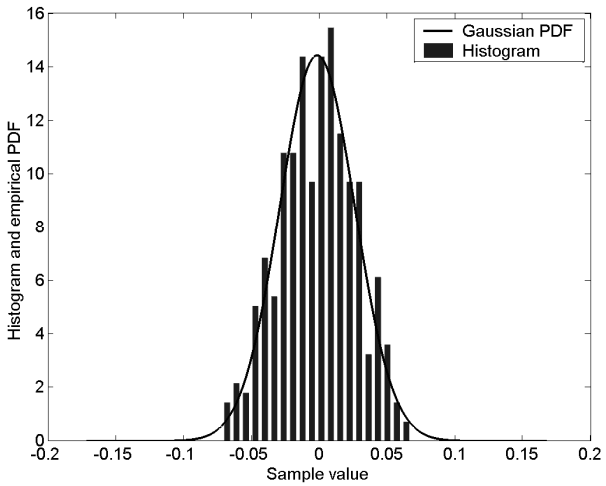


Figure 5: Histogram and Gaussian PDF with parameters estimated from speech residual after elimination of outliers

From the theoretical analysis, we see that the detection error is rather rare event, for example, when $A/\sigma = 5, M = 3$, P_{E} is approximately 10^{-80}, to simulate the rare event, we use importance sampling to reduce the variance of $\hat{P}_{\mathrm{E}}$ [10, 11]. Due to the symmetry of the detection problem and the noise distribution, we only consider the detection error under $\mathcal{H}_1$, which is

$$\begin{aligned} P_{\mathrm{E}} = P_{\mathrm{E}|\mathcal{H}_1} &= \Pr\left\{\frac{1}{K}\sum_{k=0}^{K-1} w[kP] < -A; \mathcal{H}_1\right\} \\ &= \Pr\left\{\frac{1}{K}\sum_{k=0}^{K-1} w[kP] > A; \mathcal{H}_1\right\} = \mathcal{E}_f\left(I_{\left\{K^{-1}\sum_{k=0}^{K-1} w[kP]>A\right\}}\right) \\ &= \mathcal{E}_g\left(I_{\{\bar{w}>A\}}\right) \\ &= \mathcal{E}_{\mathrm{A}}\left\{I_{\{\bar{w}>A\}} \exp\left[\left(-2A\bar{w} + A^2\right) / \left(2\sigma^2/K\right)\right]\right\} \end{aligned}$$

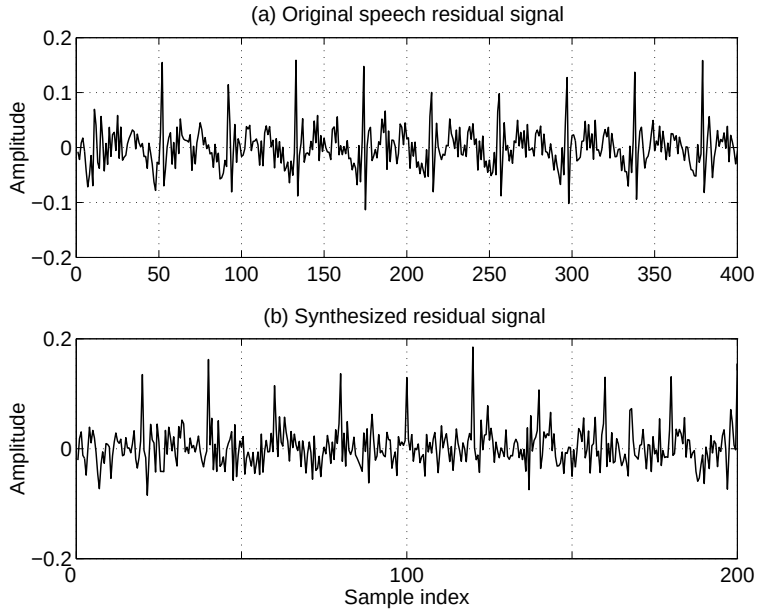


Figure 6: Synthesized speech residual using the "impulse train + WGN " model

Table 2: Theoretical P_{E} and estimated $\hat{P}_{\mathrm{E}}$ by importance sampling

$\frac{A}{\sigma}$	P_{E}	$\hat{P}_{\mathrm{E}}$	95% confidence interval
3	1.6508×10^{-31}	1.8005×10^{-31}	$[1.4065\ 2.1946]\times 10^{-31}$
4	1.9664×10^{-54}	2.0208×10^{-54}	$[1.4851\ 2.5566]\times 10^{-54}$
5	7.6301×10^{-84}	7.0208×10^{-84}	$[4.8998\ 9.1419]\times 10^{-84}$
6	9.4276×10^{-120}	1.1016×10^{-119}	$[0.7480\ 1.4551]\times 10^{-119}$

where $I_{\mathrm{D}}(x)$ is the indicator function, which is one if $x \in D$, and zero otherwise, $\mathcal{E}_f$ is the statistical expectation w.r.t. the distribution $f = \mathcal{N}(0, \sigma^2)$, $\mathcal{E}_g$ is the statistical expectation w.r.t. the distribution $g = \mathcal{N}(0, \sigma^2/K)$, $\mathcal{E}_{\mathrm{A}}$ is the statistical expectation w.r.t. the distribution $\mathcal{N}(A, \sigma^2/K)$, We use Monte Carlo simulation to estimate $\mathcal{E}_{\mathrm{A}}\left\{I_{\{\bar{w}>A\}} \exp\left[\left(-2A\bar{w} + A^2\right) / \left(2\sigma^2/K\right)\right]\right\}$, the results are shown in Table 2, the number of experiments in Monte Carlo simulation is 1000, the 95% confidence intervals are also included in the table. The Monte Carlo simulation results fit well with the theoretical result (7).

5 Conclusion and Future Work

For detection of speech polarity, the speech residual signal can be modeled as impulse train plus WGN, the optimal detector outperformes the original polarity detection algorithm by tens of order of magnitude in term of detection error. This result is validated by Monte Carlo simulation. It should be noted that in the above analysis, we have assumed that the parameters of the impulse train P, A and the AR model parameters are all assumed known, in practice, these parameters must be estimated from the speech signal, the estimation error will degrade the detector performance. The performance loss using estimated parameters is under investigation and will be reported in a future paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant 60272074 and the Spaceflight Innovation Foundation of China under grant [2002]210-6.

References

[1] S. Sakaguchi, T. Arai and Y. Murahara. The Effect of Polarity Inversion of Speech on Human Perception and Data Hiding as an Application. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, June 5-9, (2000) **2** 917-920

[2] Yardyimci, Y., Cetin, A. E., and Ansari, R.: Data Hiding in Speech Using Phase Coding. Eurospeech 97 **3** (1997) 1679-1682

[3] Ciloglu, T. and Karaaslan, S. Utku: An Improved All-Pass Watermarking Scheme for Speech and Audio. International Conference on Multimedia and Expo. July 30- Aug. 2 (2000) **2** 1017-1020

[4] Cheng, Q., Sorensen, J., "Spread Spectrum Signaling For Speech Watermarking", IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA,pp. 1337-1340.

[5] J. R. Deller, Jr. , J. G. Proakis, J. H. L. Hansen, Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, USA, 1993

[6] S. M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory. Prentice-Hall, New Jersey, 1998

[7] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification (2nd Edition), Wiley-Interscience, USA, 2001

[8] V. Barnett, T. Lewis. Outliers in Statistical Data, John Wiley and Sons, New York, 1994.

[9] H. J. Kim, T. Kim, In-Kwon Yeo. A Robust Audio Watermarking Scheme. IEEE ISCAS, Canada, May 2004.

[10] Peter J. Smith, M. Shafi, Hongsheng Gao. Quick Simulation: A Review of Importance Sampling Techniques in Communications Systems. IEEE Journal on Selected Areas in Communications, May 1997, Vol.15, No. 4, 597-613

[11] R. L. Mitchell. Importance Sampling Applied to Simulation of False Alarm Statistics. IEEE Trans. on Aerosp. Elect. Syst., Jan, 1981, Vol. AES-17, 15-24

A Probe into Ambiguities of Determinative-Measure Compounds

Shih-Min Li, Su-Chu Lin, Keh-Jiann Chen
CKIP, Institute of Information Science, Academia Sinica, Taipei
{shihmin, jess}@hp.iis.sinica.edu.tw; kchen@iis.sinica.edu.tw

Abstract

This paper aims to further probe into the problems of ambiguities in automatic identification of determinative-measure compounds (DMs) in Chinese. It is known that Chinese DMs are identifiable by regular expression rules. However, rule matching only partially solve structural and lexical ambiguities. In this paper, a deep analyses based on corpus data was studied. With the subtle analyses of error identification and disambiguation of DM compounds, we classified three types of ambiguities, i.e. structural, sense, and functional ambiguities. We also proposed resolution principles to eliminate the problems and thus to improve word segmentation and POS (Part-Of-Speech) tagging.

1 Introduction

To a speaker of English, one of the most striking features of the Mandarin noun phrase is the classifier. A classifier is a word that must occur with a number and/or a demonstrative, or certain quantifiers before the noun (Li and Thompson 1981: 104). Furthermore, Li and Thompson (1981) assert any measure word can be a classifier, so the combination of demonstrative and/or number or quantifier plus a classifier or a measure is defined as a classifier phrase or a measure phrase. For example, *san ge* in *san ge ren* (三個人), *zhe zhan* in *zhe zhan deng* (這盞燈), *ji jian* in *ji jian yifu* (幾件衣服), *liu li* in *liu li lu* (六里路), *na jin* in *na jin yangrou* (那斤羊肉) and *ji gang* in *ji gang cu* (幾缸醋) are classifier phrases/measure phrases, which are called as D-M compounds in Chao 1968. A determinative (D) and a measure normally make a compound with unlimited versatility and form a transient word of no lexical import (Chao 1968: 389). Although the demonstratives, numerals and measures may be listed exhaustedly, their combination is inexhaustible. Certain constructions of DMs are ambiguous, for example:

(1) 廣告裡全篇多具聳人聽聞的口號
guanggao li quan pian duo ju hairentingwen de kouhao
A whole advertisement mostly has shocking slogans.

(2) 取此名
qu ci ming
choose this one (person) / name this name

(3) 二十五年的審核、排隊、等待
ershiwu nian de shenhe paidui dengdai
examining, lining up and waiting in the year of twenty five / for twenty five years

The morpheme *ju* in Academia Sinica Balanced Corpus (Sinica Corpus) has two parts of speech, VJ and Nf.[1] Thus the phrase *duo ju* in sentence (1) can be either a verb phrase with the meaning of 'mostly have' or a DM. When *ju* functions as a measure, it always modifies corpses, not slogans. Since *ju* never co-occurs with slogans, the phrase *duo ju* here is certainly a verb phrase and then the lexical ambiguity of *ju* is reduced. In example (2), *ming* can function as a measure as well as a noun so that this verb phrase has two meanings. In example (3), *ershiwu nian* can be a time point specifying the event-time of the verb, or denotes the period of time delimitating the time length of the event. The former temporal adverb is tagged as Nd; the latter is separated into two morphemes and individually tagged as Neu and Nf.[2] Examples (1) to (3) show the different degree of ambiguity.

[1] The symbol in Sinica Corpus, "VJ" stands for Stative Transitive Verb and "Nf" for Measure. The detailed parts of speech can be referred to Sinica Corpus website.

[2] The symbol "Nd" stands for Time Noun and "Neu" for Numeral Determinatives.

Due to the infinite of the number of possible DMs, Mo et al. (1991) propose to identify DMs by regular expression before parsing as part of their morphological module in NLP. The adoption of DMs rules really improves the accuracy of recognition, but we still have some difficulties in segmentation as the preceding examples. In this paper, the discussion and classification of ambiguities of DMs are the focus. In addition to the typical DM structure with the combination of one or more determinatives with a measure, the reduplicative DMs and the ellipsis of determinatives will be also included under investigation. After the analyses of multiple ambiguities, we try to find out resolution principles to reduce these ambiguities.

2 Literature Review

To deal with DMs, first we have to give a proper definition to DMs; thus we can delimit the scope of our discussion. There are numerous discussions on determinatives as well as measures, especially on the types of measures.[3] The classification of measures is not the issue in this paper. To avoiding confusion between classifiers and measures, we have to pay attention to the distinction between them. Tai (1994: 480) asserts that in the literature on general grammar as well as Chinese grammar, classifiers and measures words are often treated together under one single framework of analysis. Chao (1968) treats classifiers as one kind of measures. In his definition, a measure is a bound morpheme which forms a D-M compound with one of the determinative enumerated above (Chao 1968: 584). Classifiers are defined as 'individual measures', which is one of the nine kinds of measures. As we mentioned in the section of introduction, Chao considers that determinatives are listable and measures are largely listable so D and M can be defined by enumeration, and that D-M compounds have unlimited versatility. While Li and Thompson (1981) blend classifier with measure. They conclude not only does a measure word generally not take a classifier, but any measure word can be a classifier. In Tai's opinion (1944: 481), in order to better understand the nature of categorization in a classifier system, it is not only desirable but also necessary to differentiate classifiers from measure words. In this paper, since we adopt the CKIP DM rules and symbols of POS, we inherit the term determinative-measure compounds (DMs), which have been defined as the composition of one or more determinatives together with an optional measure (Mo et al. 1991: 111).

As for the linguistic ambiguity, Crystal (1991: 17) specifies the general sense of ambiguity is a word or sentence which expresses more than one meaning. The most widely discussed type of ambiguity in recent year is grammatical (or structural) ambiguity. In the structure *new houses and shops*, it could be analysed either as *new* [*houses and shops*] (i.e. both are new) or [*new houses*] *and shops* (i.e. only the houses are new). Furthermore, according to Crystal's assertion, ambiguity which does not arise from the grammatical analysis of a sentence, but is due solely to the alternative meanings of an individual lexical item, is referred to as lexical ambiguity, e.g. *I found the table fascinating* (= 'object of furniture' or 'table of figures'). The definition of structural and lexical ambiguities can be referred to Prins (2005). Prins (2005: 1) mentions if we restrict our attention to the syntax in texts, then we may focus on ambiguity in two forms. The first is lexical ambiguity, the second is structural ambiguity. Lexical ambiguity arises when one word can have several meanings. Structural ambiguity arises when parts of a sentence can be syntactically combined in more than one way. Prins believes human can resolve most ambiguity, of both types, without even being consciously aware of alternatives. The ambiguity remains of which we are aware, knowledge about the world is used in combination with what is known about the linguistic context of the ambiguity to arrive at the most likely analysis. However, in our following analysis, we find out that only structural ambiguity and lexical ambiguity are not enough to obtain more detailed discussion on ambiguities of DMs. Structural ambiguity is caused by different segmentation of words. With the same segmentation, that string of words may still be ambiguous because the same string may have more than one meaning or may have different functions. Therefore, lexical ambiguity can be further divided into two types; the former is sense ambiguity and the latter is functional ambiguity.

In this paper, we examine and analyze Mandarin Chinese DMs in Sinica Corpus. In the subsections in section 3, we make a study of the structures and ambiguities of DMs, and then try to analyze and disambiguate these DMs.

---

[3] Chao (1968) and Li and Thompson (1981) detect measures and classifiers. He (2000) traces the diachronic names of measures and mentions related literature on measures. The dictionary of measures pressed by Mandarin Daily News Association and CKIP (1997) lists all the possible measures in Mandarin Chinese.

3 Structures and Ambiguities of DMs

The probe into DMs in this paper focuses on the typical DM, one or more determinatives following a measure, including the variant forms of DM, such as the ellipsis of the determinative and the insertion of an adjective into DM[4]. Besides, we will sketchily study the various reduplicative forms of DM, like the reduplication of only measures 'MM' and the numeral *yi* preceding the reduplicative measures '*yi*MM' (一MM). The following structures show the variants of DMs.

(4) 只有三十八位參選
zhiyou sanshiba wei canxuan
Only thirty eight persons take part in the election.

(5) 我們必須說句良心話
women bixu shuo ju liangxinhua
We have to give an absolutely honest speech.

(6) 爲一個問題作一大堆研究
wei yi ge wenti zuo yi da dui yanjiu
do a lot of research for the reason of an question

(7) 種種問題
zhong zhong wenti
all sorts of questions

(8) 一張張海報
yi zhang zhang haibao
each poster

(9) 如此一大口一大口地吃
ruci yi da kou yi da kou di chi
make such a mouthful of eating

The DM *sanshiba wei* in example (4) is the typical DM. In example (5), determinatives preceding the measure *ju* are omitted. Therefore, the demonstrative, specifying, numeral or quantitative functions of determinatives in (5) will flow away. Example (6) has the DM structure *yi da dui*, composed of DM *yi dui* and an insertion of an adjective *da*. The reduplication in (7), (8) and (9) are also our topical subjects.

As Chao (1968: 552) points out, a D-M compound is a substantive and can enter into constructions as subject, object, or attribute. In sentence (4) above, *sanshiba wei* is the subject of the verb *canxuan* and has the function as a pronoun. DMs in example (5) to (8) modify the amount of nouns, so they all have the function as an attribute. The reduplication of DMs in (9) describes the manner of eating and functions as an adverb.

3.1 Structural Ambiguities of DMs

When we identify DMs, structural ambiguity of them occurs as the following examples.

(10) 一服藥就見效
yi fu yao jiu jianxiao
Every time when he takes medicine, the illness is completely cured.
One dose is effective.

(11) 一串串珠飾品
yi chuan chuanzhu shipin
one string of beads

(12) 北市文昌國小五年一班
beishi wenchang guoxiao wunianyiban
the Fifth Grade Class One in Taipei Wen Chang Elementary school

(13) NC1 -> {NE1,NE2} {年} {NE1,NE2,ON} {班} ;

Example (10) is grammatically ambiguous and has two meanings. If *fu* functions as a verb, the meaning of (10) is the former one. If *fu* functions as a measure, the meaning of (10) is the latter. Because the combination of determinatives and measures are countless, the DM *yifu* won't be listed in the CKIP dictionary. Different word segmentation will bring structural ambiguity forth. However, Mo et al. (1991) list a resolution principle to reduce structural ambiguity. The principle asserts if

[4] The insertion of an adjective into DM has the structure of '*yi*AM'. The symbol 'A' stands for adjectives.

ambiguous word breaks occur between the words in the lexicon and the DMs, the words in the lexicon should have higher priority to get the shared characters. Therefore, the former meaning in (10) has higher priority to the latter. Similarly, the DM in (11) may be segmented as *yi chuan* or *yi chuan chua*n if the measure *chuan* is followed by the same morpheme as it. The statistics of collocation of the context help to reduce the structural ambiguity existed in example (11). Example (12) also has structural ambiguity. If *nian* and *ban* are treated as measures, *wu nian yi ban* are segmented into four. By application of the resolution principle and DM rule (13), classes in elementary schools are viewed as a unit; therefore, *wunianyiban* is restricted be a unit whose POS is Nc[5].

Another structural ambiguity exists in the ellipsis of the determinatives. The phrase *you ge ren* in sentence (14) denotes to somebody. In (14), *ge* functions as a measure without any determinatives preceding it. The same phrase in (15), however, never refers to persons. The morpheme *geren* is viewed as a unit and tagged as Nh[6] with the meaning 'individual' in (15). The morpheme *you* in (15) is a verb and means 'have', whose function is not the same as the specifying determinative *you* in (14). The resolution principle and the collation help to resolve this kind of structural ambiguities.

(14) 有一次有個人瞥見我在街上拍照
you yi ci you ge ren piejian wo zai jie shang paizhao
Once somebody got a glimpse of my taking pictures in a street

(15) 有個人空間
you geren kongjian
have individual space

When dealing with addresses, we also encounter structural ambiguities, especially indicating the floor, number, alley, lane, section and neighbourhood. The following instances show the same forms with different segmentation between DMs and addresses.

(16) 屬於１１７號公路的一段
shuyu yiyiqi hao gonglu de yi duan
belong to a part of the 117th road

(17) 羅斯福路一段八號一樓
luosifulu yiduan bahao yilou
F 1, No. 8, Roosevelt Rd., Sec. 1

(18) 行經屏市長安里竹圍巷一之一０二號時
xing jing pingshi changanli zhuweixiang yizhiyilingerhao shi
when going through No. 1-102, Zhuwei Lane, Changan Village, Pingtung City

(19) NC2 -> {NE1,NE2} {鄰,巷,弄,樓} ;

(20) NC4 -> {NE1,NE2} {之,－} {NE1,NE2} {號 } ;

(21) 日前遷至台北市信義路三段七號三樓之一
riqian qian zhi taibeishi xinyilu sanduan qihao sanlou zhiyi
a few days ago, move to 3F-1, No. 7, XinYi Rd., Sec. 3, Taipei

(22) 物理研究所則列名於３５個研究機構之一
wuli yanjiusuo ze liemingyu sanshiwu ge yanjiu jigou zhi yi
The Institute of Physics is placed among 35 research centers.

(23) 等於是一日的三分之一
dengyu shi yi ri de ssanfenzhiyi
is equal to one day of thirds

(24) NE6 -> ({NE1,NE2} {又}) {NE1,NE2} {分之} {NE1,NE2,NE5} ;

In instance (16), *hao* and *duan* are both measures specifying the fixed amount or quantity of the road, so the numerals 117 and *yi* are separated from the measures. According to CKIP Technical Report 96-01 (1996: 50), the determinative measure structures expressing time and location will be combined together as a unit. The reason why the locative DMs are conjoint is because the first joint principle of segmentation stipulates that when the meaning of a string of words is not obtained from the composition of these components, this string should be segmented as a unit. Consequently, *yiduan*, *bahao* and *yilou* in (17) and *yizhiyilingerhao* in (18) are segmented as Nc in Sinica Corpus. The DM rules (19) and (20) help us tag locative DMs as Nc, which is different from the DM structures in (16). During the process of locative DMs, another concerned structure listed in (21), (22) and (23) derives. All the three instances have the same surface structure *zhi yi*, but the functions and segmentation are

[5] The symbol "Nc" stands for Place Noun in Sinica Corpus.

[6] The symbol "Nh" stands for Pronoun.

different. The morpheme *zhi yi* in (21) is tagged as a unit whose POS is Nc, in (22) is segmented into two units whose POS is individually DE[7] and Neu, while in (23) is the part of the whole quantitative determinative *sanfenzhiyi* tagged as Neqa[8]. To reduce these structural ambiguities, the DM rule (24) is necessary.

According to the above discussions, to resolve structural ambiguities, we conclude the following resolution principles which were implemented at the word segmentation system by Ma and Chen (2003).

a) D-M compounds are expressed and matched by regular expressions.
b) Lexical words have higher precedence than D-M compounds (cf. 11).
c) Long D-M has higher precedence than short D-M (cf. 12, 16, 17, 18, 21, 22, 23).
d) Covering ambiguities are resolved by collocation context (cf. 10, 14, 15).

The structural ambiguity is caused by different possible segmentation. Although example (10) has structural ambiguities, after the application of the resolution principles, the ambiguous segmentation is resolved and the correct segmentation has higher priority.

3.2 Sense Ambiguities of DMs

Senses and semantic functions of DMs are related to the types of measures. Li and Thompson (1981: 105) claim that Mandarin has several dozen classifiers, most of which can be found in Chao (1968: sec. 7.9). Chao (1968: 584-620) divides measures into nine kinds: (1) classifiers, or individual measures (Mc), (2) classifiers specially associated with V-O constructions (Mc'), (3) group measures (Mg), (4) partitive measures (Mp), (5) container measures (Mo), (6) temporary measures (Mt), (7) standard measure (Mm), (8) quasi-measures (Mq), and (9) measures for verbs (Mv). Briefly speaking, the function of DMs is to modify the amount and quantity of abstract and concrete things, to count the frequencies of events and actions, and to indicate the event time. To testify the functions of DMs, we analyze the semantic roles of DMs in Sinica Treebank. First we use "DM" as the keyword to retrieve the Sinica Treebank data and then calculate the frequencies of the semantic roles of these DMs. The most highly frequent semantic role is quantifier, whose frequency is 6434. Quantifier is mainly used to account for the amount of things. The statistics in Sinica Treebank show the frequencies of the semantic roles of DMs, and then we get the hierarchical order of semantics roles of DMs from high to low: quantifier > Head > DUMMY > range > time > property > frequency > goal > duration > theme > DUMMY1 > DUMMY2 > agent > quantity > topic > manner > apposition > location > instrument > experiencer. As expected, quantifier is the most common semantic role played by DMs. The semantic role "property" denoting attributes and "range" referring to amount both have high frequencies. "Quantity" is also used to modify the extent of actions. The measures such as *nian* (年), *ci* (次) and *tian* (天) are related to temporal concepts, whose semantic roles may be "time", "frequency" or "duration". The semantic roles "range", "time", "frequency" and "duration" are usually concerned with the measures classified into Mm and Mq in Chao's classification of measures. The DMs always function as pronoun when their semantic roles are "goal", "theme", "agent", "topic", "apposition", "location" and "experiencer". If DMs play the semantic role "manner" and "instrument", the measures are usually classified into Chao's Mv. The sense of certain types of DMs can be identified by types of measures; however, as usual, some DMs have ambiguous senses. Their ambiguity resolution is almost equivalent to word sense disambiguation. Therefore context sensitive rules and collocation bi-grams are information for resolving lexical ambiguities. Methods for word sense disambiguation are also applicable for DMs. Here we first discuss ambiguity about temporal adverbs to illustrate the sense ambiguity and possible resolution methods.

To represent the percentage, using Chinese characters like (23) is one form, and adopting mathematical symbols like (25) is another one. The form of mathematical symbols is sense ambiguous. It can refer to either the percentage like (25) or time point like (26). Without context, the symbol "１０／２１" can be a fractional number and read as "ten over twenty one" with the POS "Neqa" and as " 10月21日 *shiyue ershiyiri*" with the POS "Nd" whose semantic role is time. To reduce this kind of sense ambiguities, we have the DM rules (28) and (29). The form in rule (28) is tagged as Neqa (numbers) while in (29) as Nd (time point). The mathematical symbol indicating a specific time usually denotes the year together so "2005／06／30" in (27) is tagged as Nd. Although we have rules (28) and (29) to help differentiate the meaning of percentage from that of time, we still have to have context to make (25) and (26) distinguishable.

---

[7] The symbol of "DE" is the POS of 的, 之, 得 and 地.

[8] The symbol "Neqa" stands for Quantitative Determinatives.

(25) ４０分的佔了２／３
sishi fen de zhan le sanfenzhier
Those of forty points occupy two-thirds.
(26) １０／２１召開全院網路工作小組第三次會議
shiyuershiyiri zhaokai quan yuan wanglu gongzuo xiaozu disan ci huiyi
convene the third conference of the network group on Oct. 21
(27) 2005／06／30更新
2005／06／30 gengxin
update on June 30, 2005
(28) NE5a -> {NE2} {—,／} {NE2} ;
(29) NE5b -> {NE2} {—,／} {NE2} {—,／} {NE2} ;

Chao (1968) gives an example about time words. The form *Guangxu sanshisinian* (光緒三十四年) can be either the phrase 'the thirty-fourth year of Guangxu (i.e., 1908)' or the sentence 'Guangxu's reign was thirty-four years (long).'. Chao believes that in most cases, the context will resolve the ambiguity. Below are examples with similar ambiguity as Chao mentions.
(30) 經過卡斯楚三十年的統治之後
jingguo sanshi nian geming de xili
after Castro's thirty-year governance
(31) 三十年秋，緝私總隊復正名爲稅警總團
sanhinian qiu qisi zongdui fu zhongmingwei shuijingzongtuan
In the autumn in the year of thirty, the anti-smuggling team is rectified to the tax policemen team

Examples (30) and (31) have the same temporal phrase *sanshi nian*, but their semantic functions and roles are different. The temporal phrase in (30) expresses time length and is segmented into two units as Neu and Nf. The semantic role of it is duration. However, *sanshinian* in (31) indicates time point and is tagged as Nd, whose semantic role is time. Although either a Chinese reign title or *Mingguo* (民國) preceding *sanshinian* is omitted, we still know *sanshinian* is a specific time, not a period, from the context. When the measures *nian* and *ri* are preceded by numerals, the temporal phrases always have sense ambiguity. Basically, we segment numeral and measure into two and then postprocess them by applying two tricks following. If DMs denote time point, they are usually preceded by key words of *Mingguo*, the Christian era like *Gongyuan* (公元) and *Xiyuan* (西元), or a Chinese reign title *Guanxu*, *Qianlong* (乾隆), *Tianbao* (天寶), *Jiajing* (嘉靖) and so on. Another trick helps to recognize DMs is its neighbouring temporal nouns. The temporal DMs usually co-occur with one or two temporal phrases such as *erlinglingwunian liuyue* (2005年6月), *liuyue sanshiri* (6月30日), *erlinglingwunian liuyue sanshiri* (2005年6月30日), etc.

Two tricks above can reduce some ambiguities of temporal phrases. But some ambiguities listed in the following examples cannot be reduced.
(32) 2005宜蘭童玩節
erlinglingwu yilan tongwan jie
I-Lan International Children's Folklore & Folkgame Festival in 2005
(33) 一要有錢
yi yao you qian
first have to have money
(34) 九二一地震
jiueryi dizhen
the earthquake 921
(35) 鼓勵534人完成319鄉之旅
guli wubaisanshisi ren wancheng sanbaiyishijiu xiang zhi lu
encourage 534 persons to accomplish the travel around 319 villages

The composition of numeral mostly functions as numeral determinatives while sometimes doesn't. The numeral 2005 in (32) refer to the year of 2005 AD; however, the numeral *yi* in (33) is a correlative conjunction. Furthermore, the similar numeral structures in (34) and (35) have different semantic meanings.

In conclusion, sense ambiguity resolution is almost equivalent to word sense disambiguation. Therefore context sensitive rules and collocation bi-grams are information for resolving lexical ambiguities. Methods for word sense disambiguation are also applicable here.

3.3 Functional Ambiguities of DMs

The semantic function of the temporal DM in (36) may be duration while (37) time. Same words and same word senses may play different semantic roles. The reason of making different assignment of semantic roles may be concerned with logical interpretation of sense collocations according to common sense and the real world knowledge.

(36) 18年的苦守
shiba nian de kushou
wait bitter for eighteen years
(37) 89年的反抗
bajiu nian de fankang
the revolt in the year of 89

When detecting DMs rules and Sinica Corpus data, we find out some interesting examples. The verb phrases (38) and (39) have the same morphemes except for the position of the temporal DM *sanshiba nian*. The semantic role of the DM in (38) is duration while in (39) is time. It seems the different position of temporal DMs will affect the meanings of sentences. Thus, we briefly calculate the data in Sinica Treebank. The totality of the semantic role time of NPs and PPs following the verb is close to that of the semantic role duration. But the totality of the semantic role time of NPs and PPs preceding the verb is much more than that of duration. It seems temporal DMs preceding verbs mostly function as time. Another different assignment of semantic role to the similar structure is shown by (40) and (41). The former DM is assigned the semantic role duration while the latter time. This kind of ambiguity has relation to situation types. The situation type of *fuxing* (服刑) is activity while that of *panxing* (判刑) is achievement. The feature [±Durative][9] of the events causes differences. The phrase *yixia* in (42) means 'for a while' and is assigned the role of duration, while in (43) is assigned frequency and composed of a numeral and a measure. The phrase in (44) is ambiguous with two meanings. One means 'bite him for a while' while another means 'bite him once'. Equal to the cause of differences between (40) and (41), the ambiguity in (44) is also due to the situation types.

(38) 親政三十八年
qinzheng sanshiba nian
take over reins of government upon coming of age for 38 years
(39) 三十八年親政
sanshibanian qinzheng
take over reins of government upon coming of age in the year of 38
(40) 34年的服刑
sanshisi nian de fuxing
serve a sentence for 34 years
(41) 34年的判刑
sanshisi nian de panxing
sentence a person in the year of 34
(42) 等我一下
deng wo yixia
Wait for me for a while.
(43) 敲他一下
qiao ta yi xia
strike him once
(44) 咬他一下
yao ta yixia
bite him for a while / bite him once

Semantic role assignment is not an easy task, since it requires not only linguistic knowledge but also world knowledge. In Yu and Chen (2004), they identify parameters of determining semantic roles and

[9] The more detailed discussion about situation types can be referred to Smith 1991.

proposed an instance-based approach to resolve ambiguities. They adopt dependency decision making and example-based approaches. Semantic roles are determined by four parameters, including syntactic and semantic categories of the target word, case markers, phrasal head, and sub-categorization frame and its syntactic patterns. The refinements of features extraction, canonical representation for certain classes of words and dependency decisions improve role assignment. To assign semantic roles of DMs, the above parameters are further refined as the features of relative positions and situation types.

The examples above show that ambiguity is unavoidable when we deal with DMs. In addition to the typical DMs, some related structures like reduplicative DMs, numerals, the ellipsis of measures, etc. are also the topics for discussion. The composition of determinatives and measures brings about ambiguity. Some ambiguities are caused by different segmentations of words, some are due to the multiple meanings of words, and others are concerned with different functions. Therefore, ambiguities of DMs are classified into structural ambiguity, sense ambiguity and functional ambiguity. Here take *yi dian* (一點)for instance.

(45) 有一點要特別注意
zhe yi dian zhuyi shixiang hen zhongyao
This point for attention is very important.

(46) 一點心意你要收下
yidian xinyi ni yao shouxia
You must receive my little thanks.

(47) 一點集合
yidian jihe
assemble at one o'clock

(48) 漂亮一點
piaoliang yidian
a little bit beautiful

(49) 快一點
kuai yidian
nearly one o'clock
more quickly

(50) 慢一點
man yidian
more slowly

The phrase *yidian* has different structures in sentences (45) to (48). In (45), *yi dian* functions as a pronoun and is segmented into two units. In (46) to (49), *yidian* is viewed as one unit. It functions as a quantitative determinative modifying *xinyi* in (46), a time noun in (47), and a post-verb adverb of degree in (48). While in (49), *yidian* is lexically ambiguous depending upon context. However, when (49) has the former meaning, (49) and (50) are functionally ambiguous. The ambiguity of DMs is complex and it is possible that one DM compound has more than one classification of ambiguities

No matter the ambiguity is the structural one, sense one or functional one, the prescription of resolution principles and DM rules are helpful in disambiguating DMs. Besides, the neighbouring morphemes and context are one another tricks in reducing ambiguity. In some cases, ambiguity is not easily resolved. Furthermore semantic role ambiguities are concerned with common sense and the resolution features also include position of temporal DMs and the situation types. Such ambiguities have to be reduced by the application of parameters of context vector models.

4 Conclusion

In section 3, we discuss the ambiguity of DMs, which is mainly divided into structural ambiguity, sense ambiguity and functional ambiguity. These ambiguities can be reduced by applying resolution principles, DM rules, context sensitive rules, collocation bi-grams and parameters of context vector models. Because language reflects the human view of the world, different personal world knowledge may result in different explanation of sentences. Some reduction of ambiguities of DMs depends upon human's common sense knowledge.

During the process of segmentation, all DM candidates are matched and classified by regular expression rules. Then structure ambiguities will be resolved by applying resolution principles and segmentation models. Sense and function ambiguities are expected to be resolved by different approaches during postprocessing. Some DMs, such as *yidian*, are three way ambiguous. The resolutions of their structure ambiguities have to be delayed until sense ambiguities are resolved. For

instance, temporal DMs are by default segmented into one unit first, which specifies time points and whose POS is Nd. If they are identified as sense of duration at postprocessing stages, the one unit DMs will be re-segmented into two units, i.e. a number followed by a measure. In future work, the debatable issue whether *yue* (月) is a quasi measure or an ordinary individual noun is our concerns. The insertion of adjectives into DMs and the reduplication of DMs are also worthy in investigation.

References

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chines*e. Berkeley: University of California Press.

Chinese Knowledge Information Processing Group. 1996. *ShouWen JieZi - A Study of Chinese Word Boundaries and Segmentation Standard for Information Processing* [In Chinese]. CKIP Technical Report 96-01. Taipei: Academia Sinica.

Crystal, David. 1991. *A Dictionary of Linguistics and Phonetics*. Cambridge, Massachusetts: Blackwell.

He, Jie (何杰). 2002. *《現代漢語量詞研究》*.民族出版社.

Li, Charles N. and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Ma, Wei-Yun and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. In *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, pp168-171.

MDNA (Mandarin Daily News Association 國語日報出版社) and CKIP (中央研究院詞庫小組). 1997.《國語日報量詞典》.

Mo, Ruo-ping Jean, Yao-Jung Yang, Keh-Jiann Chen and Chu-Ren Huang. 1991. Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation. In *Proceedings of ROCLING IV (R.O.C. Computational linguistics Conference)*. pp. 111-134.

Prins, Robbert Paul. 2005. *Finite-State Pre-Processing for Natural Language Analysis*. Art Dissertation.

Smith, Carlota S. 1991. *The Parameter of Aspect*. Dordrecht: Kluwer Academic Publishers.

Tai, James H-Y. 1994. Chinese classifier systems and human categorization. In *In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*. Edited by Matthew Y. Chen and Ovid J.L. Tzeng. Taipei: Pyramid Press. pp. 479-494.

You Jia-Ming and Keh-Jiann Chen, 2004. Automatic Semantic Role Assignment for a Tree Structure. In *Proceedings of 3rd ACL SIGHAN Workshop*. Barcelona Spain.

Website Resources

Sinica Corpus. http://www.sinica.edu.tw/SinicaCorpus/

Sinica Treebank. http://treebank.sinica.edu.tw/

Applying Maximum Entropy to Robust Chinese Shallow Parsing

Shih-Hung Wu*†, Cheng-Wei Shih†, Chia-Wei Wu†, Tzong-Han Tsai†,
and Wen-Lian Hsu†

†*Institute of Information Science, Academia Sinica, Taiwan, R.O.C*
(shwu,dapi,cwwu,thtsai,hsu)@iis.sinica.edu.tw

**Dep. Of CSIE, Chaoyang University of Technology, Taichung County, Taiwan, R.O.C*
shwu@cyut.edu.tw

Abstract

Recently, shallow parsing has been applied to various information processing systems, such as information retrieval, information extraction, question answering, and automatic document summarization. A shallow parser is suitable for online applications, because it is much more efficient and less demanding than a full parser. In this research, we formulate shallow parsing as a sequential tagging problem and use a supervised machine learning technique, Maximum Entropy (ME), to build a Chinese shallow parser. The major features of the ME-based shallow parser are POSs and the context words in a sentence. We adopt the shallow parsing results of Sinica Treebank as our standard, and select 30,000 and 10,000 sentences from Sinica Treebank as the training set and test set respectively. We then test the robustness of the shallow parser with noisy data. The experiment results show that the proposed shallow parser is quite robust for sentences with unknown proper nouns.

1. Introduction

Parsing is a basic technique in natural language processing; however, a full parser is usually costly and slow. Recently, shallow parsing has been applied to various information processing systems [12]. Compared to the performance of full parsers, a shallow parser is much faster and the parsing result is more useful for various applications, such as information retrieval and extraction, question answering, and automatic document summarization. In this paper, we adopt a machine learning approach to the Chinese shallow parsing problem.

Chinese full parsing is very challenging,[18, 22] because it is difficult to achieve high accuracy, and the performance is not suitable for online applications. Shallow parsing of Chinese, on the other hand, is promising and desirable in terms of efficiency. Researchers in Beijing, Harbin, Shenyang, and Hong Kong have also developed related techniques [10, 15, 16, 20, 21]. Most of these works use machine learning approaches, instead of the rule-based approach used in full parsing. Popular machine learning methods such as SVM, CRF, and ME, have been tested. The parsing speed of each approach is fast and the parsing accuracy is acceptable.

Currently, there is no standard for Chinese shallow parsing. Li et. al. [9] developed a Chinese shallow parsed treebank to extract Chinese collocations automatically and built a large collocation bank. There are also some works on a standard for Chinese shallow parsing [9, 19, 20]. Nevertheless, the POS standard and vocabulary in each approach are different; thus, between simplified Chinese and traditional Chinese, we cannot adopt their standard for simplified Chinese to traditional Chinese. Instead, we use the first level of the parsing results of Sinica Treebank as our shallow parsing standard [4]. Originally, Sinica Treebank was designed to provide full parsing results, whereby sentences could be labeled with POS tags and the full parsing structure. There are 54,000 sentences in Sinica Treebank, from which we randomly selected 30,000 and 10,000 sentences as the training set and test set respectively.

Since there are many unknown words in Chinese [11], a Chinese shallow parser must be robust against such words [22]. For example, it is not hard to correctly chunk the sentence “高漸離/擊筑/的/音調/忽然/急轉成/悲壯” into “高漸離擊筑的音調/NP 忽然/Dd 急轉成/DM 悲壯/VP”, if we know that “高漸離” is a proper noun. However, if the name is unknown, it could be split into three single characters and tagged with the three POS of the single characters, i.e., “高/漸/離 [VH13/Dd/P15]”. It might then be incorrectly chunked as ”高漸/NP 離擊筑的音調/PP 忽然/Dd 急轉成/DM 悲壯/VP”. In this research, we simulate unknown words by adding some noises to the corpus in order to test the robustness of the shallow parser. Since new proper nouns are normally unknown, we design three ways to add noises to the training and testing sets by treating proper nouns as unknown words.

2. Shallow Parsing Standard

Sinica Treebank provides a full parse tree for each sentence. Here, we use the first-layer parsing results of Sinica Treebank as the standard for shallow parsing. Instead of using all the phrase tags in Sinica Treebank, we annotate five of them for chunking; all other phrases (including single words not in any phrase) are tagged as others (X). The five tags, namely, noun phrase (NP), verb phrase (VP), preposition phrase (PP), geographic phrase (GP), and clause (S), are the major tags in Sinica Treebank, and therefore play significant syntactical roles. Thus, the constituents of the root node of a parse tree are NP, VP, PP, GP, S, and X. Table 1 lists examples of the six types of constituent.

Table 1. Chunk Tags

Chunk Tag	Description	Example
NP	Noun Phrase	前十名 / 的 / 選手 [DM / DE / Nab]
VP	Verb Phrase	傳遞 / 區運 / 聖火 [VD1 / Nad / Nac]
PP	Preposition Phrase	在 / 旅客 / 口 / 中 [P21 / Nab / Nab / Ncda]
GP	Geographic Phrase	一個 / 星期 / 以後 [DM / Nac / Ng]
S	Clause	窗戶 / 玻璃 / 破掉 [Nab / Nab / VH11]
X	Others	０ / 到 / ２度 [DM / Caa / DM]

3. A Maximum Entropy-based Shallow Parser

Parsing is a fundamental technique in natural language processing, the results of which can be used to improve various natural language tasks, such as word-sense disambiguation (WSD) [3] and part-of-speech (POS) tagging [12].

Many natural language processing tasks, such as part-of-speech tagging, named-entity recognition, and shallow parsing, can be viewed as sequence analysis tasks. Shallow parsing identifies the non-recursive core of each phrase type in a text as a precursor to full parsing or information extraction [1, 6]. The paradigmatic shallow parsing problem is called NP chunking, which finds the non-recursive cores of noun phrases called base NPs. Ramshaw and Marcus introduced NP chunking as a machine-learning problem [14].

Machine learning techniques, such as maximum entropy (ME) and conditional random fields (CRF), are quite popular for sequential tagging. We adopt ME to build a robust Chinese shallow parser.

3.1 The B-I-O Scheme of Our Shallow Parser

In this work, we regard each word as a token, and consider a test corpus and a set of *n* phrase categories. Since a phrase can have more than one token, we associate two tags, *x*: *x_begin* and *x_continue*, with each category. In addition, we use the tag *others* to indicate that a token is not part of a phrase. The shallow parsing problem can then be redefined as a problem of assigning one of $2n + 1$ tags to each token. This is called the B-I-O scheme. There are 5 named entity categories and 11 tags: *NP_begin, NP_continue, VP_begin, VP_continue, PP_begin, PP_continue, GP_begin, GP_continue, S_begin, S_continue,* and *X(others).*

3.2 Maximum Entropy Formula

ME is a flexible statistical model that assigns an *outcome* to each token based on its *history* and *features* [2]. The outcome space is comprised of the tags for an ME formulation. ME computes the probability $p(o|h)$ for any o from the space of all possible outcomes, O, and for every h from the space of all possible histories, H. A *history* is composed of all the conditioning data that enables one to assign probabilities to the space of outcomes. In shallow parsing, *history* can be viewed as all the information derived from the test corpus relevant to the current token.

The computation of $p(o|h)$ in ME depends on a set of binary-valued *features*, which is helpful in making a prediction about the outcome. For instance, one of our features is as follows: when the current token is a verb, it is likely to be the leading character of a verb phrase. More formally, we can represent this feature as

$$f(h,o) = \begin{cases} 1: \text{if Current - token - verb(h)} = \text{true and } o = VP\_begin \\ 0: \text{else} \end{cases} \quad (1)$$

Here, *Current-token-verb(h)* is a binary function that returns the value *true* if the *current token* of the history *h* is a verb.

Given a set of features and a training corpus, the ME estimation process produces a model in

which every feature f_i has a weight α_i. This allows us to compute the conditional probability as follows:

$$p(o \mid h) = \frac{1}{Z(h)} \prod_i \alpha_i^{f_i(h,o)}, \qquad (2)$$

where Z(h) is a normalization factor. Intuitively, the probability is the multiplication of the weights of active features (i.e., those f_i (h,o) = 1). The weight α_i is estimated by means of a procedure called Generalized Iterative Scaling (GIS) [8], which improves the estimation of the weights at each iteration. The ME estimation technique guarantees that, for every feature f_i, the expected value of α_i will be equal to the empirical expectation of α_i in the training corpus. ME allows the designer to concentrate on finding the features that characterize the problem, while letting the ME estimation routine deal with assigning relative weights to the features.

3.3 Decoding

After an ME model has been trained and the proper weight α_i has been assigned to each feature f_i, decoding (i.e., *marking up*) a new piece of text becomes a simple task. First, the model tokenizes the text and preprocesses the test sentence. Then, for each token, it checks which features are active and combines the α_i of the active features according to Equation 2. Finally, a Viterbi search is run to find the highest probability path through the lattice of conditional probabilities that does not produce any invalid tag sequences. Further details of the Viterbi search can be found in [17].

4. Experiment

By comparing models with and without noisy training data, we can determine whether our Chinese shallow parser is noisy-data-tolerant. In this section, we describe how we add noisy data to maximum entropy models and evaluate the tolerance of our system to Chinese chunking.

4.1 Data and Features

Sinica Treebank contains more than 54,000 sentences, from which we randomly extract 30,000 for training and 10,000 for testing. The tokenized results and the corresponding part-of-speech sequences of these sentences are extracted into a feature file, and the top-level chunks of the parsing tree structure can be taken as the standard for training and evaluation. The information in the feature file is translated into machine learning features by ME model in both the training and testing phrases. The features we adopted are: words, adjacent characters, prefixes of words (1 and 2 characters), suffixes of words (1 and 2 characters), word length, POS of words, adjacent POS tags, and the word's location in the chunk it belongs to.

To analyze the performance of our shallow parser under noisy conditions, we build a standard model and various noisy models. Training data consisting of the tokenization and POS information derived from the manually annotated Sinica Treebank is used as the standard model in our experiments. The accuracy of chunking in this model is then compared with that of models containing noise to

observe the difference.

4.2 Noise Model Generation

The most important issue in noisy model generation is how to mix noisy features with correct features as smoothly as in a real parsing system. We design three methods for adding noise to generate different types of models with noisy tokenization and POS sequences.

The first two approaches are based on unknown word replacement. We find that unknown words are one of the major causes of noisy data in real world system processing, because most unknown words are proper nouns. Theoretically, we can pick a certain number of proper nouns in the selected data and substitute them with noisy data to simulate real world input. In our experiment, “Nb” and “Nc”, which are defined as “proper nouns” and “proper location nouns” respectively in the Sinica Treebank tagging guideline [5], are chosen as replacement targets. Words with these two target POS are regarded as replacement target strings and replaced by noisy data.

We adopt two types of noisy data for unknown word replacement. The first is the split character sequence of a replacement target string in a sentence. Initially, we extract the correct tokenization results and POS sequences of all data in the Sinica Treebank with “Nb” and “Nc”. Then, wherever applicable, we split the replacement target string in a sentence into single Chinese characters. The corresponding POS tag of each split character is re-assigned by selecting the most frequent POS tags of these single characters in Sinica Treebank. For example, “馬來西亞” (Malaysia) would be split into “馬”, ”來”, ”西”, and “亞”, and the original POS tag “Nca” would be replaced by the pos tags of four single characters: “Nab”, “Dbab”, “Ncda”, and “Nca”. In this experiment, we control the amount of noisy data in models to observe the relation between the percentage of imprecise data and the chunking performance. The model generated by this approach is called a Type 1 noise model. Another approach, called the Type 2 noise model, tokenizes the replacement target with AUTOTAG, which may produce segmenting boundaries and POS tags that differ from those in Sinica Treebank. The information is then used as noisy features and replaces the target string. For instance, the replacement target string “太白金星” with POS tag ”Nb” would be tagged by AUTOTAG as “太白/Nb” and “金星/Nb”. The above noise-adding approaches are used to generate training data, as well as various kinds of noisy information in the test sets.

In addition, we adopt an automatic tool, CKIP AUTOTAG [7], to obtain the tokenization information and POS features for generating models. This is a Chinese tokenizing tool that can deal with word segmentation in both the training and testing sets. CKIP AUTOTAG provides the POS sequences of the sentences. The tokenized sentences and POS sequences produced by AUTOTAG are used to generate feature files for ME processing.

5. Results and Discussion

In our experiment, we adopt the B-I-O scheme to identify the boundaries of Chinese chunks and the position of each element word in the chunks. In addition, we employ the following four standards

when calculating the accuracy of Chinese shallow parsing: evaluation by token, by chunk boundary, by chunk category sequence, and by chunks. Token evaluation is based on the number of Chinese words. All words in the test data can be verified independently to determine if they have the correct boundaries and belong to the right chunks. Evaluation by chunk boundary only checks the boundaries of each chunk, while evaluation by chunk category sequence only checks if all the chunks in a sentence can be identified successfully and disregards the constituents. By contrast, in chunk evaluation, the basic unit is the whole chunk, and only a chunk with the right constituents and tagged with proper categories can be considered correct. We use an example to demonstrate the evaluation process. The input sentence is "小朋友 換成 你 來 試試看", which consists of five tokens; and the standard parsing result is "小朋友/NP 換成/VC 你/NP 來-試試看/VP", which contains four chunks. The parsing result we obtain from the system is "小朋友/NP 換成/VC 你/NP 來/Db 試試看/VE", which contains five chunks. In this case, the accuracy of the chunk boundary and the chunk category are both 3/4=0.75, because the first three chunks in the sentence have the correct boundaries and phrase tags, and the last VP chunk is separated by two units. The token number in this sentence is 5 and the last two tokens have incorrect phrase category tags. Therefore, the accuracy of the token is 3/5=0.6. In chunk evaluation, three of the four chunks are identified successfully and the chunk accuracy is 3/4=0.75. We adopt these evaluation methods in all the experiment configurations in Tables 2 to 5.

5.1 Performance on Noisy Data

Table 2 shows the accuracy rates using Type 1 noisy models with different scales of noisy data for chunking clean test data. The columns show the percentage of 'Nb' and 'Nc' replaced by single character noisy data in the training model, and the rows indicate the four evaluation methods. We find that the accuracy in this series decreases slightly, while the percentage of single character noisy data increases.

Table 2. Results of chunking clean test data with the Type 1 noise model

	Boundary	Category	Tokens	Chunks
0 (%)	84.83	70.10	69.14	70.47
10 (%)	84.74	69.92	69.04	70.30
20 (%)	84.80	69.94	69.03	70.26
30 (%)	84.77	69.88	69.10	70.20
40 (%)	84.70	69.77	68.97	70.13
50 (%)	84.64	69.65	69.02	70.00
60 (%)	84.56	69.57	68.78	69.82
70 (%)	84.42	69.39	68.76	69.59
80 (%)	84.53	69.67	68.99	69.77
90 (%)	84.38	69.44	68.58	69.72
100 (%)	84.26	69.51	68.57	69.75

Table 3 shows the accuracy rates using the Type 1 model with different scales of noisy data for chunking test data with single character noise (Type 1). It is quite interesting that the curve is not monotonically increasing or decreasing. This indicates that the accuracy in this series decreases until the percentage of noise reaches 60%, and then it increases. Figures 1 to 4 show the differences between the clean test data and the noisy test data in Tables 2 and 3. We can observe the trends in the experiment results more intuitively.

Table 3. Results of chunking test data containing Type 1 noisy data with the Type 1 noise model

	Boundary	Category	Tokens	Chunks
0 (%)	83.73	69.09	65.16	66.51
10 (%)	83.65	69.00	65.33	66.36
20 (%)	83.72	69.11	65.34	66.30
30 (%)	83.69	69.20	65.37	66.27
40 (%)	83.62	69.14	65.42	66.25
50 (%)	83.58	69.05	65.52	66.13
60 (%)	83.57	69.07	65.36	66.00
70 (%)	83.52	69.24	65.70	66.07
80 (%)	83.63	69.46	65.83	66.25
90 (%)	83.65	69.49	65.69	69.30
100 (%)	83.77	69.67	65.85	66.42

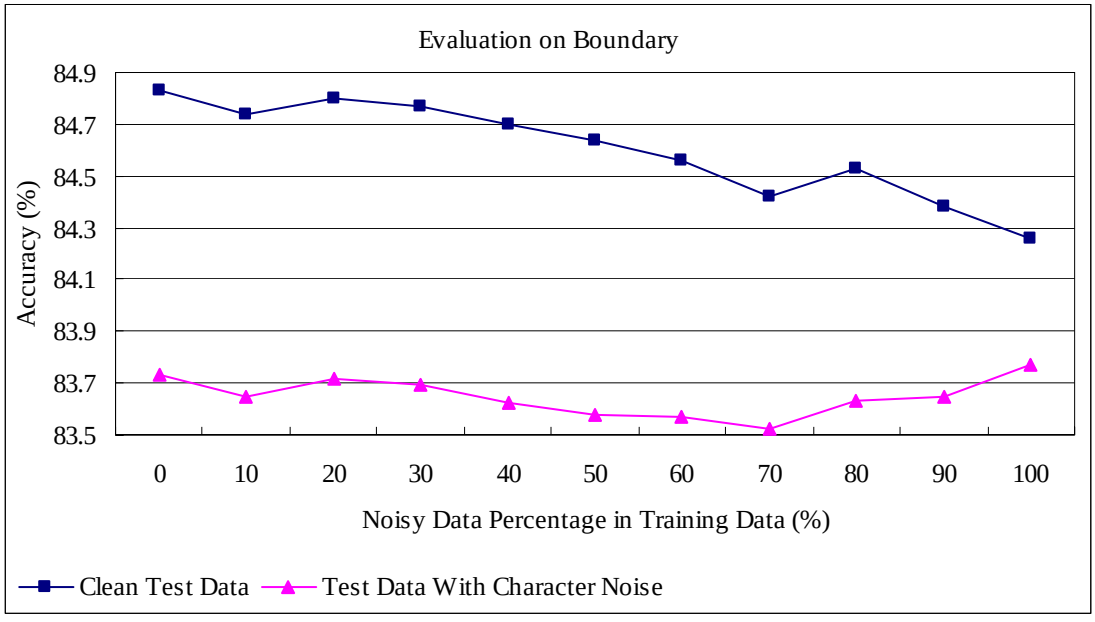


Figure 1. Evaluation of the boundaries in different experiment configurations

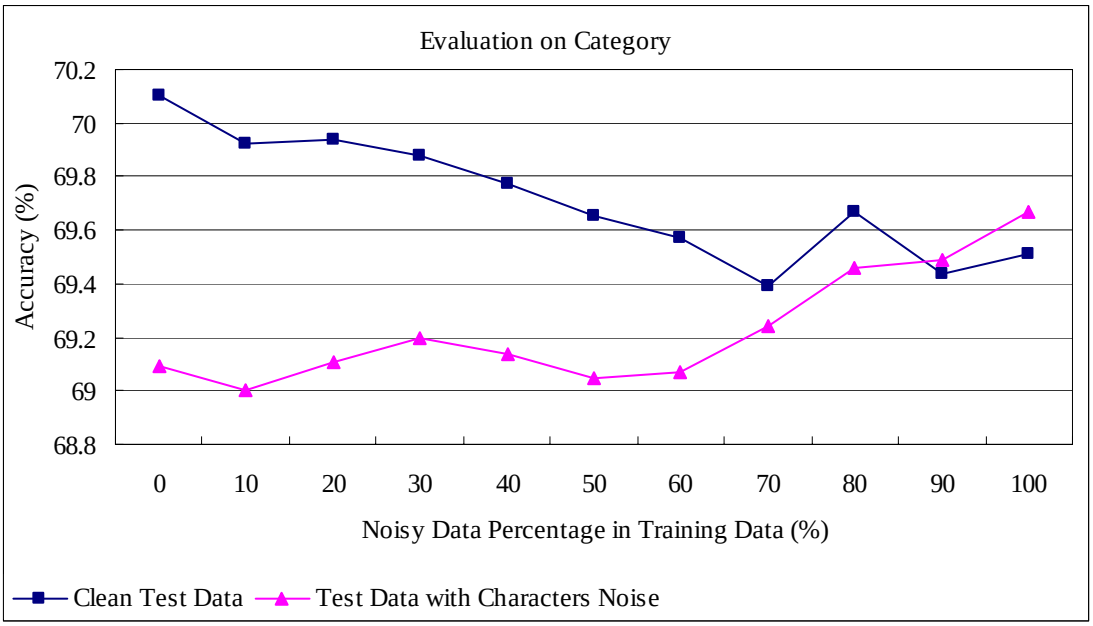


Figure 2. Evaluation of the chunking category in different experiment configurations

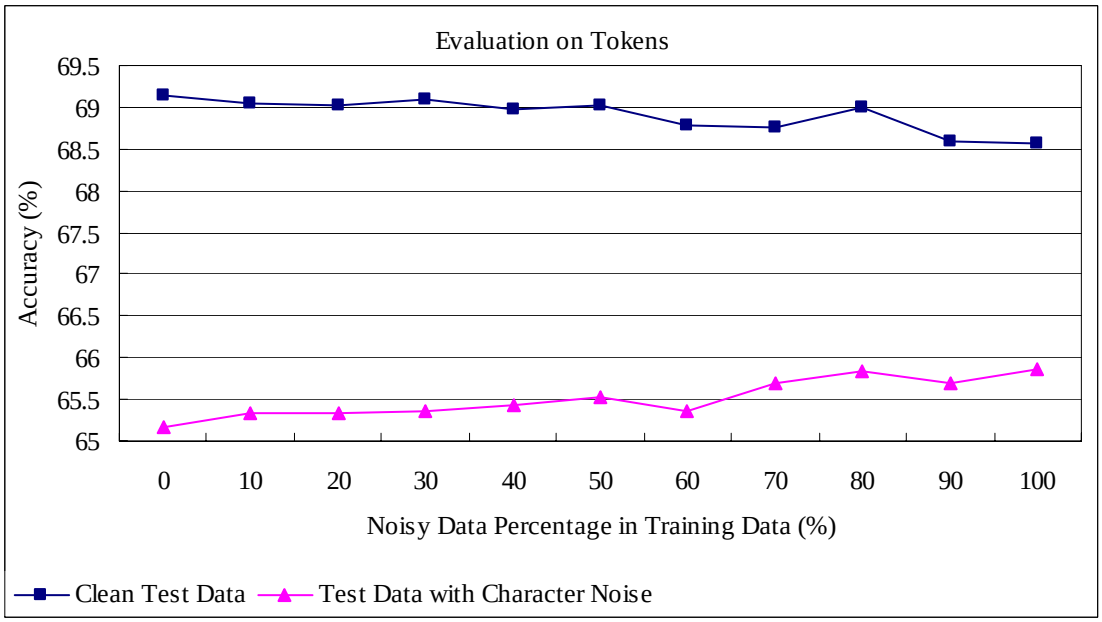


Figure 3. Evaluation of tokens in different experiment configurations

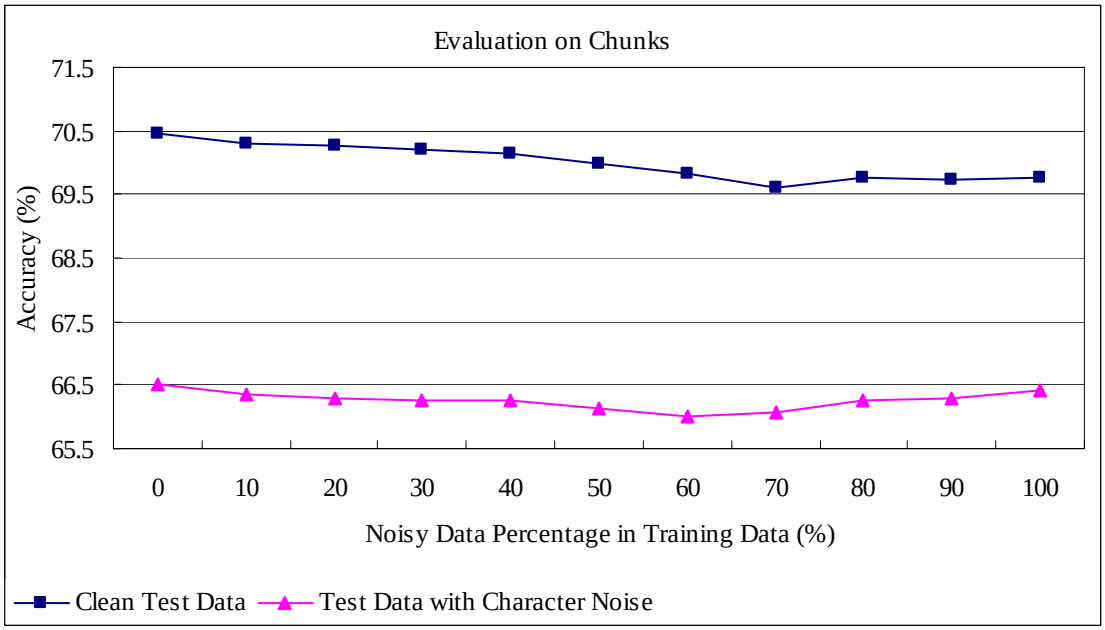


Figure 4. Evaluation of chunks in different experiment configurations

Table 4 shows the accuracy rates using the Type 2 noise model with and without tokenized strings for chunking clean test sentences and test data with tokenized strings. There are four configurations:

- C-C: Using a clean training model and clean test data.
- C-N: Using a clean training model and noisy test data in which all 'Nb' and 'Nc' are replaced by tokenized results.
- N-C: Using a training model with noisy data in which all 'Nb' and 'Nc' are replaced by the tokenized results of chunking clean test data.
- N-N: Both the training model and the test data have noisy data in which all 'Nb' and 'Nc' are replaced by tokenized results.

Table 4 also shows that noisy training data yields better accuracy for both clean and noisy test data, although the difference is quite small.

Table 4. Results of chunking with the Type 2 noise model

	Boundary	Category	Tokens	Chunks
C-C	84.83	70.10	69.14	70.47
C-N	84.84	70.09	69.04	70.37
N-C	84.89	70.13	69.15	70.51
N-N	84.90	70.11	69.02	70.38

Table 5 shows the accuracy rates using the model generated by AUTOTAG-parsed data and Sinica Treebank chunking tags. Both the training and the test sets are preprocessed by AUTOTAG. This experiment is designed for open testing; thus, we can use the AUTOTAG program to tokenize any

sentence and give it POS tags. However, compared to the standard model, the chunking accuracy is lower. The parsing results of the AUTOTAG-parsed model and the Type 2 noise models are shown in Figure 5.

Table 5. Accuracy using the model generated by AUTOTAG-parsed data

	Boundary	Category	Tokens	Chunks
Fully AUTOTAG	81.42	64.81	61.80	61.30

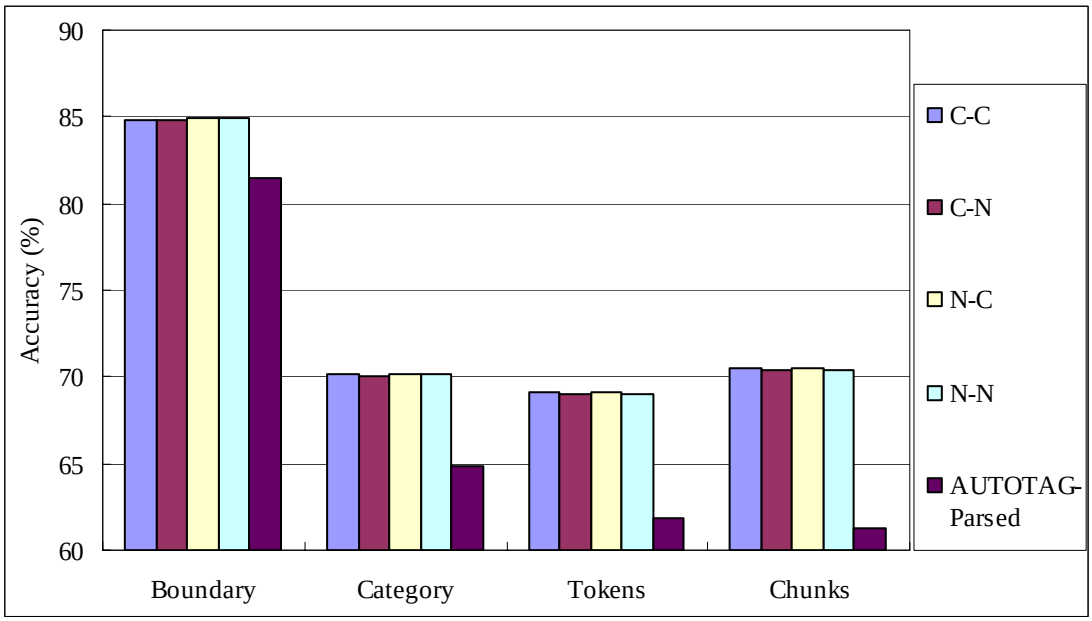


Figure 5. Comparison of various experiment configurations using tokenized string noisy data (the Type 2 noise model) and the AUTOTAG-parsed model

In Tables 6, 7, and 8, we give examples of the correct and incorrect shallow parsing results of four sentences. In each table, the left column contains the original sentences tokenized and tagged with POS tags; the center column shows the standard chunking result from Sinica Treebank; and the right column shows the shallow parsing result of our system. Table 6 shows the parsing examples with Type 1 noise. The shallow parsing results of the first two sentences are correct, while those of the last two sentences are incorrect.

Table 6. Shallow parsing examples with Type 1 noise

Sentence and POS sequences with Type 1 noise	Chunking standard from Sinica Treebank	Chunking results of our system
女性/形象/在/**台/灣**/和/**中/國/大/陸**/小說/是/解放/的/過程 [Nab/Nac/P21/**Nca/Nab**/Caa/**Ng/Ncb/VH13/Nab**/Nac/V_11/VC2/DE/Nac]	女性形象/NP 在台灣和中國大陸小說/PP 是/V_ 解放的過程/NP	女性形象/NP 在台灣和中國大陸小說/PP 是/V_ 解放的過程/NP

與/**中/華**/及/**日/本/隊**/在/伯仲/之間 [P35/**Ng/Nca**/Caa/**Nca/Nes/Nab**/VC1/Nhac/Ng]	與中華及日本隊/PP 在/VC 伯仲之間/GP	與/P3 中華及日本隊/NP 在/VC 伯仲之間/GP
首先/義賣/的/是/**黑/將/軍/史/東**/的/手套 [Cbbb/VC31/DE/V_11/**VH11/Dd/Nab/Nad/Ncda**/DE/Nab]	首先義賣的/NP 是/V_ 黑將軍史東的手套/NP	首先/Cb 義賣的/NP 是/V_ 黑將軍史東的手套/NP *
學藝/股長/**王/文/星**/站起來/說 [Nad/Nab/**Nbc/Nab/Nab**/VA11/VE2]	學藝股長王文星/NP 站起來/VA 說/VP	學藝股長王文星站起來/NP 說/VP *

Table 7 shows the parsing examples with Type 2 noise. The shallow parsing results of the first and the last sentences are correct, while those of the second and the third sentences are incorrect.

Table 7. Shallow parsing examples with Type 2 noise

Sentence and POS sequences with Type 2 noise	Chunking standard from Sinica Treebank	Chunking results of our system
女性/形象/在/台灣/和/**中國/大陸**/小說/是/解放/的/過程 [Nab/Nac/P21/Nca/Caa/**Nc/Nc**/Nac/V_11/VC2/DE/Nac]	女性形象/NP 在台灣和中國大陸小說/PP 是/V_ 解放的過程/NP	女性形象/NP 在台灣和中國大陸小說/PP 是/V_ 解放的過程/NP
與/中華/及/**日本/隊**/在/伯仲/之間 [P35/Nba/Caa/**Nc/Na**/VC1/Nhac/Ng]	與中華及日本隊/PP 在/VC 伯仲之間/GP	與/P3 中華及日本隊/NP 在/VC 伯仲之間/GP *
首先/義賣/的/是/**黑/將軍**/史東/的/手套 [Cbbb/VC31/DE/V_11/**VH/Na**/Nba/DE/Nab]	首先義賣的/NP 是/V_ 黑將軍史東的手套/NP	首先/Cb 義賣的/NP 是/V_ 黑將軍史東的手套/NP *
學藝/股長/**王/文星**/站起來/說 [Nad/Nab/**Nb/Nb**/VA11/VE2]	學藝股長王文星/NP 站起來/VA 說/VP	學藝股長王文星/NP 站起來/VA 說/VP

Table 8 shows the parsing results using AUTOTAG-parsed training data and test data. The results of the first and last sentences are correct, while those of the second and the third sentences are incorrect. We replace the original word segmentation and POS tags of all the sentences with AUTOTAG-parsed word segmentation and POS tags. The word segmentation of the last sentence provided by AUTOTAG is incorrect; however, the chunking result is correct.

Table 8. Shallow parsing examples with AUTOTAG-parsed training data and test data

AUTOTAG-parsed Sentence and POS sequences	Chunking standard from Sinica Treebank	Chunking results of our system

女性/形象/在/台灣/和/中國/大陸/小說/是/解放/的/過程 [Na/Na/P/Nc/Caa/Nc/Nc/Na/SHI/VC/DE/Na]	女性形象/NP 在台灣和中國大陸小說/PP 是/V_ 解放的過程/NP	女性形象/NP 在台灣和中國大陸小說/PP 是/V_ 解放的過程/NP
與/中華/及/日本/隊/在/伯仲/之間 [P/Nc/Caa/Nc/Na/P/Nh/Ng]	與中華及日本隊/PP 在/VC 伯仲之間/GP	與中華及日本隊/PP 在伯仲之間/PP *
首先/義賣/的/是/黑/將軍/史東/的/手套 [D/VC/DE/SHI/VH/Na/Nb/DE/Na]	首先義賣的/NP 是/V_ 黑將軍史東的手套/NP	首先/Cb 義賣的/NP 是/V_ 黑將軍史東的手套/NP *
學藝/股長王/文星/站起來/說 [Na/Nb/Nb/VA/VE]	學藝股長王文星/NP 站起來/VA 說/VP	學藝股長王文星/NP 站起來/VA 說/VP

The experiment results show the noise-tolerance of our Chinese shallow parser with two different kinds of noise from unknown proper nouns. The system's performance is only degraded slightly when noisy data is added. Most sentences, such as "六十年代的台灣是怎樣的形貌" in which "台灣" is split into two characters and assigned with incorrect POS tags, can still be identified. However, the token accuracy is a little lower than the chunk accuracy, which indicates that our system needs to be improved for chunking longer phrases. In contrast, the chunking accuracy obviously decreases if models fully generated by AUTOTAG-parsed data are used. The difference between the AUTOTAG and Sinica Treebank tag sets probably causes the accuracy to decrease. Furthermore, this suggests that, while the shallow parsing system can deal with unknown nouns, it has difficulty dealing with other kinds of noisy data. For example, data preprocessing errors, such as, incorrect tokenization or wrong tagging in other POS categories, affect the performance of shallow parsing substantially We can not comment on which part-of-speech tags are the major factors in Chinese chunking without conducting additional experiments.

5.2 Use of Our Shallow Parser on News Articles

For the first application of our shallow parser, we collect some news articles as the test set. The articles did not have standard word segmentation, POS tagging, and parsing results; therefore, we cannot report on the accuracy. However, we find the results interesting. Some examples are given in Table 9. The left column shows the original sentences tokenized and tagged with POS tags by AUTOTAG. The right column shows the shallow parsing results using our system.

One interesting point is that the shallow parser tends to group named entities into a phrase. Therefore, the shallow parsing result can be used as a feature for boundary detection in named entity recognition (NER). In sentence 1, "中鋼公司" is grouped as one phrase, and in sentence 9, "中鋼公司 88 年盈餘" is grouped as one phrase, without first recognizing that "中鋼公司" is an entity by NER. Another example, in sentence 2 is that "益華在花蓮的三棟大樓" is grouped as one phrase, without first recognizing that "益華" is a company name.

Table 9. Shallow parsing results for news articles

	Tokenization and POS of Sentences	Shallow Parsing Result
1	中 鋼 / 公 司 / 是 / 台 灣 / 鋼 鐵 業 / 龍 頭 [Nc/Nc/SHI/Nc/Na/Na]	中鋼公司/NP 是 台灣鋼鐵業龍頭/NP
2	益/華/在/花蓮/的/三棟/大樓/有/二/棟/是/七層/建築 [VJ/ Nc/P/Nc/DE/Nb/Na/V_2/Neu/Nf/SHI/Na/Na]	益華在花蓮的三棟大樓/NP 有 二棟/NP 是 七層建築/NP
3	大陸/仍/有/廣闊/發展/空間 [Nc/D/V_2/VH/VC/Na]	大陸/NP 仍 有 廣闊發展空間/NP
4	光/是/中共/國家/主席/江澤民/就/出/訪五次 [Da/SHI/Nb/Na/Na/Nb/D/VC/Na]	光/NP 是 中共國家主席江澤民/NP 就出訪五次/PP
5	許 多 / 地 區 / 都 / 出 現 / 新 舊 / 共 存 / 的 / 景 觀 [Neqa/Nc/D/VH/Na/VH/DE/Na]	許多地區/NP 都 出現 新舊共存的景觀/NP
6	過去/一年/是/兩岸/關係/比較/困難/、/且/希望/落空/的/一年 [Nd/Nd/SHI/Nc/Na/Dfa/VH/PAUSECATEGORY/Cbb /VK/VH/DE/Nd]	過去一年/NP 是 兩岸關係比較困難、且希望落空的一年/NP
7	恆生/指數/創下/歷史/新高 [Nb/Na/VC/Na/VH]	恆生 指數/NP 創下 歷史新高/NP
8	將/資本主義/及/投機/氣息/帶入/大陸/內部 [P/Na/Caa/VH /Na/VCL/Nc/Ncd]	將資本主義及投機氣息/PP 帶入大陸內部/NP
9	中鋼/公司/88/年盈餘/可望/達到/140 億/元/左右 [Nc/Nc/Neu/Na/VK/VJ/Neu/Nf/Ng]	中鋼公司 88 年盈餘/NP 可望達到 140 億元/VP 左右
10	企業界/已/開始/尾牙/聚餐 [Nc/D/VL/Nd/VA]	企業界/NP 已 開始 尾牙聚餐/VP
11	投資/人/靜候/美國/聯邦/準備/理事會/(/Fed/)/21 日/的/利率/決策 [Nc/Na/VJ/Nc/Na/VC/Na/PARENTHESISCATEGORY/FW /PARENTHESISCATEGORY/Nd/DE/Na/Na]	投資人/NP 靜候 美國聯邦準備理事會(Fed)21 日的利率決策/NP
12	央 行 / 總 裁 / 及 / 理 監 事 / 都 / 有 / 一 定 / 的 / 任 期 [Nc/Na/Caa/Na/D/V_2/A/DE/Na]	央行總裁及理監事/NP 都 有 一定的任期/NP

6. Conclusion and Future Works

In this paper, we propose a Chinese shallow parser that can chunk Chinese sentences into five chunk types. We test the noise tolerance of the shallow parser and found that the accuracy of data with simulated unknown words only decreases slightly in chunk parsing. We also test our Chinese shallow parser on an open corpus, and found that it yields interesting chunking results.

Tolerance of unknown words is an essential characteristic of a Chinese shallow parser. In this paper, we demonstrate our parser's robustness in handling noisy data from proper nouns. However, we could not verify the robustness of chunking noisy data from other kinds of POS. Thus, adopting other POS systems, such as the Penn Chinese Treebank tagset, for Chinese shallow parsing could prove both

interesting and useful. In the future, we will improve our model by adding more types of noise, such as random noise, filled noise, and repeated noise proposed by Osborne [13]. In addition to Sinica Treebank, we will extend our training corpus by incorporating other corpora, such as Penn's Chinese Treebank.

Acknowledgements

This research was supported in part by the National Science Council under GRANT NSC94-2752-E-001-001-PAE.

References

1. Abney, S.P. Parsing by Chunks. in Berwick, R.C., Abney, S.P. and Tenny, C. eds. *Principle-Based Parsing: Computation and Psycholinguistics*, Kluwer, Dordrecht, 1991, 257-278.
2. Berger, A., Della Pietra, S.A. and Della Pietra, V.J. A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*. 39-71.
3. Bikel, D.M., A Statistical Model for Parsing and Word-Sense Disambiguation. in *the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, (Hong Kong, 2000), 155-168.
4. Chen, F.-Y., Tsai, P.-F., Chen, K.-J. and Huang, C.-R. 中文句結構樹資料庫的構建. *Computational Linguistics and Chinese Language Processing*, *4* (2). 87-104.
5. Chen, K.-J., Huang, C.-R., Chen, F.-Y., Luo, C.-C., Chang, M.-C., Chen, C.-J. and Gao, Z.-M. Sinica Treebank: Design Criteria, Representational Issues and Implementation. in Abeille, A. ed. *Treebanks Building and Using Parsed Corpora. Language and Speech series*, Kluwer, Dordrecht, 2003, 231-248.
6. Church, K.W., A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. in *the Second Conference on Applied Natural Language Processing*, (1988), 136-143.
7. CKIP. Autotag, Academia Sinica, 1999.
8. Darroch, J.N. and Ratcliff, D. Generalized iterative scaling for log-linear models. *Annals of Mathematicl Statistics*, *43*. 1470-1480.
9. Li, B., Lu, Q. and Li, Y., Building a Chinese Shallow Parsed TreeBank for Collocation Extraction. in *CICLing*, (2003), 402-405.
10. Lu, Q., Zhou, J. and Xu, R.-F., Machine Learning Approaches for Chinese Shallow Parsers. in *International Conference On Machine Learning And Cybernetics*, (Xi'an, 2003), 2309- 2314.
11. Ma, W.-Y. and Chen, K.-J., A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. in *the Second SIGHAN Workshop on Chinese Language Processing*, (2003), 31-38.
12. Müller, F.H. and Ule, T., Annotating topological fields and chunks - and revising POS tags at the same time. in *Nineteenth International Conference on Computational Linguistics (COLING 2002)*, (Taipei, Taiwan, 2002), ACM, 695-701.

13. Osborne, M. Shallow Parsing using Noisy and Non-Stationary Training Material. *Journal of Machine Learning Research*, *2*. 695-719.
14. Ramshaw, L.A. and Marcus, M.P., Text chunking using transformation-based learning. in *The ACL Third Workshop on Very Large Corpora*, (1995), 82-94.
15. Tan, Y., Yao, T., Chen, Q. and Zhu, J., Applying Conditional Random Fields to Chinese Shallow Parsing. in *CICLing*, (2005), 167-176.
16. Tan, Y., Yao, T., Chen, Q. and Zhu, J., Chinese Chunk Identification Using SVMs plus Sigmoid. in *The First International Joint Conference on Natural Language Processing (IJCNLP-04)*, (2004), 527-536.
17. Viterbi, A.J. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, *IT*. 260-269.
18. XIA, X. and WU, D., Parsing Chinese with an almost-context-free grammar. in *EMNLP-96, Conference on Empirical Methods in Natural Language Processing*, (Philadelphia, 1996).
19. Xu, R.-F., Lu, Q., Li, Y. and Li, W., The Construction of A Chinese Shallow Treebank. in *the Third SIGHAN Workshop on Chinese Language Processing*, (2004), 94-101.
20. Zhang, L.roach to Extract Chinese Chunk Candidates from Large Corpora. in *20th International Conference on Computer Processing of Oriental Languages (ICCPOL03)*, (ShenYang, P.R.China, 2003).
21. Zhao, T.-J., Yang, M.-Y., Liu, F., Yao, J.-M. and Yu, H., Statistics Based Hybrid Approach to Chinese Base Phrase Identification. in *Second Chinese Language Processing Workshop*, (Hong Kong, China, 2001), 73-77.
22. Zhou, M., A block-based robust dependency parser for unrestricted Chinese text. in *The second Chinese Language Processing Workshop attached to ACL2000*, (Hong Kong, 2000).

異體字語境關係的分析與建立

周亞民
台灣大學資訊管理研究所
milesymchou@yahoo.com.tw

黃居仁
中央研究院語言學研究所
churen@gate.sinica.edu.tw

摘要-利用計算機處理漢語，實際上是透過漢語的書寫形式，而異體字是漢語書寫形式的特性，但是長久以來異體字的關係沒有被適當的表達，而且將異體字關係過度的簡化爲全同異體字，而實際上不同的異體字在使用上並不完全相同，本研究之目的是表達異體字的關係，並提出詞義、聲韻、構詞、時間、空間、構詞的語境(context)模型，根據此模型分析和建立異體字的關係，作爲中文資訊處理的基礎資源。

1. 簡介

以計算機處理自然語言，需要解決語言形式與語意之間的關係，對於資訊檢索和機器翻譯皆是如此，WordNet 的重要性就是因爲它建立了不同的詞彙形式和語意之間的關係，不過，異體字和異體詞的關係，並不是藉由 WordNet 可以建立的，主要原因是漢語的書寫形式不同。漢語的書寫形式是表意書寫系統(Ideographic Writing System)[Miller 1991][Coulmas 2003]，表意文字是概念的具體表徵，由於漢字並不是特定的人所創造，因此不同字形表示相同的概念，或不同的概念使用相同的字形，是很普遍的現象，而且隨著時間的變遷，使用的範圍廣，字形所表示的意義擴大或縮小，最後交織成複雜的一詞多形關係，這些關係將不同的字形連結成爲異體字。

異體字關係是文字學的主要議題，異體字可以分爲全同異體和部份異體，全同異體指的是音義完全相同而字形不同的字，而部份異體只需有部份用法相同即可，用法完全相同的稱爲狹義異體字，廣義異體字則包含部份異體字和全同異體字[裘錫圭 1995]。也有文字學家認爲只有全同異體字才能被認定是異體字 [董琨 1993][洪成玉 1995] 文字學中討論的正字、俗字、通假字、假借字、古今字、重文等，都是與異體字有關，但是都只由一個面向討論，缺乏整體的架構可以分析和比較異體字的關係，因此，也使得要分析異體字的關係不容易。因爲計算機的編碼系統採用一個字形一個字碼，因此對於計算機而言，只要字碼不同就是不同的字，例如：群≠羣、說≠説，而造成中文資訊處理上的問題，尤其是檢索，但是它們都是同一個字的異形或異寫的異體字。爲了要在計算機中表達異體字的關係，最早的是謝清俊教授設計的 CCCII(Chinese Character Code for Information Interchange)[謝清俊、黃克東 1989]，其後則是漢字構形資料庫[莊德明 1999][莊德明、謝清俊 2005]，以及中央研究院歷史語言研究所袁國華教授提出的建立 UNICODE 漢字異體字表與異體字辭典之相關研究[袁國華、曾黎明 2005]。這些研究所建立的異體字關係，大部份都將異體字關係簡化爲全同異體字，但是真正用法完全相同的異體字不多，其它都是只有部份用法相同的部份異體字[裘錫圭 1995]，因此，我們提出一個能夠表達部份異體字關係的模型，並且據以表達異體字的關係，作爲中文資訊處理的基本資源。本研究是漢字知識本體研究中的一個部份，漢字知識本體除了描述異體字關係，還描述書寫形式、語音、字義、變異、詞彙衍生，以 OWL-DL 描述漢字的知識，並且與 IEEE SUMO(Suggested Upper Merged Ontology)和

GOLD(General Ontology for Linguistic Description)整合，以提供計算機在自然語言處理所需的書寫、構詞、語法知識[周亞民 2005] 。

2. 異體字的語境模型

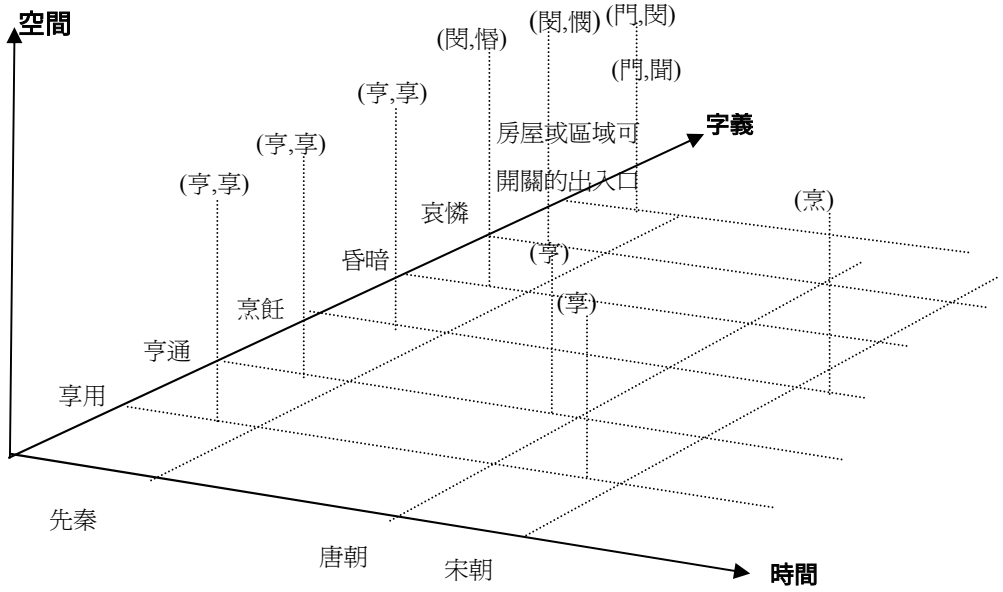
本研究對異體字的定義是音義全部或部份相同而字形不同即是異體字，甲骨、金、籀、篆、隸、楷等不同的文字發展不是異體字，因爲其改變是全面性的，不是發生在某一個字形的改變。我們將異體字的語境區分爲意義、時間、空間和構詞，分述如下：

(1)詞義

第一個重要的語境是兩個字形在什麼詞義時可以互相使用，如「蛇」古作「它」,「它」的本義爲蛇，但因常常假借爲第三人稱，因此，另加意符「虫」表示其本義，這兩個字只有「蛇」的意義爲古今字，第三人稱則不是古今字。又如陣戰的「陣」古作「陳」，論語衛靈公：「衛靈公問陳於孔子」，後改「東」爲「車」，說文沒有「陣」但收「陳」,「陳」與「陣」只有在陣戰的意義爲古今字。又如「戚」的本義爲兵器，假借爲憂慼的「慼」，論語：「小人長戚戚」，後造本字「慼」，故「慼」與「戚」在憂慼是假借關係的異體字。又如「拾」與「十」在數字的意義爲異體，其它的意義並沒有異體字關係，在「路不拾遺」這個詞彙中就不能寫「路不十遺」，因爲「拾」的意思是撿取而非數字，因此，必需區別「拾」與「十」在那些意義爲異體，以判斷兩個字形是否有異體字關係。

(2)時間

異體字關係的第二個重要的語境是時間。現代漢語的書寫形式中不存在著異體字關係，並不表示古代漢語不是異體字，例如「亨」與「享」，現在不是異體字，但是漢朝以前兩字可以通用，說文解字注：「亨，隸書作亨作享，小篆之變也」,「亨」與「享」是隸變後寫法不同，但意思相同的字，干祿字書：「亨」是亨通、亨宰,「享」是祭亨，因此，到了唐朝仍然以「亨」表「享」，而「享」已經沒有「亨」的意思，而「烹」爲亨飪的俗字，說文、玉篇和廣韻都沒有「烹」字，直到類篇和集韻才有，因此，可能到了宋朝才用「烹」表示烹飪的意思(圖一)。



圖一 異體字之意義、時間與空間的關係

再以正俗字爲例，正俗字大部份的詞義原來都是相通的，但是後來可能變成用法不同的字，如「邪」與「耶」原爲用法相同的異體字，干祿字書說：「耶邪：上通下正」，「邪」本義爲地名，後假借爲疑問詞「耶」，史記：「羽豈其苗裔邪？」又因爲「耳」與「牙」在漢朝時兩個字形接近，成爲沒有區別的異體字，但是現在兩個字形變成用法不同，疑問詞的「耶」不作「邪」，正邪不作「耶」。

每個漢字所表達的意義，會隨著時間而改變，什麼時候開始有某個意義，會影響與其它字形之間是否爲異體字的判斷，要知道某個字形何時有此意義並不容易，如果要進一步知道兩個字形在特定的時間，有那些共同的意義就更加困難，但也不是不可能，由於本研究關心的是異體字之間的關係，所以並不需要將造字以來曾經有的意義都加以分析整理，最重要的異體字之間有交集的意義，盡可能的由字書和例證中去找尋證據。

(3)空間

由於漢字在不同地區的使用差異，也造成了異體字的現象，尤其是國家處於分裂的情形，異體字的現象更爲顯著，以春秋戰國時期來說，當時處於分裂的國家，文字的差異很大，說文：「田疇異畮，車涂異軌，律令異法，衣冠異制，言語異聲，文字異形」，文字異形指的就是戰國的字形因地而異的現象，同一個概念，有的國家用本字，其它國家用假借字，或不同國家用不同的假借字，例如「門」本作「門」，但是齊國假借「聞」作「門」，燕國假借「閔」作「門」，另外，還有部件使用不同，如：秦國的「廚」字爲从广从尌，而三晉卻是从广朱聲[裘錫圭 1995]。中國大陸於 1956 年開始進行了漢字簡化方案後，使漢字的結構大幅的改變，也造成了兩岸在文字使用上的差異，產生簡繁字形的異體字。

(4)構詞

異體字雖然音義相同，但是在構詞上可能不同，例如「記」和「紀」爲部份異體，在「記載」這個概念時，兩字爲異體，但是古漢語在這個概念的使用上沒有區別，但是現代漢語卻有區別，例如: 「筆記本」不寫爲「筆紀本」，「記者」不寫「紀者」，「紀念日」不寫「記念日」，「記憶」不寫「紀憶」[裘錫圭 1995]。

又如「昇」與「升」雖然有異體字關係，但是在構詞上並不一定能夠相替換，例如當作上升之意時，「升高」可作「昇高」，「提升」也可寫「提昇」，但是「昇華」不作「升華」，又如字義爲登進時，「升學」不作「昇學」，「升官」不作「昇官」[洪嘉馡、巫宜靜、黃居仁 2005]。簡化字與繁體字也有相同的情形，例如「藉」簡化爲「借」，但是「狼藉」不可簡化爲「狼借」，「慰藉」也不可簡化爲「慰借」，又如「乾」簡化爲「干」，然而「乾坤」不簡化爲「干坤」，「乾隆」不簡化爲「干隆」[江藍生、陸尊梧 2004]

與異體字構詞有關係的還有異體詞，例如：「按語」和「案語」、「梅雨」和「霉雨」， 異體詞又可以分爲有異體字關係和沒有異體字關係，如「梅雨」和「霉雨」，「梅」和「霉」並不是異體字，兩個字的意義並不相同，因江南一帶梅子成熟季節連綿降雨而得「梅雨」一詞，「梅」的概念是梅子，而「霉雨」是因爲連日下雨東西容易發霉而得霉雨，「霉」的概念是發霉，故「梅

雨」與「霉雨」是不存在異體字關係的一詞異形，這個問題是中文詞網可以解決的，在本研究只要描述「雨」可以衍生「梅雨」和「霉雨」，在中文詞網將「梅雨」和「霉雨」放入同一個同義詞集，計算機就知道梅雨與霉雨同義，而本研究關心的異體字問題，因爲中文詞網並不處理異體字問題，而有異體字關係的一詞異形，本身就是異體字，當然可以交換使用，只要整理異體字關係就可以解決。

3. 異體字關係的建立

我們採用中文電腦基本用字的常用字集、說文解字和漢語大字典的異體字表的共同交集共二千七百組作爲分析的對象。分析的重點包括不同字形在那些字音、字義和時間有異體字關係、字書描述異體字關係的體例和異體字構詞的限制。字音依古音、中古音和現代音分別建立，字義則區分本義、引伸義和假借義，對於異體字使用的時間和字義，則以例證的根據。 異體字關係的建立分爲三個部份：

(1) 建立字書對異體字關係的描述

確認異體字關係最基本的證據來自於字書，因此，建立異體字關係的第一步就是必須能夠描述字書對異體字關係的解釋。不同字書對異體字的描述體例不同，我們整理出來字書對異體字的體例如下：

A. 古作

「古作」描述異體字的古今字關係，例如：

集韻：「嶽，古作。」

龍龕手鑑：「巖，古作岩。」

玉篇：「巨，古作丘。」

B.古文

「古文」也是描述古今字關係，例如：

說文：「仅，古文奴。」

一切經音義：「氾，古文泛。」

龍龕手鑑：「  ，古文米字。」

C.今作

「今作」描述古今字關係，例如：

玉篇：「  ，導也，今作唱。」

玉篇：「  ，移也，今作徒，同。」

龍龕手鑑：「  ，今作艱。」

D.後作

「後作」描述的也是古今字關係，例如：

漢語大字典：「或，後作國。」

漢語大字典：「暴，後作曝。」

E.本作

「本作」大部份描述的是本字，例如：

正字通：「瀏，本作。」」

玉篇：「噚，本作粵。」

F.本字

「本字」大部份也是描述本字，例如：

字彙：「喪，喪本字，从哭从亡。」

正字通：「弘，彈本字。」

G.通

「通」在異體字描述的關係比較複雜，最早使用「通用字」或「通」的體例是干祿字書[章瓊2004]，干祿字書的體例關係是通用的俗體字，其它字書有的是同源字，也有近義字，最多是指通假字，漢語大字典的體例也是通假字。

集韻：「儃，讓也，通作禪、嬗。」

正字通：「疑，又與擬通。」

干祿字書：「虛，虛的通行體。」

漢語大字典：「卿，通慶。」

H.俗作

「俗作」描述的是正俗字的關係，例如：

龍龕手鑑：「峨，俗作峩。」

字彙：「兔，俗作兎。」

正字通：「健，俗作健。」

干祿字書：「突，突，上俗，下正。」

I.用同

「用同」在漢語大字典中所表示的異體字關係爲後起同音替代字，例如：

漢語大字典：「咨，用同齜。」

漢語大字典：「廢，用同費。」

J.也作

「也作」在漢語大字典用的很多，只能描述兩字有異體字關係，例如：

漢語大字典：「忍，也作肭。」

漢語大字典：「轏，也作棧。」

K.亦作

「亦作」只能描述兩字有異體字關係，但無法明確描述何種異體字關係，例如：

說文通訓定聲：「發，亦作發。」

集韻：「背，違也，亦作貲。」

古今韻會舉要：「膳，亦作善。」

L.或作

「或作」也是只能描述兩字有異體字關係，例如：

集韻：「剡，利耜也，或作覃。」

龍龕手鑑：「峭，或作峭，峭，正。」

M.或從

「或從」以說文和集韻使用較多，並不能明確描述何種異體字關係，例如：

集韻:「榜，進船也，或从手。」

說文：「𨒫，正行也，从辵，正聲，征，或从彳。」

N.同

「同」只能描述兩字有異體字關係，並不一定是全同異體字，例如：

龍龕手鑑：「弘，同弘。」

廣韻：「弘，同引。」

漢語大字典:「巖，同壧。」

字彙補:「恖，與憂同。」

正字通：「㑹，與嶺同。」

O.籀文

「籀文」描述異體字爲籀文，例如：

說文:「童，男有辠曰奴，奴曰童，女曰妾，从辛，重省聲，𠇍，籀文童。」

字彙補:「畬，籀文畬字。」

R.簡化字

簡化字是漢語大字典的體例，描述繁簡關係的異體字，例如：

漢語大字典：「发，發，髮的簡化字。」

漢語大字典：「欢，歡的簡化字。」

這些體例有些比較明確的表達異體字的關係，包括：古作、古文、今作、俗作、籀作、通、用同和簡化字，其它的則只能知道有異體字關係，包括：同、亦作、也作、或作、也作、或從，都無法判斷是什麼關係，對於我們判斷異體字關係，最好能找到比較明確的關係，而且相同的異體字，不同的字書描述內容也不一定全然相同，例如漢語大字典：「騱，也作奔」，玉篇：「騱，今作奔」，篇海類篇：「騱，與驎同，亦作奔」，所以，每一組異體字，都盡可能同時將多本字書的考證加入，可以幫助我們得到較完整的輪廓。

另外，不同字書使用相同的體例，可能表示的是不同的異體字關係，這種情形主要發生在「通」這個體例，例如干祿字書將異體字分爲正、俗、通，「通」指的是通用字，但是漢語大字典的「通某」指的是通假字，集韻的「通作某」則不一定是通假關係。

對計算機而言，古今字、正俗字、假借字和通假字等都只能提供非常有限的異體字描述，應該將異體字的關係做更詳細的描述，才能提供計算機處理異體字的根據，字書對異體字的描述通常都不會很詳細，但是字書提供了漢字關係的基礎。

(2)異體字的時間面向

此面向描述什麼時間那些漢字有異體字關係，彼此又有那些共同的意義，而我們主要的就是字書和例證。字書對於漢字的使用情形，與例證所提供的資訊不同，前者只能確定若某字爲字書所收，則此字應出現在該字書成書之前，但是此字於字書成書時是否仍在使用，則必需根據例證。特定朝代的文字書寫習慣，會表現在該時代的文獻中，如果能夠找到出更多在該時代的文獻作爲例證，則更有充分的證據可以確定這些文字的使用並非源自於版本相異或作者個人的風格。例如「歸」的初義是女子出嫁，說文：「歸，女嫁也」，引申爲返回和歸依，但在古文獻中多假借爲「饋(贈送)」，書經：「唐叔得禾……王命唐叔歸於東」，又如論語：「陽貨欲見孔子，孔子不見，歸孔子豚」，又詩經：「自牧歸荑，洵美且異」。由這些先秦文獻中，可以得知在秦以前，「饋」有贈送

之義，而說文皆收有「歸」和「饋」，而依廣韻的記載，「饋」是求位切，「歸」依集韻記載亦爲求位切，有了這些證據就可以充分說明在先秦時代，「歸」和「饋」爲部份異體，當作贈送時兩字有異體字關係。

要找出異體字的使用情形，必須從大量的古文獻中尋找，我們可考慮使用數位化的古籍資料庫，但是處理古籍資料時，以今字取代古字是很常見的現象，而且古今皆是如此，檢索數位化的古籍時，對於所出現的文字，很難判斷是否真的是原來的用字，例如以中研院漢籍電子文獻檢索「茶」，可以在史記/列傳/卷九十三/韓信盧綰列傳第三十三找到：「漢十一年秋，陳豨反代地，高祖如邯鄲擊豨兵，燕王綰亦擊其東北。當是時，陳豨使王黃求救匈奴。燕王綰亦使其臣張勝於匈奴，言豨等軍破。張勝至胡，故燕王臧荼」子衍出亡在胡……，但是「荼」字最早應該出現在隋唐之間，但利用資料庫檢索，卻可以在西漢找到「茶」字，主要的原因也可能以今字代古字，因此，如果要找出異體字在不同朝代的使用，利用現有的資料庫雖然有效率，但是所使用的版本是非常關鍵的問題。

因爲兩漢以前的文獻由於能夠留下來的較少，現在我們可以看到的大部份都是隋唐以後所抄寫或刻印，很多文字都已經改用隋唐的習用字，因此，引用兩漢以前的古籍，不一定能夠完全反應先秦和兩漢的文字使用情形，但是傳抄和刻印仍有相當多的古文字被留下來，因此，還是有很好的參考價值，只是我們在引用時，常會懷疑字書所引用的正確性，必須要多持保留的態度，書證的引用必須要找到好的例子，參考來源的版本要經過校對。

目前台灣所編的字典對例證方面並不重視，除了字義的解說外，少數的字典有例句，但是對於我們建立異體字的時間關係沒有任何的幫助，因爲這些例句大部份都是出版社自行造句，最多只能反映現代漢語的使用情形，而沒有其它時代的使用情形。相較之下，康熙字典與漢語大字典則收錄了大量古籍書證。如果將康熙字典與漢語大字典比較，我們可以發現最主要的差異之一，後者有較多的考證資料，更重要的是包括近百年來的許多考古發現，例如敦煌莫高窟藏經的發現、山東銀雀山西漢墓漢簡、馬王堆帛書、睡虎地秦墓竹簡等考古發現，對於我們建立異體字的時間關係非常的重要，漢語大字典也搜集了部份的新證據，例如「復」字，古代借爲「腹(肚子)」，什麼時侯開始「復」有腹義，從睡虎地秦墓竹簡中可以找到使用的例子：早到室即病復痛，因此，我們可以說至少在先秦以前，「復」假借爲「腹」，最晚什麼時侯仍借「復」表「腹」，漢書中可以找到「復」心弘道，惟賢聖兮，所以至少到了東漢「復」仍然借爲「腹」，現在則已經沒有此義。

(3)異體字字義面向

異體字字義面向描述漢字之間有那些意義存在著異體字關係，目前異體字整理的較好的是漢語大字典與教育部的異體字資料庫，但是對於異體字之間究竟有那些共同的意義，並沒有清楚的描述。如果將漢語大字典與教育部的異體字資料庫比較，前者的異體字表雖然簡略，但是可以在字典中找出它們的關係，而且除了有字書的書證還有例證，不過異體之間有那些共同的意義仍然不夠清楚，必須要仔細比較兩個漢字之間的字形結構、字音、本義、引申、通假和假借關係，或考證其它資料，才能確認是否有異體字關係，以及有那些共同意義。

字書中對異體字關係的描述，究竟兩字爲通用或部份異體判斷上並不容易，例如龍龕手鑑：「㑹，古文，音財」，字彙補：「㑹」，古文財字」，即「㑹」和「財」爲古今字，但是這兩字只有「財物」和「財產」基本意義是相同，而財有很多通假義，包括「才能」、「剛剛」、「材料」等，

無法證明「𠂤」也有「財」的通假義[趙振鐸 2003]，因此，我們將「財」與「𠂤」的關係描述只有在解釋爲「財物」和「財產」爲部份異體字。

我們在分析異體字的共同字義時，同時會描述字義是本義、引申或假借，這些描述對於異體字的關係，是很重要的依據，例如甲乙兩字本義不同，乙字出現的字書較早，而甲字的引申義與乙字的本義相同，乙字現在已經失去本義，而前者的引申義仍在使用中，我們就可以判斷乙爲古字，甲爲今字，共同的意義亦非本義，則可知道甲乙應該不是一字異體。

那些意義是引申義，必須要先確認本義，但是要探求漢字的初義並不容易，因爲很多漢字都是數千年前所創造的，再加上字形經過了幾次重大的改變，更增加了考證的困難。但是掌握漢字的本義非常重要的，因爲本義與字形之間的關係密切，掌握本義則能夠找出詞義引申的擴展和脈絡，以及漢字彼此之間的關係，進一步確定形成異體字的原因。

由於確認漢字的本義有相當的困難，所以同一個漢字其本義存在多種不同的看法，主要的因素端看所發現的證據，以及對這些證據所作的詮釋。我們針對這個問題的解決方法，以說文的解釋爲主，說文以小篆作爲研究本義的基礎，難免會有錯誤，但是到目前爲止，要探求本義，說文仍是非常重要的參考。

漢字除了本義之外，還有很多意義，這些意義大部份都是由本義所引申而來的。在古漢語中，由於溝通的需要，因而產生新的詞彙，這些詞彙除了可以創造新字來表達，另外就是直接透過字義的引申，當然也可以利用產生雙字詞或多字詞的方式表達，不過古漢語使用單字詞較現代漢語普徧，引申是成爲表達概念的重要方式，這種方式最重要的好處是可以控制新字產生的數量。引申義有時候較本義更爲常用，因此失去了本義，就只好另外造字，或假借其它的字表達本義，例如「北」的小篆爲兩人背對背，後來引申爲北方(與中原相背)之後，本義逐漸失去，所以又另外造了「背」字表示本義，「北」與「背」成爲古今字。如果不描述它們的字義是本義、引申或假借，就無法描述這個層面關係。

(4)異體字字音面向

建構異體字關係必需要考慮字音，尤其是通假與假借關係，都是建立在字音相同或相近的基礎。漢字有一字多音和義隨音轉的特性，因此，異體字的共同意義的部份，一定要找出究竟字音爲何，才能確定異體字關係。不過很多異體字音已經改變，因此，不能以現代音作爲判斷的基礎，而必須從中古音和古音著手，才能確認異體字的關係。例如說文：「容，盛也，从宀，谷」。依說文的解釋「容」的本義爲容納盛載，說文：「庸，用也，从用，从庚，庚，更事也」，意思是採用或需要，但釋名：「容，用也」。依釋名的解釋，「庸」與「容」有共同的意義，但是從字形分析來看，兩字皆爲會意字，彼此之間沒有增減形符或改變形符，從金文和小篆字形也找不到共同的形符與演變關係，因此，如果釋名的解釋是正確的，那麼只有一個可能就是假借或是通假，但是「容」的現代音爲ㄖㄨㄥ 2，而「庸」的現代音爲ㄩㄥ，似乎沒有假借或通假的可能，然而依廣韻兩字的中古音皆爲餘封切，又古韻皆爲東韻，由此，可以確定兩者之間應爲假借或通假關係。

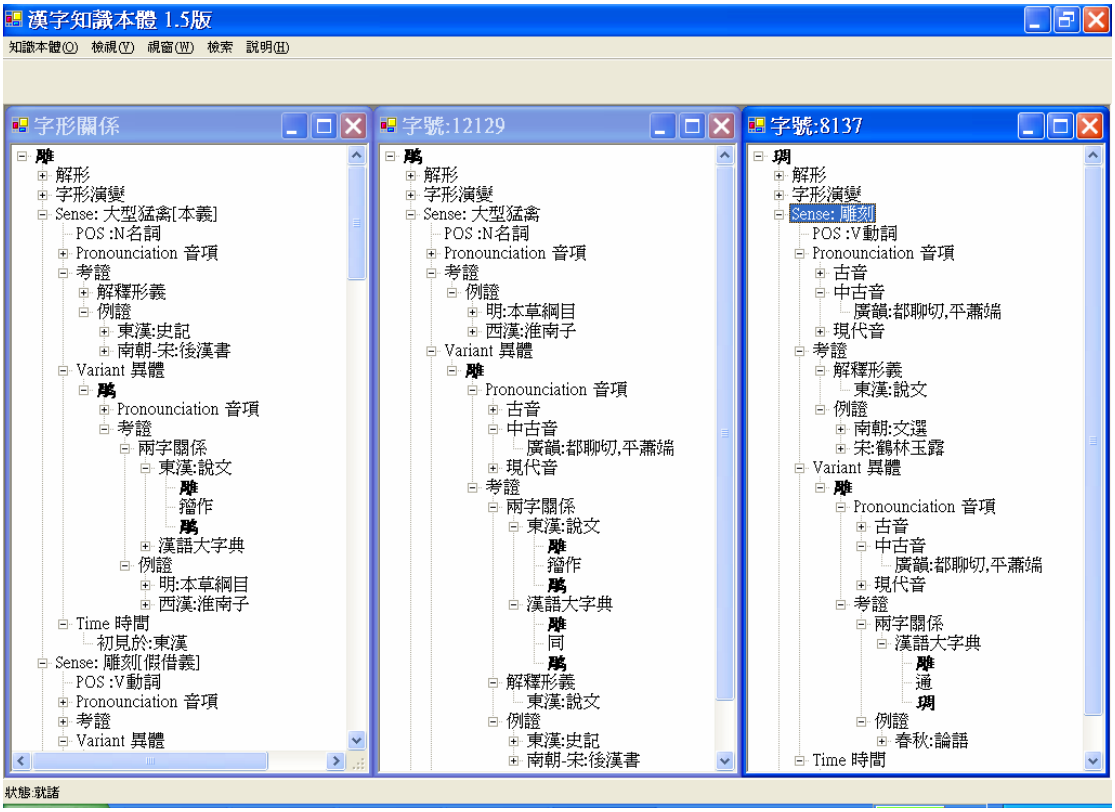
古音反應的是先秦的語音系統，由於時間久遠，研究上較爲困難，因此意見較分歧，如果仔細比較，古音韻部的研究較聲類有較爲一致的研究成果，故我們古音部份只先描述韻部。雖然古韻的研究較古聲類有一致的結果，但是仍然有不同的分部方法，例如顧炎武分爲十部，王念孫分爲二十一部[董同龢 1979]，王力則分爲三十部，漢語大字典也分爲三十部。由於漢語大字典有相

當的權威性，且收字較多，比較容易找出古韻的資料，因此異體字字音知識的部份，我們以漢語大字典的古韻三十部作爲古韻的分部系統。中古音最有代表性的字書是廣韻和集韻，其注音方式採用反切音，因爲同韻或同聲鈕可以使用不同的反切下字或上字，確認異體字字音不能只依賴反切上字與下字是否相同，而是應該要依賴聲鈕和韻調，因此，我們也將廣韻的反切上字與下字整理，做爲確認異體字的參考。

4. 研究結果與應用

爲了驗證本研究的異體字語境模式，我們以兩年的時間，描述了三千個異體字關係，並且以本研究提出的模型進行表達，以雕爲例說明異體字關係描述的結果(圖二)，對於「雕」與「鵰」的描述是「鵰」爲「雕」的籀文，又根據本研究所對於意符「鳥」與「隹」都表達鳥類的概念，而漢字在使用時經常會更換其同表達相同概念的意符，因此，不僅知道「鵰」爲「雕」的籀文，更可以知道爲什麼它們形成不同的字形結構。另外，再依描述「鵰」與「雕」爲部份異體字，只有當作「大型猛禽」爲異體關係，因爲「雕」有「用彩畫裝飾」和「雕刻」的意義，而「鵰」沒有這些意義，另外「雕」與「鵰」在東漢以前就有「大型猛禽」的意義。

而「雕」與「琱」的關係，根據異體字關係的描述，「雕」假借爲「琱」，且爲有本字假借，而且早在春秋時「雕」即有「雕刻」的意義，「琱」的意符是「玉」，表達的概念是「玉石」，而「雕」的意符是「隹」，表達的概念是「鳥」，兩個字的意符概念差異表示兩字的造字本義應該不同，古韻皆爲幽部，因此，兩字應爲假借關係，雕與琱只有在「雕刻」這個意義有異體關係，因此爲部份異體關係。



圖二　雕的異體字關係

爲了說明本研究對於異體字關係表達的優點，我們將本研究與電腦漢字字形與詞彙整合知識庫比較，該計劃是是數位典藏國家型計劃技術分項計劃－建立 Unicode 漢字異體字表與異體字辭典之相關研究的成果之一，研究動機是因爲 Unicode 的異體字很多，對於計算機的應用造成處理上的困擾，並且嘗試要表達部份異體字關係[袁國華、曾黎明 2005]。此研究是文獻中與本研究最相關的研究，缺乏其它相關的研究，也顯示出異體字關係在計算機長久以來不受重視的事實．我們以「雕」、「鵰」、「琱」三個異體字爲例進行比較，電腦漢字字形與詞彙整合知識庫對於這三個字的關係描述代碼是 H，即它們都是漢語大字典所收的異體字，這個描述計算機只能知道「雕」、「鵰」、「琱」有異體字關係，但是究竟是什麼關係，就沒有任何的描述。相同的異體字，本研究能夠充分的表達在那些意義和什麼時間兩個字形是異體字，以及在當時的字音是什麼，以了解是否有假借關係，本研究的異體字描述架構優於電腦漢字字形與詞彙整合知識庫的異體字描述。

圖三　電腦漢字字形與詞彙整合知識庫的異體字關係[袁國華、曾黎明 2005]

我們將本研究所建立的異體字關係應用在異體字的檢索，且將重點放在部份異體字的檢索。本研究將異體字檢索分爲共時(synchronically)與歷時(diachronically)，共時檢索對於不同年代的文獻，均視爲同一個斷代，而歷時檢索則會根據被檢索文獻的時代產生不同的異體字。歷時異體字檢索考慮的是異體字關係，如果不是異體字關係的差異則不是本研究能夠提供的知識，也不是歷時異體字檢索，例如現代漢語用「衣架」，古漢語用「桁」，如樂府詩集中有「還視桁上無懸衣」，那麼如果檢索詞是「衣架」，就必需以「桁」進行古漢語文獻的檢索，又如現代漢語用「黑馬」，古漢語用「驪」，這些詞彙關係並非異體字。無論是共時檢索或歷時檢索，如果根據本研究所建立的異體字關係知道檢索詞有部份異體字，檢索系統會要求檢索者選擇檢索詞的詞義，根據詞義決定適當的候選異體字，如果被檢索文件的年代是候選異體字使用的時段，就會將後選異體字一起加入檢索字，反之如果並不在異體字的使用時段，就不會加入檢索字。如果檢索詞是雙字詞或多字詞，則會分別找出異體字，加入檢索詞彙後進行檢索。

4.1 共時異體字檢索

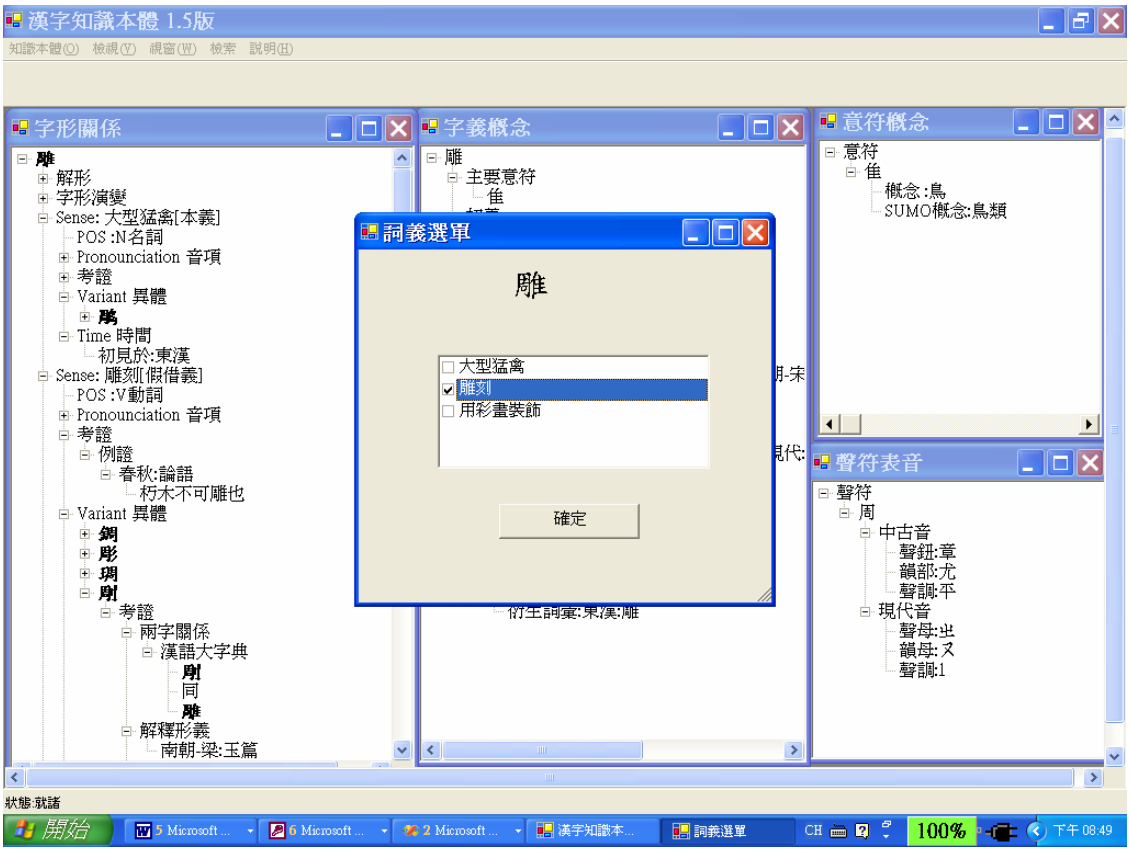
我們以「雕」作爲共時異體字檢索的例子，根據漢語大字典「雕」的意義有:一種大型凶猛的鳥、兇猛、雕刻和用彩繪裝飾等，根據我們所建立的異體字關係描述了「雕」分別在不同的意義有異體字，包括「鵰」、「彫」、「琱」、「剛」、「鍋」，異體字中的「鵰」只有當大型猛禽時與「雕」爲異體，「彫」與「雕」則是在雕刻和用彩繪裝飾是異體字，「琱」、「剛」、「鍋」等皆與「雕」在雕刻的意義爲異體。如果以「雕」進行檢索，由於「雕」有部份異體字，檢索系統會要求確認檢索字義，若以雕刻爲字義，檢索詞會加入異體字「彫」、「琱」、「剛」和「鍋」。表一是作爲雕的異體字檢索的文件集合，這些文件分別用了不同的異體字，來源包括中央研究院漢籍電子文獻的二十五史資料庫、中央研究院平衡語料庫和聯合報聯合知識庫。

檢索結果可以發現出現「雕」、「彫」、「琱」、「剛」和「鍋」的文件都被找到，但是「睽違 12 年白尾海鵰現蹤」、「新校本隋書/紀/卷四　帝紀第四/煬帝下/大業十二年」、「新校本舊唐書/列傳/卷八十八　列傳第三十八」、「新校本宋史/志/卷一百四十九　志第一百二/輿服一/五輅」、「平衡語料庫」等有異體字「鵰」的文件不會被當作通用異體字而被檢索出來，因爲「鵰」與「雕」只在大型猛禽爲異體字。

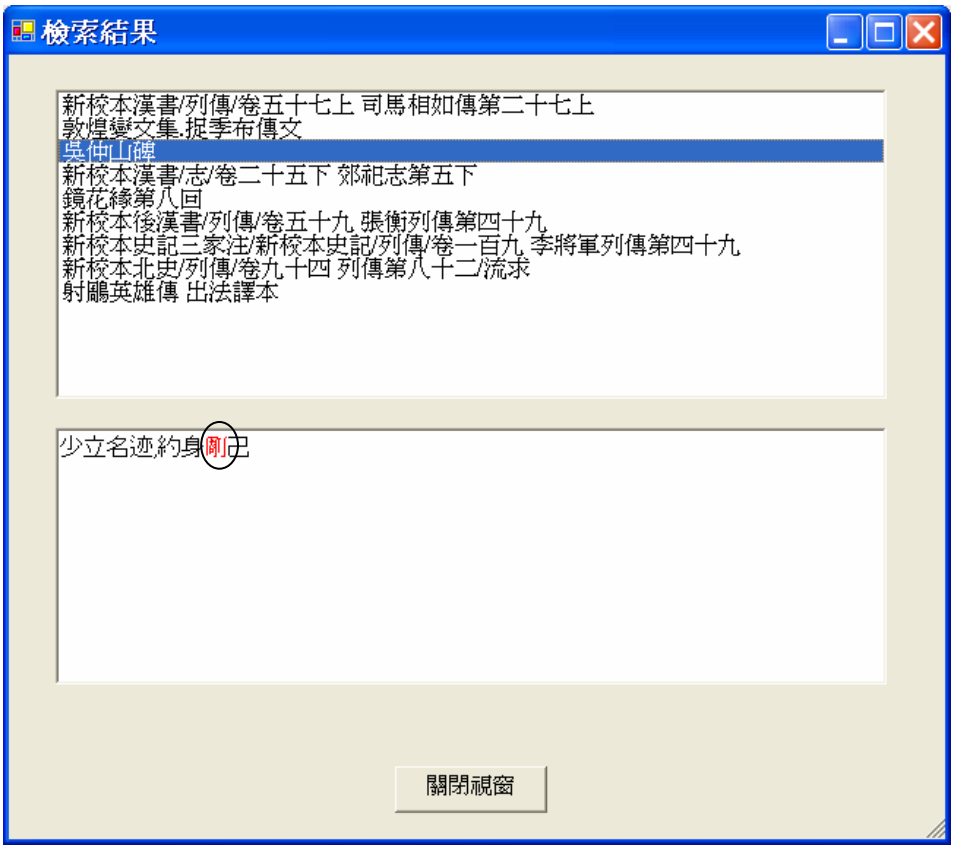
表一　檢索雕的異體字文件集合

編號	來源	標題	文獻部份內容
1	中央研究院 廿五史資料庫	新校本史記三家注/新校本史記/列傳/卷一百九　李將軍列傳第四十九	匈奴大入上郡，天子使中貴人從廣[一]勒習兵擊匈奴。中貴人將騎數十縱，[二]見匈奴三人，與戰。三人還射，[三]傷中貴人，殺其騎且盡。中貴人走廣。廣曰：「是必射雕者也。」
2	中央研究院 廿五史資料庫	新校本漢書/志/卷二十五下　郊祀志第五下	則一梁豐鎬之間周舊居也，固宜有宗廟壇場祭祀之臧。今鼎出於一東，中有刻書曰：『王命尸臣：「官此栒邑，[六]賜爾旂鸞黼黻琱戈。」[七]尸臣拜手稽
3	中央研究院 廿五史資料庫	新校本後漢書/列傳/卷五十九　張衡列傳第四十九	願竭力以守義兮，雖貧窮而不改。執雕虎而試象兮，阽焦原而跟止。[五]庶斯奉以周旋兮，要既死而後已。[六]俗遷渝而事化兮，泯規矩之圜方。
4	中央研究院 廿五史資料庫	新校本北史/列傳/卷九十四　列傳第八十二/流求	所居曰波羅檀洞，塹柵三重，環以流水，樹棘爲藩。王所居舍，其大一十六間，琱刻禽獸。多裸鏤樹，似橘而葉密，條纖如髮之下垂
5	中央研究院 廿五史資料庫	新校本隋書/紀/卷四帝紀第四/煬帝下/大業十二年	二月己未，真臘國遣使貢方物。甲子夜，有二大鳥似鵰，飛入大業殿，止于御幄，至明而去。癸亥，[一七]東海賊盧公暹率众萬餘，保于蒼山
6	中央研究院 廿五史資料庫	新校本舊唐書/列傳/卷八十八　列傳第三十八	辭辯縱橫，音旨明暢，高宗深納之。思謙在憲司，每見王公，未嘗行拜禮。或勸之，答曰：「鵰鶚鷹鸇，豈尐禽之偶，奈何設拜以狎之？且耳目之官，固當獨立也。」
7	中央研究院 廿五史資料庫	新校本宋史/志/卷一百四十九　志第一百二/輿服一/五輅	駕六青馬，馬有金面，插鵰羽，鞶纓，攀胸鈴拂，青繡屜，錦包尾。又誕馬二，在輅前，飾同駕馬。餘輅及副輅皆有之。駕士六十四人。金輅色以赤，駕六赤馬
8	中央研究院 廿五史資料庫	新校本元史/列傳/卷一百三十一　列傳第十八/完者都	完者都許以爲副元帥，凡征蠻之事，一以問之。且慮其姦詐莫測，因大獵以耀武，適有一鵰翔空，完者都仰射之，應弦而落，遂大獵，所獲山積，華大悅服
9	漢語大字典	吳仲山碑	少立名迹,約身剛己

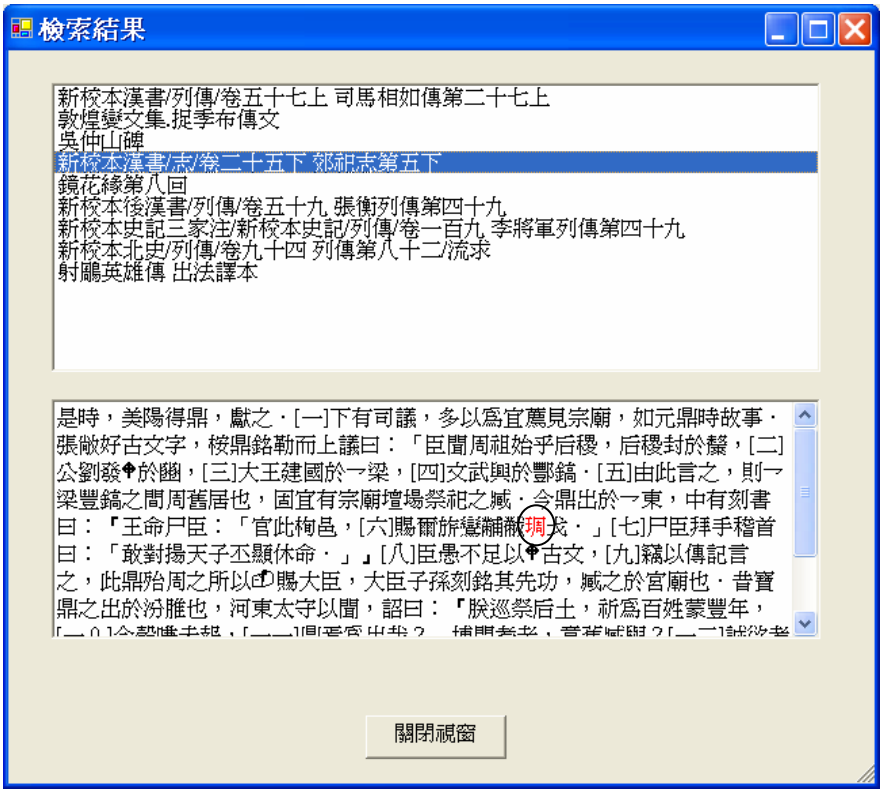
10	漢語大字典	敦煌變文集捉季布傳文	駿馬剚鞍穿鏁甲,旗下依依認得真
11	中央研究院廿五史資料庫	新校本漢書/列傳/卷五十七上　司馬相如傳第二十七上	於是乎乃使剸諸之倫，手格此獸。[一]楚王乃駕馴駼之駟，[二]乘彫玉之輿，[三]靡魚須之橈旃，[四]曳明月之珠旗，[五]建干將之雄戟，[六]左烏號之
12	鏡花緣	鏡花緣第八回	忽見山旁又走出一只小虎，行至山坡，把虎皮揭去，卻是一個美貌少女。身穿白布箭衣，頭上束著白布漁婆巾，臂上跨著一張雕弓。走至大蟲跟前
13	平衡語料庫	平衡語料庫	地域整備法」，其內容乃是想達成一箭三鵰之目的：1依民間活力來擴大國內需求
14	聯合知識庫(20050311)	睽違 12 年 白尾海鵰現蹤	高雄市野鳥學會會員在鳳山水庫發現一隻猛禽類，經與鳥類相關資料比對，確認是台灣罕見的白尾海鵰，鳥會紀錄中，上一次在鳳山水庫發現白尾海鵰，已是12 年前的事
15	聯合知識庫(20041211)	射鵰英雄傳 出法譯本	備受華人愛戴的著名武俠小說作家金庸（查良鏞），其作品「射鵰英雄傳」首次被翻譯爲法文；這部翻譯作品早先已面世。據稱，翻譯者用了三年時間將四冊的「射雕英雄傳」全部翻譯成爲兩冊的法文版



圖四 輸入檢索詞並確認「雕」的意義

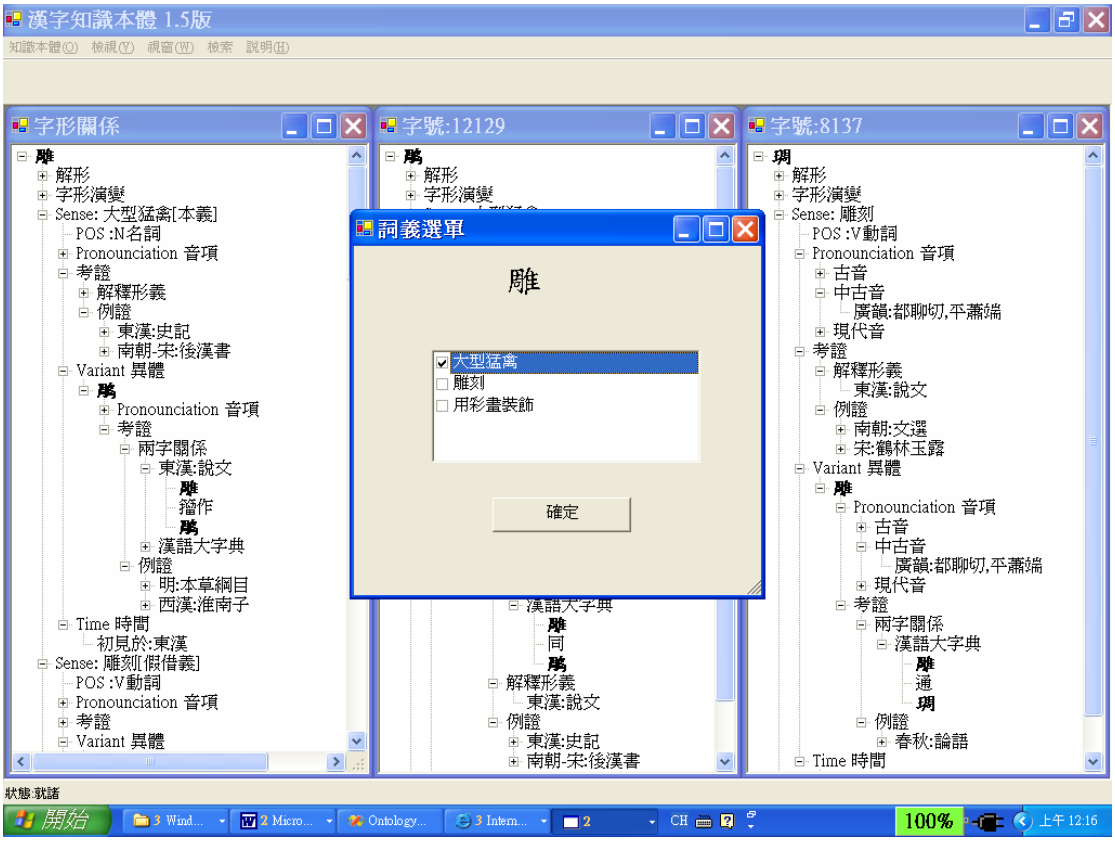


圖五 檢索「雕」(雕刻)可查到部份異體「剛」的文件

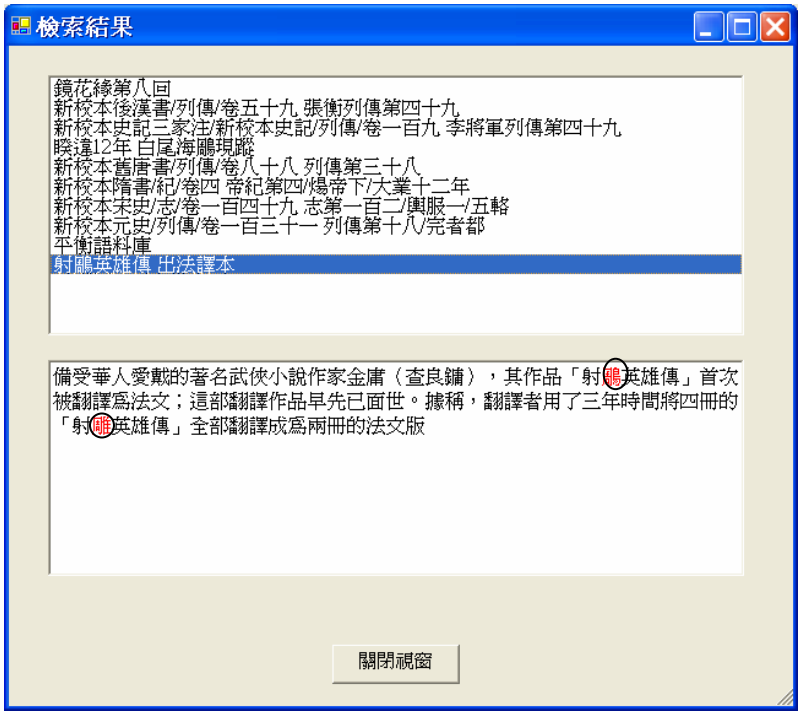


圖六 檢索「雕」(雕刻)可查到部份異體「琱」的文件

如果將檢索的字義條件改爲大型猛禽，異體字中只有「鵰」會被加入檢索詞，而其它的異體字「彫」、「琱」、「剛」和「鋼」都不會被加入，所以檢索的結果沒有出現「彫」、「琱」、「剛」和「鋼」的文件。



圖七　檢索「雕」(大型猛禽)



圖八 檢索「雕」(大型猛禽)可找到出現「鵰」的文件

4.2 歷時異體字檢索

共時異體字檢索並不考慮異體字使用的時間和檢索文件的時間，而歷時異體字檢索則會依據已經建立的異體字關係提供歷代異體字的差異，根據被檢索文件的時間進行檢索詞的修正，例如「獅」本假借「師」，後來才增加意符「犬」造了分化字「獅」以明確其假借義[裘錫圭 1995]，「師」的初義爲軍隊編制單位，且爲最高編成單位[宋子然 2002]，在漢以前就被借假爲「獅」，如漢書：「鉅象、師子、猛犬、大雀之仓食於外囿」，到了北宋新唐書都還是借「師」表示「獅」，而本字「獅」較早出現在玉篇，因此，以「獅子」爲檢索詞，如果被檢索文件是北宋以前的，就應該改以「師子」進行檢索。

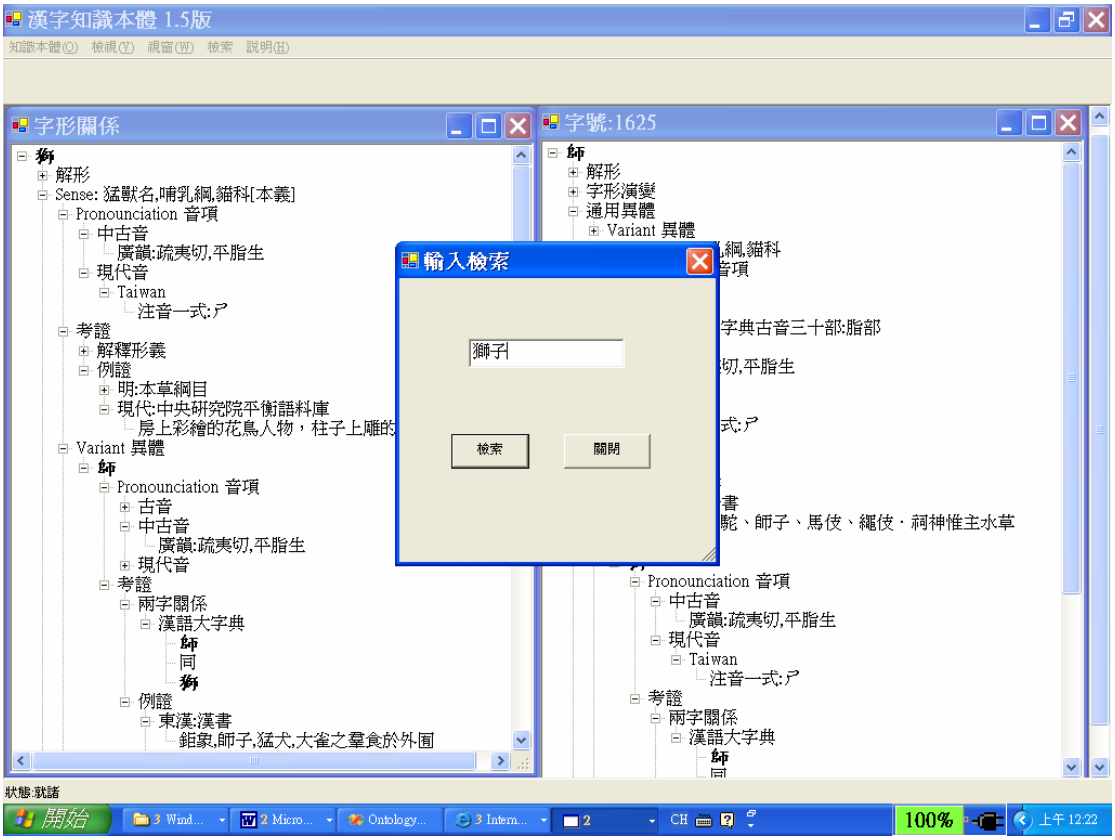
表二 是作爲「獅」的異體字檢索的文件集合，來源是中央研究院漢籍電子文獻的二十五史資料庫、中央研究院平衡語料庫和聯合報聯合知識庫，由於歷時檢索會根據文件的時間和異體字的使用時間決定檢索詞的擴展，因此文件的時間必需被放入計算機，這些文件的時間範圍由東漢至現代。

表二　獅的異體字檢索文件集合

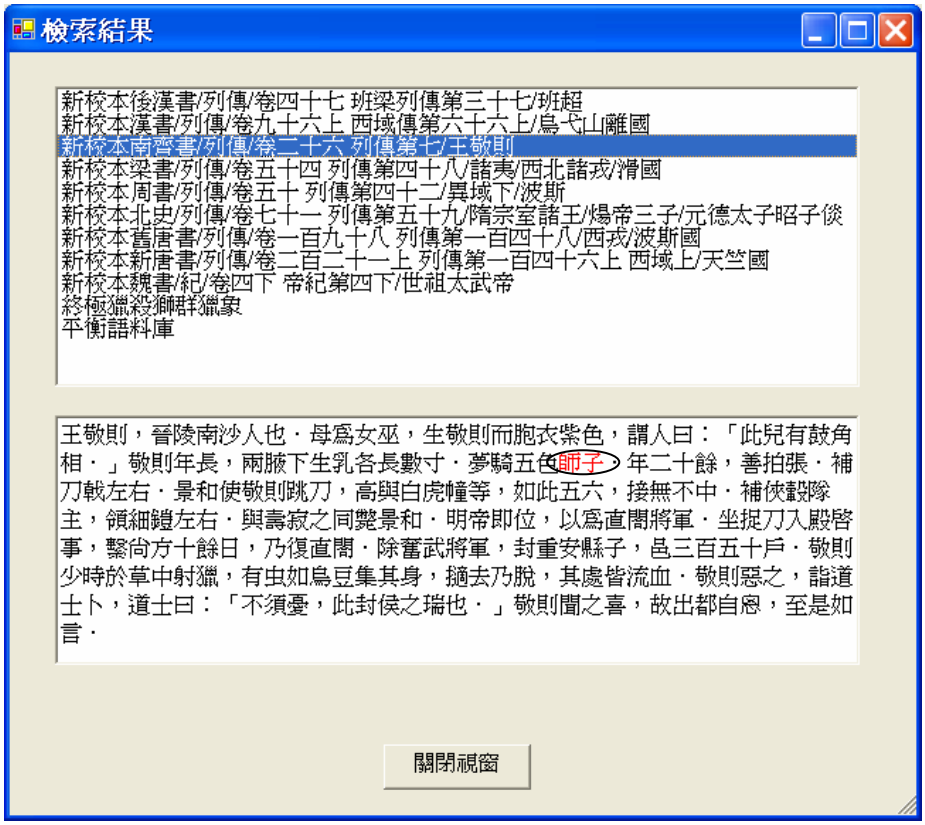
編號	來源	標題	時間	文獻部分內容
1	中央研究院廿五史資料庫	新校本漢書/列傳/卷九十六上 西域傳第六十六上/烏弋山離國	東漢	果菜、食飲、宮室、市列、錢貨、兵器、金珠之屬皆與罽賓同，而有桃拔、師子、犀牛。[二]俗重妄殺。[三]其錢獨文爲人頭。幕爲騎馬。以金
2	中央研究院廿五史資料庫	新校本後漢書/列傳/卷四十七 班梁列傳第三十七/班超	南朝宋	初，月氏嘗助漢擊車師有功，是歲貢奉珍寶、符拔、師子，[一]因求漢公主。超拒還其使，由是怨恨。永元二年，月氏遣其副王謝將兵七萬攻超。超兵少，皆大恐。超譬軍士曰：「月氏兵雖
3	中央研究院廿五史資料庫	新校本魏書/紀/卷四下 帝紀第四下/世祖太武帝	北朝北齊	十有一月辛卯，至于鄒山，劉義隆魯郡太守崔邪利率屬城降。使使者以太牢祀孔子。壬子，次于彭城，遂趨盱眙。頍盾國獻師子一。十有二月丁卯，車駕至淮。詔刈藿葦，汎筏數萬而濟。
4	中央研究院廿五史資料庫	新校本南齊書/列傳/卷二十六 列傳第七/王敬則	南朝齊	王敬則，晉陵南沙人也。母爲女巫，生敬則而胞衣紫色，謂人曰：「此兒有鼓角相。」敬則年長，兩腋下生乳各長數寸。夢騎五色師子。年二十餘，善拍張。補刀戟左右。
5	中央研究院廿五史資料庫	新校本梁書/列傳/卷五十四列傳第四十八/諸夷/西北諸戎/滑國	唐	漢永建元年，八滑從班勇擊北虜有功，勇上八滑爲後部親漢侯。自魏、晉以來，不通中國，至天監十五年，其王厭帶夷栗　始遣使獻方物。普通元年，又遣使獻黃師子、白貂裘、波斯錦等物
6	中央研究院廿五史資料庫	新校本周書/列傳/卷五十 列傳第四十二/異域下/波斯	唐	富室至有數千頭者。又出白象、師子、大鳥卵、珍珠、離珠、頗黎、珊瑚、琥珀、笔璃、馬瑙、水晶、瑟瑟、金、銀、　石、金剛、火齊、鑌鐵、銅、錫、朱沙、水銀、綾、錦、白疊、氍、氈毹
7	中央研究院廿五史資料庫	新校本舊唐書/列傳/卷一百九十八列傳第一百四十八/西戎/波斯國	五代	大驢、師子、白象、珊瑚樹高一二尺、琥珀、車渠、瑪瑙、火珠、玻张、琉璃、無食子、香附子、訶黎勒、胡椒、華撥、石蜜、千年棗、甘露桃
8	中央研究院廿五史資料庫	新校本新唐書/列傳/卷二百二十一上 列傳第一百四十六上 西域上/天竺國	北宋	天竺國，漢身毒國也，或曰摩伽陀，曰婆羅門。去京師九千六百里，都護治所二千八百里，居葱嶺南，幅圓三萬里，分東、西、南、北、中五天竺，皆城邑數百。南天竺瀕海，出師子、豹、龜、
9	中央研究院廿五史資料庫	新校本北史/列傳/卷七十一列傳第五十九/隋宗室諸王/	唐	三歲時，於玄武門弄石師子，文帝與文獻皇后至其所。文帝適患腰痛，舉手馮后，昭因避去，如此者再三。文

		煬帝三子/元德太子昭子倓		帝歎曰：「天生長者，誰復教乎！」由是大奇之。文帝嘗謂曰：「當爲爾娶婦。」應聲而泣。
10	中央研究院廿五史資料庫	新校本宋史/列傳/卷四百八十九列傳第二百四十八/外國五/占城	元	大中祥符三年，國主施離霞離鼻麻底遣使朱浮禮來貢。四年，遣使貢師子，詔畜于苑中。使者留二蠻人以給豢養，上憐其懷土
11	中央研究院平衡語料庫	平衡語料庫	現代	有一隻淘氣的小老鼠，看見獅子睡著了，就從獅子的爪子上，爬到他的鼻子上，把獅子弄醒了
12	聯合知識庫	終極獵殺獅群獵象	現代	由曾獲多項艾美獎的生態影片導演休貝爾夫婦，歷時八年時間追蹤拍攝的紀錄片「終極獵殺」，首度拍攝到獅群獵捕大象的珍貴畫面，也推翻科學界認爲獅子不會攻擊大象的既有觀點

若以「獅子」爲檢索詞，根據本研究建立的異體字關係可以知道自東漢至北宋期間，借「師」表示「獅」，因此，如果被檢索文獻的時間是介於這段期間，檢索詞會擴展爲「師子」，如果不在東漢至北宋期間，只會以「獅子」作爲檢索詞，所以由檢索的結果可以發現北宋以前的文獻都被找出來，北宋以後只有出現「獅子」的文獻被找到，但宋史(元朝)中出現「師子」的文獻卻未被找到，因爲它們不是本研究建立的異體字關係描述「師」作「獅」的時間。



圖九　輸入檢索詞「獅子」



圖十 檢索「獅子」可找到中古漢語文件中的「師子」

5. 結論

本研究的困難在於如何將複雜的異體字關係很有系統的方式加以表達，對於異體字的規範和整理，二千多年來是文字學研究之一，雖然已經累積了很多異體字的知識，但是卻一直無法在計算機表達，我們所提出異體字的模型和建立異體字關係的方法，可以將異體字的關係表達在計算機，不僅能夠表達全同異體字關係，更重要的是能夠表達部份異體字關係，改變過去計算機將異體字視爲全同關係的缺失，與過去異體字的描述方式有顯著的不同，我們分析和表達了三千個異體字，發現本研究提出的異體字描述架構能夠充分的表達異體字的關係。本研究可以表達異體字關係包括古字、今字、正字、俗字、有本字假借、無本字假借等，並且描述異體字在那些詞義、時間、構詞和聲韻可以互相使用。 WordNet 的詞彙關係只建立在詞義，而且沒有考慮語言的變遷，我們不僅分析詞義，還加入時間、構詞和聲韻對異體字關係的影響，同時也要分析字形和字音的變化，如此才能釐清異體字關係，再以系統化的方式描述，因爲需要分析的變數很多，而 WordNet 只由詞類和詞義建立詞彙的關係，本研究的分析更複雜，但是，利用 WordNet 建立的中文詞網不能夠表達漢字的特性，而異體字是非常漢字重要的特性，不過因爲無法利用 WordNet，只能逐字進行分析，因爲需要很長的時間。

異體字是漢語書寫系統的重要特性，不過卻沒有能夠將異體字的關係表達在計算機，造成許多中文資訊處理上的問題，本研究將建立的異體字關係應用在異體字檢索問題，目的是提供檢索系統檢索詞彙的不同形式，但是檢索詞彙在被檢索文件中的詞義，必需要由前後文決定，尤其是多義詞的歧義性，這是資訊檢索的問題，並不是本研究要解決的問題，但是透過本研究的實例，可以呈現異體字關係可能的應用，除此之外，我們還利用本研究成果，設計一個檢索介面，用來檢索 Google 的文件，但是提供異體字的檢索。本研究還可以應用在很多其它問題，例如中文網

域名稱(domain name)和缺字的問題。中文網域名稱目前還沒有完全解決異體字造成無法解析的問題，只能同時註冊多個可能異體字網域名稱，主要原因就是計算機沒有異體字關係的知識，而缺字問題實際上大部份都是缺異體字字形[謝清俊 1996]，如果有異體字字形可以使用，就不一定要使用缺字字形，也不需造字，而造字所產生的交換和檢索問題也同時可以被解決。

本研究只是開始，必定還有其它的異體字關係未能表達，如同 WordNet 仍有許多需要詞彙關係沒有被表達，但是即使如此，運用 WordNet 已經足夠改進很多自然語言處理的問題[Fellbaum 1998][Miller 1995]，我們也開始應用本研究建立的異體字關係，但是還需要更多的應用，才能確認本研究對於異體字關係的描述是否足夠。文字是語言的形式表達，文字整理和分析是非常基礎，因爲它是基礎的研究，其影響非常廣泛，但是願意投入這個基礎研究的很少，希望未來能夠有更多的研究資源投入。

誌謝

感謝吳玲玲教授、謝清俊教授、簡立峰教授、季旭昇教授、高照明教授和何瑁鎧教授給予本研究許多的意見。

參考文獻

1.江藍生、陸尊梧(2004)，簡化字繁體字對照字典，漢語大詞典出版社，第六次印刷。

2.宋子然(2002)，訓詁理論與應用，巴蜀書社，第一版。

3.林樹(1972)，中文電腦基本用字研究，國立交通大學工學院計算與控制學系出版。

4.周亞民(2005)，漢字知識本體－以字爲本的知識結構與其應用示例，國立台灣大學資訊管理學系博士論文。

5.段玉裁(1813)，說文解字注，黎明，1990 印刷。

6.洪成玉(1995)，古今字，北京，語文出版社。

7.洪嘉馡、巫宜靜、黃居仁(2005)，異體字與異體詞詞彙語意初探，第六屆漢語詞彙語意學研討會，廈門。

8.袁國華、曾黎明(2005)，建立 UNICODE 漢字異體字表與異體字辭典之相關研究，數位典藏國家型計劃技術分項計劃，NSC93-2422-H001-018，中央研究院歷史語言研究所。

9.許慎(121)，說文解字，徐鉉校定，北京，中華書局，2004，第一版，第二十二刷。

10.徐中舒(1992)，漢語大字典，建宏出版社。

11.張如瑩、黃居仁(2004)，中央研究院中英雙語知識本體詞網(Sinica BOW)：結合詞網、知識本體與領域標記的詞彙知識庫，ROCLING XVI: Conference on Computational Linguistics and Speech Processing, 台北，9 月 2-3 日。

12.莊德明(1999)，漢字印刷字形的整理，電子古籍中的文字問題研討會，臺北， 6 月 14-16 日。

13.莊德明、謝清俊(2005)，漢字構形資料庫的建置與應用，漢字與全球化國學術研討會，台北，1 月 28-30 日。

14.章瓊(2004)，現代漢語通用字對應異體字整理，巴蜀書社，第一版。

15.董同龢(1979)，漢語音韻學，文史哲出版社，第七版。

16.董琨(1993)，漢字發展史話，台灣商務。

17.裘錫圭(1995)，文字學概要，臺北，萬卷樓，4 月再版。
18.謝清俊(1996)，從缺字問題談漢字交換碼的重新設計－漢字的字形與編碼，漢字，字碼與資料庫國際研討會，京都，東京，10 月 4 日(修正版 1996 年 12 月 20 日)
19.謝清俊、黃克東(1989)，國字整理小組十年，十二月。
20. Coulmas, F.(2003), *Writing Systems:An Introduction to their Linguistics Analysis*, Cambridge University Press.
21.Fellbaum, C.(1998), *WordNet:an Electronic Lexical Database*, The MIT press.
22.Miller, G. A.(1991) *The Science of Words*, Scientific American Library.
23.Miller, G. A.(1995)"WordNet:A Lexical Database for English", *Communications of the ACM*, Vol.38, No.11, Nov., pp.39-41.

台語變調系統實作研究

A Study on Implementation of Taiwanese Tone Sandhi System

楊允言 Iû° Ún-giân[1],李盛安 Li Sheng-an[2],劉杰岳 Lâu Kiãt-gãk[3],高成炎 Kao Cheng-yan[4]
國立台灣大學資訊工程系 台北 台灣
d93001[1] d93005[2] cykao[4] @csie.ntu.edu.tw kiatgak[3]@gmail.com

摘要

台語羅馬字在過去近兩百年來，累積了數量相當可觀的文本，然而目前能流利閱讀台語羅馬字者並不多，使這些資料的利用價值大大降低。

本文主要處理台語的變調問題，實作出台語變調系統。我們採用台語羅馬字書寫的台文語料，以句子爲單位，透過台華對譯辭典找出中文翻譯，再從中研院資訊所詞庫小組的八萬目辭典中取得詞類訊息，接著利用我們訂出的變調規則，標記出每個音節的變調註記。台語變調情形有很多種，文中也有較詳盡的敘述。

研究結果顯示，訓練語料得到 97.56%的變調正確率，測試語料則有 88.90%的變調正確率。我們討論了錯誤的原因，希望持續做改進，以達到更高的正確率。

關鍵詞 台語文 Written Taiwanese、變調規則 Tone Sandhi Rule、台語羅馬字 Taiwanese Romanization

1 背景說明

在台灣，台語是日常生活中常被使用的語言，而台語書面語則較不常見；雖然如此，台語書面語已有百年以上的歷史。[Tiuⁿ 2001]目前台灣社會則存在數十、甚至超過百套的台語書寫系統。[Iû° & Tiuⁿ 1999]本文採用的書寫系統，爲台語羅馬字（又稱爲白話字、教會羅馬字）。

根據國家台灣文學館籌備處委託成功大學台灣文學系所執行的「台灣白話字文學資料蒐集整理計畫」，[1]雖然歷經政治變局，許多資料已遺失，但在該計畫的努力下，仍蒐集到近兩千種台語羅馬字相關書刊，出版地也遍及台灣、廈門、上海、廣州、香港、新加坡、菲律賓、倫敦、日本、...等地；根據蒐集到的出版品其出版年代的統計，1950、1960 年代是相關書刊在台灣出版的高峰。除了正式出版的書刊，也有一般民眾的書信、醫師所寫的患者病歷資料等使用台語羅馬字。然而後來政府以阻礙國語推行爲由強加禁止，導致台語羅馬字的急速式微。

上列計畫成果，我們希望透過資訊科技，讓更多人能夠運用這些資料，促成台語文的基礎或

[1] 計畫主持人爲呂興昌教授，執行期間是 2001.5~2004.12。

應用研究。鑑於一般人對於台語羅馬字並不熟悉，如果利用語音合成技術，將這些文字資料轉成語音，可以讓這些寶貴的資料獲得更高的使用價值。

不過台語的文字要轉成聲音，最大的挑戰在於台語的變調問題。台語羅馬字表記本調，實際發音時，在語詞的層次，大多數情形是最後一音節讀本調，其餘讀變調。然而在句子的層次，大部分情形是在詞組或是結構標誌的分界處的最後一音節讀本調，其餘讀變調（包含語詞的最後一音節也讀變調）。實際上，除了規則變調外，變調又有好幾種情形，本文將一一討論。

本文將重點放在變調規則的處理，這將是其後輸出的語音、聲調是否正確的主要關鍵。本文主要採用上列計畫所蒐集到的語料做爲輸入，實作台語變調系統，輸出爲輸入的語料加上變調註記。由於沒有正確變調結果的訓練語料集，我們暫時不採用統計學習的方法，而是用規則式學習，並由本文其中兩位長久參與台語文工作的作者，來評估變調結果正確率。[2]

2 台語變調說明

台語的聲調，依據傳統的說法，平、上、去、入分陰陽，但上聲不分，所以共有七個聲調，根據陰平、上、陰去、陰入、陽平、陽去、陽入的順序，[3]用數字表示分別是 1(高平)、2(高降)、3(低)、4(中短)、5(低升)、7(中平)、8(高短)。刮號中描述調值。聲調符號請參酌下面的例子。

變調是台語非常重要的特色。在語詞層次，通常最後一音節讀本調，其餘讀變調。下例中的五個語詞，畫底線者讀本調，其餘則讀變調：[4]

例 1　tâi 台　/　Tâi-gí(gú)台語　/　Tâi-gí(gú)-bûn 台語文
Tâi-gí(gú) bûn-hāk 台語文學　/　Tâi-gí(gú) bûn-hāk-sú 台語文學史

實際上，在音節或語詞的層次，台語變調至少包括下列幾種：

(1) 規則變調：以疊詞做說明，刮號內的數字爲實際讀出的聲調。

例 2　(i)　1 聲→7 聲：如「chheng-chheng 清清」(7,1)
(ii)　7 聲→3 聲：如「chëng-chëng 靜靜」(3,7)
(iii)　3 聲→2 聲：如「chhiò-chhiò 笑笑」(2,3)
(iv)　2 聲→1 聲：如「léng-léng 冷冷」(1,2)
(v)　5 聲→7 聲或 3 聲(台北)：如「âng-âng 紅紅」(7/3,5)
(vi)　4 聲→8 聲(-p/t/k)或 2 聲(-h)：如「sip-sip 濕濕」(8,4)「phah-phah 打打」(2,4)
(vii)　8 聲→4 聲(-p/t/k)或 3 聲(-h)：如「直直 t…t-t…t」(4,8)「熱熱 jõah-jõah」(3,8)

(2) 隨前變調：一般爲代名詞或人名的後綴，前面一音節讀本調，此音節的聲調視前面聲調而定，爲 1 或 3 或 7 聲。

例 3　(i)　「A-eng--a 阿瑛 a」(7,1,1) (第二個"a" 是後綴)
(ii)　「góa lâi khòa° -- i 我來看伊」(1,7/3,3,3) (「i 伊」原來是第 1 聲)
(iii)　「h³--lí [給]你」(7,7)(「lí 你」原來是第 2 聲)

[2] 本文第一作者從事台語文工作將近 20 年，第三作者將近 10 年，除了發表技術性論文及相關軟體開發，也有台語刊物編輯、台語文章創作等各方面相關的經歷。
[3] 這樣的順序，方便與漢語系其它語言（如客語等）做對應。
[4] 「台語文學」和「台語文學史」是一個詞還是兩個詞（「台語 / 文學」、「台語 / 文學史」）也許仍值得商榷，在這裡我們暫時視爲一個詞。

(3) 輕聲：輕聲前讀本調，輕聲的部分讀 3 聲或 4 聲(入聲)。

例 4 (i) 「Tân--sian-si°(sin-se°)陳先生」(5,3,3)(「sian-si°(sin-se°)先生」原來聲調是 7,1)
(ii) 「kiâ°--chhut-lâi 行出來」(5,4,3) (「chhut-lâi 出來」原來聲調是 8,5)

(4) 再變調：多半出現在喉塞音(-h) 4 聲，規則變調兩次(4→2→1)。

例 5 (i) 「beh thăk-chu[要]讀書」(1,4,1) (「beh[要]」4 聲應變 2 聲，實際變 1 聲)
(ii) 「khì göa-kháu 去外口」(1,3,2) (「khì 去」3 聲應變 2 聲，實際變 1 聲)

(5) á[仔]前變調：á 前的音節，只有 1、2 聲同規則變調，其餘不同。

例 6 (i) 1 聲→7 聲：如「sun-á 孫仔」(7,2)
(ii) 2 聲→1 聲：如「chháu-á 草仔」(1,2)
(iii) 3 聲→1 聲：如「tà°-á 擔仔」(1,2)
(iv) 4 聲→8 聲(-p/t/k) 或 1 聲(-h)：如「tek-á 竹仔」(8,2)「thih-á 鐵仔」(1,2)
(v) 5 聲→7 聲：如「l²-á 爐仔」(7,2)
(vi) 7 聲→7 聲：如「ph³-á 簿仔」(7,2)
(vii) 8 聲→4 聲(-p/t/k)或 7 聲(-h)：如「chhăt-á 賊仔」(4,2)「hiõh-á 葉仔」(7,2)

(6) 三連音變調：三連音疊詞的第 1 音節，2、3、4 聲同規則變調，其餘不同。

例 7 (i) 1 聲→5 聲：如「chheng-chheng-chheng 清清清」(5,7,1)
(ii) 2 聲→1 聲：如「ún-ún-ún 穩穩穩」(1,1,2)
(iii) 3 聲→2 聲：如「hèng-hèng-hèng 興興興」(2,2,3)
(iv) 4 聲→8 聲(-p/t/k)或 2 聲(-h)：如「sip-sip-sip 濕濕濕」(8,8,4)
「bah-bah-bah 肉肉肉」(2,2,4)
(v) 5 聲→5 聲：如「kôa°-kôa°-kôa°寒寒寒」(5,7/3,5)
(vi) 7 聲→5 聲：如「chëng-chëng-chëng 靜靜靜」(5,3,7)
(vii) 8 聲→5 聲：如「t…t-t…t-t…t 直直直」(5,4,8)「pçh-pçh-pçh 白白白」(5,3,8)

(7) 升調：通常發生在日語借詞，變調後是 5 聲。

例 8 「ŏai-siak-chù [白襯衫]」(5,8,3)「khăn-páng[看板]」(5,2)「hăn-t‾-lù [方向盤]」(5,1,3)[5]

3 文獻探討

在台灣，從事台語文計算語言學的研究團隊，包括長庚大學資訊系呂仁園教授主持的多媒體訊號處理實驗室、[6]台大資訊系陳信希教授主持的自然語言處理實驗室、交大電信系陳信宏教授主持的語音處理實驗室、成大電機系王駿發教授主持的多媒體通訊 IC 系統設計實驗室、成大資訊系吳宗憲教授主持的多媒體人機通訊實驗室、台大資訊系高成炎教授主持的台語文研究室、...等。以下探討有提出變調正確率的兩篇論文。

[Lîm 1997]是較早實作的台語變調系統，輸入是中文，輸出是台語文及發音，語料爲中文新聞資料，利用中研院資訊所詞庫小組的斷詞、標記結果，以及鄭良偉提供的台華對譯辭典，用資料庫查詢華文所對應的台文，台文有漢字及台語羅馬字。變調規則用到： a)句尾讀本調 b) "ê[的]" 前讀本調 c) 名詞詞尾讀本調 d)其餘規則變調。其變調的正確率有 82.53%。不過系統並非使用台語文做爲輸入，將中文轉成台文時，語詞排列順序及詞義排歧並沒有處理，翻譯出來的台語文，

[5]正式的台語羅馬字聲調符號並不包括 ŏ ă，我們參酌[Tiuⁿ 2001]採用此符號。
[6] 清大統計所江永進教授也加入此研究團隊。

與實際 Native Speaker 所講的台語有些差距。儘管如此，仍然是先驅性的台語變調實作系統。

[Liang et. 2004] 是最近發表的台語語音合成實作系統，輸入爲大量的中文新聞語料，去除多於 20 個音節的句子，利用辭典轉成台文後，經過斷詞、標記發音、做變調規則處理後轉成聲音檔並播放。因爲資料量大，所以挑選前兩百句，由兩位台語專家做正確率的評估，結果是：斷詞正確率超過 97%，標記發音有 89%的正確率，變調規則處理則有 65%的正確率。

本文所實作的系統與上述論文所提的系統，主要的不同點在於，我們採用台語文語料，文學類與非文學類大致各半，沒有觸及中文翻譯成台文的問題，且任何長度的句子都處理；另外，因爲使用台語羅馬字，也少了斷詞及標記發音的問題。不過，與漢字相比，發音相同的歧異語詞數大大增加，特別是單音節語詞，這是較大的挑戰。

4 研究步驟

4.1 語料

本文所採用的台語文語料以台語羅馬字書寫，台語羅馬字的書寫形式是以詞爲單位，同一個語詞的音節以連字符(hyphen)連接，語詞和語詞間以空白間隔。

語料由上列計畫提供。訓練語料部分，我們挑選四本書，分別是：

- 1913 年出版的《Sin-bûn ê chãp-liõk 新聞的雜錄》(不知作者，類別：報導)；[7]
- 1924 年出版的《Chãp-häng kóan-kiàn 十項管見》(作者: 蔡培火, 類別: 論述)；[8]
- 1955 年出版的《Chháu-tui téng ê bîn-bäng -- jî-tông chong-kàu k±-sü 草堆頂的眠夢—兒童宗教故事》(作者：黃懷恩，類別：小說) ；
- 1961 年出版的《Tang-p³ thôan-tö kiàn-bûn kì 東部傳道見聞記》(作者：陳降祥，類別：報導)[9]

上述語料涵蓋日本時代和國民政府時代，每本書挑出兩段，總共 614 音節。

測試語料除了上列計畫所提供的之外，部分來源爲我們蒐集的台語羅馬字語料。同樣挑選四份資料，分別是：

- 1885 年出版的《Pçh-öe-j,, ê l,,-ek 白話字的利益》(作者：葉牧師，類別：論述)[10]
- 1905 年出版的《Kau-chiàn ê Siau-sit 交戰的消息》(作者：教會公報編輯室，類別：報導)[11]

---

[7] 雖然的報導類，不過其寫作風格與現在一般的報導文章不盡相同，我們挑選的兩段，都是以第一人稱在敘述，看起來很像一般的散文。

[8] 蔡培火是日本時代台灣政治社會運動的重要人物之一，文化協會成立後，他曾寫文章、演講、辦夏季學校來鼓吹台語羅馬字的使用，可惜其理念不見容於台灣總督府。戰後，蔡培火加入國民黨，成爲籠絡台灣人的象徵，位高卻無權。有興趣者可參考台灣史料基金會出版的《蔡培火全集》http://www.twcenter.org.tw/a02/a02_08/a02_08_01.htm。

[9] 對整個台灣來說，東部是相對特殊的地方，漢人比例少，清國時代只有領台最後 20 年才真正管理東部，日本時代有計畫地移入大量日本人；因爲邊陲，對於東部的（漢字、日文）文字論述多只有官方立場，反而台語羅馬字書寫的相關東部消息，提供比較貼近尋常百姓的觀點。

[10] 這篇文章在談論台語羅馬字和「孔子字」（漢文）的優缺點，是篇擲地有聲的論述，出處爲台灣府城教會報（目前的名稱是台灣教會公報）。

[11] 這篇文章在談日俄戰爭，出處也是台灣教會公報。討論日本時代的台灣政治社會運動，會提及台灣人所辦的第一份刊物是 1920 年的《台灣青年》，當時還無法在台灣發行。如果也把台語羅馬字文獻也考慮進來，這些政治社會運動的論述

- 1954 年出版的《Thià° lí iâ° kè thong sè-kan 疼你贏過通世間》(作者：賴仁聲，類別：小說)[12]
- 1997 年發表在 BBS 的《Ài lí kap ài i pî°-á chöe 愛妳及愛伊平仔多》(作者：盧誕春，類別：散文)

同樣是涵蓋兩個時代，年代的範圍更廣。

4.2 詞類標記

詞類標記的部分，由於目前尚未有台語詞類的標準，我們只好暫時借用華語的成果。我們將語料，以語詞爲單位，查詢台文華文線上辭典（有台語羅馬字、台語漢羅寫法、華語對譯、查詢頻率、...等欄位），對應到華語的詞彙後，再從中研院詞庫小組的八萬詞目資料中找到此語詞的詞類標記。這裡會遇到歧義(ambiguity)問題，包括：

(a) 同音詞，特別是單音節同音詞；
(b) 台語對華語的翻譯是一對多；
(c) 華語語詞本身有多重詞類。

同音詞的問題，我們只取查詢頻率最高的，[13]實際檢視資料發現，在大部分情形下是對的；因爲一個台語詞可能對應到多個華語詞，而且一個華語詞本身可能有多重詞類，因此一個台語詞可能會有多重詞類，目前在詞類標記階段我們皆加以保留，留待變調標記階段取捨。八萬詞目的詞類標記，是簡化後的詞類標記，有 46 個詞類，實際上我們只取一層，其中某些詞類做了些調整（第二層的訊息、且會影響變調結果的詞類），如將 Vh（狀態不及物動詞、狀態使動動詞）改成 A（形容詞），Nh（代名詞）改成 R，Ng（後置詞）改成 G，Nd(時間詞)改成 S。

至於未知詞，如果是「XX」或「XXX」（音節重覆）的形式，我們暫時標記成 A(形容詞)，其餘標記爲 N(名詞)。

所以，我們所用的詞類標記包括：A 形容詞、C 連接詞、D 副詞、G 後置詞、I 感嘆詞、M 特別標記、N 名詞、P 介詞、R 代名詞、V 動詞、S 時間詞、T 語助詞等 12 個標記。

4.3 變調規則

變調規則是本研究最重要的部分。目前的變調規則演算法請參考表 1：

表 1 變調規則演算法

1 變調註記 lóng 先填 t (規則變調)
2 Siöng 尾一個改做本調
3 (語詞層次)"ê" ê 處理：kä 頭前 ê 詞尾改做 #(本調)

也許有需要改寫。

[12] 賴仁聲在 1920 年代曾出版兩本台語羅馬字的小說《A-niâ ê bāk-sái 阿娘的目屎》及《Khó-ài ê siû-jîn 可愛的仇人》。1950 年代出版的這本書是 2002 年才「出土」的。台語文學史上應有其重大意義。

[13] 一般的詞頻是指某語詞出現在語料中的頻率，這裡的查詢頻率指的是此語詞在線上辭典被使用者查詢到的次數，使用者查詢方式，包括利用台語羅馬字、台語漢字以及華語等三種。

4 (詞類層次) chhë A/A Pair (無 ambiguity ê 情形)

4.1 chhë A / A Pair (頭前 A、後壁 A) ：kä 頭前 A ê 詞尾改做 #(本調)

5 (詞類層次) chhë N/V N/A N/P N/R N/D Pair (無 ambiguity ê 情形)

5.1 Chhë N / V Pair (頭前 N、後壁 V)：kä N ê 詞尾改做 #(本調)

5.2 Chhë N / A Pair (頭前 N、後壁 A) ：kä N ê 詞尾改做 #(本調)

5.3 Chhë N / P Pair (頭前 N、後壁 P) ：kä 頭前 N ê 詞尾改做 #(本調)

5.4 Chhë N / R Pair (頭前 N、後壁 R) ：kä 頭前 N ê 詞尾改做 #(本調)

Chhë N / D Pair (頭前 N、後壁 D) ：kä 頭前 N ê 詞尾改做 #(本調)

6 (詞類層次)C(連接詞) ê 處理

詞類是 C，kä 頭前 hit 個詞 ê 詞尾改做 #(本調)

7 (詞類層次) G(後置詞) ê 處理

詞類是 G，kä 頭前 hit 個詞 ê 詞尾改做 #(本調)，G 這個語詞 ê 詞尾 mä 改做#(本調)

8 (詞類層次) S(時間詞) ê 處理

詞類是 S，kä 這個語詞 ê 詞尾改做#(本調)

9 (語詞層次) R(代名詞) "góa / lí / i / gún|góan / lán / lín / in" ê 處理(愛 t,, 第三條後壁)

9.1 語詞是"i / in"包括 t,,句尾：變調註記改做 "t"(規則變調) (這條以後需要進一步討論)

9.2 語詞是"góa / lí / gún|góan / lán / lín "而且無 t,,句尾：變調註記改做 "t"(規則變調)

10 (語詞層次)句尾 "kóng[講]" ê 處理：假使 Delimeter 是[，, ：:"] ，而且"kóng" ê 頭前 ê 語詞（ùi 頭前一直檢查 kàu 句首）有發現詞類是 R(代名詞) ê，變調註記改做 "t"(這個規則以後需要進一步修改，處理人名(Unknown Word)出現 t,,頭前 ê 情形)

11 (音節層次)á 前變調處理

所有 ê 內容，若有包含"á"而且˜是 t,,詞頭 ê，將 "á"頭前 hit 個音節的變調註記改做 & (á 前變調)

12 再變調處理

12.1 (音節層次)"beh" ê 再變調處理：假使"beh" ˜是出現 t,,句尾，變調註記改做 $(再變調) (包括 t,,詞內底 ê "beh",親像 "強 beh/ tih-beh/愛 beh" ...)

12.2 (語詞層次) "khì[去]" ê 再變調處理："khì[去]"若˜是出現 t,,詞尾，伊 ê 後壁 ê POS 若是 N iah 是 V，tö 標記$(再變調) (這個規則以後可能需要進一步修改)

12.3 (音節層次)"koh" ê 再變調處理：假使"koh" ˜是出現 t,,句尾，變調註記改做 $(再變調) (包括 t,,詞內底 ê "koh",親像 "koh 再/ chiah-koh/iáu-koh" ...)tö 標記「再變調」(這個規則以後需要進一步修改)

12.4 (語詞層次)"kah" ê 再變調處理：假使"kah" ˜是出現 t,,句尾，變調註記改做 $(再變調)

13 (語詞層次)輕聲處理：若是語詞內底出現"--" ê 所在，kä "--"頭前 ê 第一個音節標記本調，"--"後所有 ê 音節 lóng 標記做輕聲

14 (語詞層次)三連音處理：假使一個語詞 lóng 總三個音節，而且三個音節 lóng käng 款，第一個標記做"~"

15 (語詞層次) 特殊詞處理

15.1 句內若出現"sím-mih / sím-m…h " chia ê 語詞，kä 語詞改做 "sím-mí"，變調註記 mài 修改。

15.2 句內若出現"án-ni /àn-ni"，kä 語詞改做"án-ni"，變調註記是 t#；若出現"an-ni / an-n,,"，kä 語詞改做"án-ni"，變調註記是 t#。

16 句內 ê Marker 處理

16.1 (語詞層次)假使句內有 "iah-s,, / ah-s,, / iăh-s,, / ăh-s,, /á-s,,"，頭前 hit 個語詞 ê 詞尾 ê 變調註記改做#(本調)

16.2 (句型層次)"s,,[是]" ê 處理：假使句內出現"s,,[是]"、而且頭前 hit 個語詞 ê 詞類是 V(動詞)、而且頭前 hit 個詞 t,, "s,,"後壁（一直 kàu 句尾）koh 有出現，將頭前 hit 個語詞 ê 詞尾 ê 變調註記改做#(本調)

16.3 (語詞層次)假使句內有 "che / he / chia / hia"，變調註記改做#(本調)

16.4 (語詞層次)假使句內有"ü-sî[有時]/put-sî[不時]/ kui-khì[kui 氣] / óan-jiân[宛然] / gôan-lâi[原來] chiong-lâi[將來] / chiông-lâi[從來] / sui-jiân[雖然] / sui-bóng[雖罔] / sî-siông[時常] / hui-siông[非常] / s…t-chäi[實在] / s,,-chün[時陣]"，這個語詞 ê 詞尾 ê 變調註記改做#(本調)

16.5 (語詞層次)假使句內有"chiü tö[就] "，而且頭前語詞 ê 詞類是 A(形容詞)，頭前語詞 ê 詞尾 ê 變調註記改做#(本調)

16.6 (語詞層次)假使句內有"sî-kàu[時 kàu]"，這個語詞 ê 兩個音節 ê 變調註記 lóng 改做#(本調)

17 (詞類層次) T(語助詞)處理：若是 siöng 尾 ê 詞類是 T(語助詞)，將頭前 hit 個語詞 ê 詞尾變調註記改做#(本調)

18 其它變調 ê 處理：

18.1 (語詞層次)"teh[在]" ê 處理：語詞若是"teh / t,,-teh"，kä "teh" ê 變調註記改做$(以後改做^)(其它變調，暫時 kah 再變調 käng 款)

19 (語詞層次)輕聲 ê 處理：

19.1 (語詞層次)句尾若有"chhut-khì[出去] chhut-lâi[出來] lõh-lâi[落來] lõh-khì[落去] kòe-lâi/kè-lâi[過來] kòe-khì/kè-khì[過去]" 而且頭前 hit 個語詞 ê 詞類是 V，請將頭前 hit 個詞 ê 詞尾 ê 變調註記改做#(本調)，句尾這個詞 ê 所有音節 ê 變調註記 lóng 改做%(輕聲)

19.2 句內若"sian-si°/sin-se°/sian-se°[先生]"˜是出現 t,,第一個語詞，而且頭前一音節是單音節、第一字母大寫，請將頭前音節變調註記改做#(本調)，這個語詞 ê 所有音節 ê 變調註記 lóng 改做%(輕聲)

19.3 (語詞層次)句尾 ê "bô[無]" (是˜是 beh 輕聲 ê 處理)

19.3.1 頭前一個詞若是"á / á-s,, / iah / iah-s,, / ah /ah-s,, [或是]"，無需要修改(維持原來，讀本調)

19.3.2 其它 ê 情形，頭前 ê 語詞詞尾修改做#(本調)，"bô[無]"改做%(輕聲) (部分會錯誤)

19.4 (語詞層次)句尾 ê "bë/böe[不會]" (是˜是 beh 輕聲 ê 處理)

19.4.1 句內若有出現"ë/öe[會] ë-hiáu/öe-hiáu[會曉]"，修改做%(輕聲)(t,, 19.5.2 進前做)

19.4.2 頭前一個詞若是"á / á-s,, / iah / iah-s,, / ah /ah-s,, [或是]"，改做#(本調)

19.4.3 (其它 ê mài 修改，因爲有可能是 ambiguity(賣))

20 (語詞層次) R(代名詞) "góa / lí / i / gún|góan / lán / lín / in" t,,句尾 ê 隨前變調處理(愛 t,,第 9 條以後做)

假使這個代名詞 t,,句尾，而且頭前 ê 語詞是動詞（只要有出現 tö 會 sái），變調註記改做'@'(隨前變調)

我們利用下列資源建立變調規則演算法，包括：

(a) 語言學家整理的台語變調規則；

(b) 從訓練語料歸納出的規則；

(c) 我們本身對台語變調規則的理解；

(d) 中研院資訊所詞庫小組的中文斷詞系統（參考其詞類標記結果）；

(e) 台語文語詞檢索系統（看台語某些語詞的變調情形）。

值得一提的是，語言學家整理的台語變調規則，有的只針對某些狀況處理而非全體適用，有的規則存在許多例外情形，這些因素導致實作上的困難。因此，語言學家整理的台語變調規則之外，本文其中兩位作者，兼具資訊背景及台語文背景，我們對台語變調規則有所了解，針對訓練語料及現有變調規則演算法得出錯誤變調註記的部分，進行錯誤分析，進一步來補充變調規則。補充變調規則時，也根據經驗，考慮到其它沒有在訓練語料中出現，但是相關的變調規則，也一併補充進來。訂定變調規則時，以適用大部分情形爲原則（例如可以讓語料庫中80%以上正確），不要求完全正確。而新規則加入後，可能影響部分原來的規則，於是，詞庫小組的中文斷詞系統及台語文語詞檢索系統成爲我們決定是否加入新規則的重要參考工具。

變調規則處理音節、語詞、詞類、句型等四種層次的變調問題，舉例來說：

(1) 音節層次，例如「koh[又]」、「beh[要]」，不管是否爲語詞的一部份（如「kiông-beh 強[要]」、「koh-chài[又]再」等），都標記爲再變調。

(2) 語詞層次，例如「che[這]」、「he[那]」，不管出現在何處，一律標記爲本調。有的情形則是，某個語詞出現時（如「ê[的]」，）會去改變前面語詞的變調標記。

(3) 詞類層次，例如詞類爲 N(名詞)，之後的詞類若爲 A(形容詞)、D(副詞)、P(介詞)、R(代名詞)或 V(動詞)，則此名詞詞尾音節標記爲本調。有時是某個詞類出現時（如 G 後置詞），也會改變前面語詞的變調標記。

(4) 句型層次，有些語詞出現時（如「iah-s,,[或]是」）此句子此語詞前的部分，可以視爲一個子句；又例如「ë...bë 會...[不會]」的句型出現時（「bë」出現在句尾，句中出現「ë 會」），則將「bë」標記爲輕聲。

有些規則有先後順序，後面的規則可覆蓋原來的規則，如代名詞（「lí 你」、「góa 我」、「i 伊」）的變調處理規則可以覆蓋「ê[的]」的變調處理規則；或是上述句型層次的兩個例子中，第一個規則可以覆蓋第二個規則：

例 9 「Lí ë khì kok-göa bë 你會去國外[不會]」，句尾的「bë」標記爲輕聲
「Lí ë khì kok-göa iah-s,, bë 你會去國外[或]是[不會]」，句尾的「bë」標記爲本調

另外，因爲詞類歧異的情形沒有處理，所以這些規則在詞類層次的部分，有的規則註明要在沒有歧異時才適用，有的規則是只要存在此詞類就適用。

目前我們訂定 20 條變調規則，並打算持續修改及增加。

舉例來說，下列的訓練語料：

例 10 Chhin-chhiü° án-ni lâi kóng , chäi lán Tâi-ôan k,,n-k,,n ch…t-tiap-á-kú ê kang-hu , ài soa° chiü ü soa° , ài hái chiü ü hái , beh jõah chiü ü jõah ,kôa° chiü ü kôa° . S¯-í thang kóng Tâi-ôan s,, ch…t-ê sió Tang-iû° . Lán Tâi-ôan ü chit-khóan thian-jiân ê hó-kéng ,hó khì-häu , chiong-lâi nä-s,, ëng-sim ke lâng ê kang-hu töa-töa lâi chéng-tùn , tek-khak ë chiâ°-chò Tang-iû° ê töa kong-h£g , h³ Tang-iû° ê lâng ch…p-óa lâi hióng-hok an-lõk .

(親像 án-ni 來講，在咱台灣近近一 tiap 仔久 ê 工夫，愛山就有山、愛海就有海；beh 熱就有熱、寒就有寒。所以 thang 講台灣是一個小東洋。咱台灣有這款天然 ê 好景、好氣候，將來若是 koh 用心加人 ê 工夫大大來整頓，的確會成做東洋 ê 大公園，h³ 東洋 ê 人集倚來享福安樂。)[14]

經過詞類標記和變調規則處理後，輸出爲：

例 11 Chhin -chhiü°(D) án-ni#(D;N) lâi(D;V) kóng#(V), chäi(D;A;P;V) lán(R) Tâi-ôan#(N) k,,n-k,,n(A) ch…t-tiap&-á-kú#(N) ê(M) kang-hu#(A;N), ài(D;V) soa°# (N) chiü(D) ü(D;P;V) soa°#(N), ài(D;V) hái#(N) chiü(D) ü(D;P;V) hái#(N), beh$(D) jõah#(A) chiü(D) ü(D;P;V) jõah#(A), kôa°#(A) chiü(D) ü(D;P;V) kôa°#(A). S¯-í(C) thang(D) kóng(V) Tâi-ôan#(N) s,,(D;V) ch…t-**ê#**(N) sió(D;A) Tang-iû°#(N). Lán(R) Tâi-ôan#(N) ü(D;P;V) chit-**khóan#**(D;N) thian-jiân#(A) ê(M) hó-kéng#(N), hó(D;A;C;V) khì-häu#(N), chiong-lâi#(S) nä-s,,(C) ëng-sim#(N) ke(V) lâng#(N) ê(M) kang-hu#(A;N) töa-töa(A) lâi(D;V) chéng-tùn#(V), tek-khak(D) ë(D;V) chiâ°-chò(V) Tang-iû°#(N) ê(M) töa(A;N) kong-h£g#(N), h³(D;P;V) Tang-iû°#(N) ê(M) lâng#(N) ch…p-óa(V) lâi(D;V) hióng-hok#(A) an-lõk#(A).

其中，刮號內爲詞類標記，台語羅馬字之後若沒有符號表示規則變調，標記「#」表示讀本調，「&」表"á"前變調，「$」表再變調，台語羅馬字以粗體加框，表變調規則錯誤的部分（正確應該讀規則變調）。目前所使用的變調註記請參考表 2：

表 2 變調註記

(t)	#	@	%	$	&	~	^
規則變調	本調	隨前變	輕聲	再變調	á 前變調	三連音第一音節	其它變調

[14] 出處爲 1924 年出版的《Chāp-häng kóan-kiàn 十項管見》，作者是蔡培火。
http://iug.csie.dahan.edu.tw/TG/chu/10HKK/10HKK.asp

4.4 評估正確率

如前所述，本文其中兩位作者是台語文專家，我們有能力確認變調結果的正確率。有些句子，不同的兩種變調結果都是可接受的（有的人變其中一種，有的人變另外一種），則系統的輸出，只要符合其中一種都視爲正確。例如「góa ài lí 我愛你」，「góa ài# lí@」（「góa 我」規則變調，「ài 愛」本調，「lí 你」隨前變調）和「góa ài lí#」（「góa 我」、「ài 愛」規則變調，「lí 你」本調）都是正確的，這當中牽涉的語意問題或句子的焦點問題（如上例，句子的重點是動作「ài 愛」還是對象「lí 你」，變調結果應該不同）我們暫不考慮（而有些例子，不牽涉語意和焦點，仍有兩個不同的合法變調結果）。

5 初步研究結果

初步結果請參考表 3。訓練語料共有 614 個音節，經過人工檢查，共有 15 個音節變調標記錯誤，正確率爲 97.23%；測試語料共有 955 個音節，共有 106 個音節變調標記錯誤，正確率爲 88.90%。

表 3 變調標記正確率

	音節數(A)	標記錯誤數(B)	正確率(1-B/A)
訓練語料	614	15	97.56%
測試語料	955	106	88.90%

其中，測試語料中的錯誤，有些是因爲變調規則尚未完整，沒有處理到的，如果把這些變調規則再補上，應該至少可以提高 2.5%的正確率。

6 錯誤分析及相關問題討論

我們希望再做改進，讓變調系統的正確率能夠提高。在此提出我們所遇到的問題。

6.1 台語的詞類

目前我們使用中文的詞類，中文的詞類是不是適合台語，可能需要語言學家給我們答案。[15] 日本時代，1934 年陳輝龍出版《台灣語法》，[16]戰後，1950 年李獻璋在日本出版《福建語法序說》，[17]這些書都有討論台語的詞類；此外，1984 年中華語文研習所出版、Embree 的《台英辭典》，詞條有詞類的資料，這些應該都是可供參考的資料，當然，這些資料需要建立電子檔，還需要語言學家的協助，檢視這些資料對於處理台語變調問題是否合用。

6.2 台語的分詞標準及辭典

[Chan 1997]根據中研院詞庫小組的中文分詞標準，提出了台語分詞標準，可惜沒有引起進一步的討論。如果台語分詞標準可行，我們還需要一部符合分詞標準的辭典，目前還沒有，希望將來可以建置完成。

[15] 鄭良偉曾訂定台語的詞類，這是我們之後要斟酌的方向之一。
[16] 陳輝龍所列出的詞類有名詞、代名詞、數詞及助數詞、形容詞、動詞、助動詞、副詞、前置詞、語尾詞、接續詞、感嘆詞等 11 類。

6.3 書寫的標準化問題

如果是漢字表記的台語文，這個問題十分嚴重。至於台語羅馬字的文獻、語料，腔調基本上沒有問題，已可透過辭典查詢來解決，連字符號(hyphen)的使用則不完全一致，這多少造成一些系統在處理變調問題的麻煩，把一個語詞分開寫（如「chit-ê 這個」寫成「chit ê」）或把兩個語詞連起來寫（如「s¯ siá 所寫」寫成「s¯-siá」）對系統而言，會產生不同的變調標記，其中一個是錯的。然而，動手修改語料並不是好方法。

不過，從資訊處理的角度而言，也許可以將這些問題做較完整的整理，並提出書寫的建議規範。

6.4 詞類無法解決的變調問題

以目前的作法，在檢視變調標記錯誤的地方時，有些可能無法透過詞類的排列順序來決定變調與否，目前看到的，包括並列的動詞及並列的名詞，例如：

例 12　「phah-pià°(V) chò(V) khang-khòe(khè) (N)打拼做空課」(2,2,2,7,3)
「kiãh-bãk(V) khòa°(V) h£g(N)舉目看園」(3,8,2,5)

這是並列動詞的情形，第一個例子中，前面的動詞其詞尾音節需規則變調；而第二個例子則是讀本調，我們無法從詞類中確認該怎麼變調。當然這裡有一些線索，第一例中的「phah-pià° 打拼」若拆成個別音節，「phah 打」和「pià°拼」都是動詞，第二例中的「kiãh-bãk 舉目」拆開的話，「kiãh 舉」是動詞「bãk 目」是名詞，不過這樣的分析，在實作上可能太繁瑣。

例 13　「tiän-chú lêng-kiä°電子零件」
「thâng-thöa chiáu-chiah 蟲豸鳥隻」[18]

這是並列名詞的情形，第一例中，前面的名詞其詞尾音節需規則變調；第二例中則讀本調。目前我們暫時想不出解決的方法。

6.5 錯誤分析

我們檢視錯誤的部分，包括前面已經討論過的，大致有以下幾種情形（刮號中的敘述表可能的解決方式）：

- 因爲辭典沒有此語詞而導致的錯誤；（增加詞條）
- 因爲少了標點符號；[19]
- 因爲同音詞造成的詞類標記錯誤；
- 因爲詞類歧異而無法判別；
- 連字符號書寫問題；（修改語料，或是，部分情形可用程式處理，例如，看到「chit ê 這個」就改成「chit-ê 這個」）
- 變調規則不夠完整；（可以繼續修改變調規則，不過還要考慮 side effect）

---

[17] 李獻璋所列出的詞類有名詞、代名詞、(以上屬實體詞)動詞、助動詞、(以上兩類屬敘說詞)形容詞、副詞、(以上兩類屬限定詞)介詞、連接詞、(以上兩類屬關係詞)助詞(屬情態詞)等 9 類。

[18] 「thâng-thöa 蟲豸」就是昆蟲。

[19] 例如「t,, ke-l³ teh kiâ° ü tú-tiõh ch…t-ê ...[在]街路[在]行有[遇]著一個.....」,「kiâ°行」後面如果有逗點，結果會正確，可是語料中沒有逗點，導致錯誤。

- 定量詞問題；（加上定量詞處理應可解決）
- 專有名詞問題；（加上專有名詞處理應可解決）
- 句型問題；（繼續修改變調規則，不過此部分難度較高）
- ...

另外，當然還可能有不少我們尚未遇到的疑難雜症。

7 未來工作

以目前的成果而言，台語變調規則系統應該還有很長一段路要走，未來的工作包括：

- 藉助語言學家的專長，希望語言學家能討論制訂出台語的詞類、分詞規範，並建立出符合分詞規範的分詞辭典；
- 語料的處理上，我們還需要處理構詞、定量詞、專有名詞等問題；
- 詞類標記的處理上，我們還需要處理詞類排歧(disambiguity)等問題；
- 詞類標記的部分，我們也許還可以考慮直接以 Embree 台英辭典的詞類來做做看，省去對應到中文這個步驟，也許讓詞類歧異降低，增加變調結果正確率。
- 變調處理部分，變調規則的改進當然是必須的；
- 目前的變調規則，算是 bottom-up 的方式，也可考慮運用語法理論，用 top-down 的方式實作看看。鄭良偉教授提出語法模版理論，認爲可以解決翻譯、台語變調等問題。實作上看起來似乎很困難，不過也值得一試。

對 Native Speaker 而言，一個三歲小孩對於台語變調幾乎沒有任何困難。台語變調系統，要如何達到三歲小孩的水準，需要持續的努力。在華文的大環境下，台語的資源很有限，不過，一步一腳印，我們希望這些點點滴滴的成果能夠累積，也期盼更多研究資源及人力的投入。

誌謝

本研究獲得國家台灣文學館籌備處委託計畫「台語文數位典藏資料庫計畫—台語文全羅文字語音輸出系統Tâi-gu-bûn S±-u„-tián-chông Chu-liäu-kh± Kè-öe -- Tâi-gú-bûn Chôan-lô Bûn-j„ Gú-im Su-chhut Hë-thóng」的經費支持，特此致謝。也感謝兩位審查者對本文所提出的建設性改進意見。

參考資料

1. 專書、論文

[Chan 1997] 曾金金，1997，〈台語斷詞原則討論〉，《台語文學出版物收集、目錄、選讀編輯計畫結案報告》p47-73，文建會委託計畫。
http://iug.csie.dahan.edu.tw/TG/CompLing/hunsu/hunsu.htm

[Iû° 2003] 楊允言，2003，《台文華文線上辭典建置技術及使用情形探討》，2003 第三屆全球華文網路教育國際學術研討會論文集 p132-141，台北，圓山大飯店，2003/10/24-26。
http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/Lunbun/THsutian/

THsoaNtengsutian.htm 。

[Iû° & Tiuⁿ 1999] 楊允言、張學謙，〈台灣福佬話非漢字拼音符號的回顧與分析〉，《第一屆台灣母語文化重生與再建學術研討會論文集》p62-76，台南：台南市文化基金會。http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/Lunbun/Huho/huho-0.htm

[Lîm 1997] 林川傑，1997，《國語-閩南語機器翻譯系統之研究》，台北，台灣大學資訊工程系(碩士論文)。

[Liang et. 2004] Min-siong Liang、Jui-Cheng Yang、Yuang-Chin Chiang、Ren-Yuan Lyu，2004，〈*A Taiwanese Text-to-Speech System with Applications to Language Learning*〉，《Proc. of the 4th IEEE Int. Conf. on Advanced Learning Technologies (ICALT'04) 》p91-95，Finland : Joensuu http://msp.csie.cgu.edu.tw/pmwiki.php/PublishMedia/PubPaper1

[L² 1999] 盧廣誠，1999，《臺灣閩南語詞彙研究》，台北，南天。

[Ông et. 1999] 王駿發、黃保章、林順傑，1999，〈國語文句翻台語語音系統之研究〉，《第十二屆計算語言學研討會》p37-53，新竹：交通大學。

[Sia et. 1999] 佘永吉、鍾高基、吳宗憲，1999，〈臺語多音調音節合成單元資料庫暨文字轉語音雛型系統之發展〉，《第十二屆計算語言學研討會》p15-36，新竹：交通大學。

[Tëⁿ 1997] 鄭良偉，1997，《台語、華語的結構及動向 I 台語的語音與詞法》，台北：遠流。

[Tëⁿ 2002] 鄭良偉，2002，〈語法模板上的聲調變化—認知及測驗〉，《2002 台語羅馬字教學及研究國際學術研討會論文集》，台東：台東大學。http://iug.csie.dahan.edu.tw/iug/ungian/POJ/siausit/2002/2002POJGTH/lunbun/K1-Liong-ui.pdf

[Tiuⁿ 2001] 張裕宏，2001，《白話字基本論：台語文對應&相關的議題淺說》，台北：文鶴。(第一章導論：http://iug.csie.dahan.edu.tw/iug/Ungian/patlang/POJkpl/POJkpl01. htm)

2. 網站資料

Taiwanese Package http://www.phahng.idv.tw

中研院資訊所詞庫小組中文斷詞系統 http://ckipsvr.iis.sinica.edu.tw/

台華線上辭典 http://iug.csie.dahan.edu.tw/TG/sutian

台語文語詞檢索系統 http://iug.csie.dahan.edu.tw/TG/concordance

台語羅馬字教學進修網站 http://elearning.lib.nttu.edu.tw/tglmj/index.htm

白話字&萬國碼：字型及軟體開發 http://iug.csie.dahan.edu.tw/TG/Unicode/

白話字書目資料 http://iug.csie.dahan.edu.tw/iug/Ungian/Soannteng/subok/poj.htm

白話字文物展覽 http://www.de-han.org/pehoeji/exhibits/index.htm

長庚大學 多媒體訊號處理實驗室 http://msp.csie.cgu.edu.tw/pmwiki.php

台大資訊系 台語文研究室 http://tb.csie.ntu.edu.tw/

台大資訊系 自然語言處理實驗室 http://nlg3.csie.ntu.edu.tw

交大電信系 語音處理實驗室 http://speech.cm.nctu.edu.tw/

成大資工系 多媒體人機通訊實驗室 http://chinese.csie.ncku.edu.tw/chwu/home.htm

成大電機系 多媒體通訊 IC 系統設計實驗室 http://140.116.156.179/

利用雙語學術名詞庫抽取中英字詞互譯及詞義解歧

白明弘[1,2]、陳克健[1]、張俊盛[2]

1 中央研究院資訊科學研究所

2 國立清華大學資訊工程研究所

mhbai@sinica.edu.tw, kchen@iis.sinica.edu.tw, jschang@cs.nthu.edu.tw

摘要

語意的研究十分依賴語意知識庫所提供的訊息，由於語意研究逐漸變得熱門，相對的語意知識庫的建構也變得十分迫切。WordNet 是目前最廣爲人知的英語語意知識庫，許多語意解歧(word sense disambiguation)的研究都以 WordNet 爲共同標準。由於 WordNet 的成功，使得許多其他語系的 WordNet 建構計畫也紛紛出現。本文提出一個自動從雙語學術名詞庫中抽取中文語意訊息的方法，這個方法利用一個詞和詞的對應(word-to-word alignment)演算法抽取中英詞對譯的訊息，再用語意解歧的方法，將中文詞連結到 WordNet synset，以建構中文 WordNet。

1. 緒論

近年來在自然語言處理領域中，語意研究受到了廣泛的重視。語意解歧的技術不斷推陳出新，進而使得語意的應用也受到鼓舞。然而，語意的使用必需仰賴語意知識庫提供語意訊息，這些訊息包括一個詞彙有多少不同的語意，以及一個語意和另一個語意是否有同義關係或是上下位關係等。例如：「分子」可以表示化學上的「粒子」(如「水分子」)，也可以表示「一群人」(如「激進分子」)；而「拉布拉多」和「大麥町」由其上位詞可知都是一種「狗」。

WordNet 是一部訊息豐富的語意知識庫[Miller 1990]，其中收錄了爲數極多的詞彙。在結構上它將所有的相同的語意集成 synset，並以 synset 爲基礎進一步連結語意之間的關係，如上位關係(hypernym)、下位關係(hyponym)、整體關係(holonyms)及部分關係(meronyms)等。目前 WordNet 已經被應用在許多的研究上，如語意解歧(word sense disambiguation)、資訊檢索(information retrieval) 及電腦輔助語言學習(computer-assisted language learning)等領域，儼然成爲語意研究的共同標準。

由於 WordNet 的成功使得許多其他語系的 WordNet 建構計畫相繼出現。例如：EuroWordNet (EWN)，該計畫目標爲建構包含多種歐洲語的 WordNet，及中文詞網計畫[CKIP 2003]，以建構中文語意知識庫爲目標。從零開始建構一個 WordNet 是一項艱鉅的任務，所以有許多的研究嘗試以自動的方式將詞彙連結到 WordNet。例如：[Atserias et al. 1997]、[Daude et al. 1999]以及[Jason et al. 2003]都是利用雙語詞典所提供的翻譯，自動將詞彙連結到 WordNet。使用一般雙語詞典的翻譯最大的問題在於用詞過度典型化。例如："plant" 在 WordNet 中的第一個語意 "plant, works, industrial plant"，在雙語詞典中翻譯成「工廠」。但實際上在文章中可能翻譯成「廠」、「工廠」、「廠房」、「所」(如「power plant/發電所」)及「工場」等詞。用詞過度典型化的現象，使得許多文章中的用詞無法找到適當的翻譯連結到 WordNet。

在本實驗中，我們選擇以雙語學術名詞庫作爲抽取語意訊息的資料來源。由於學術名詞庫中包含了大量的複合詞，所以很多詞會搭配不同的詞一再出現，並對應到不同的翻譯。因此不但可以避免一般雙語詞典翻譯過度典型化的問題，而且多樣化的翻譯結果可以幫助語意解歧 [Diab et al, 2002][Bhattacharya, 2004]。在本實驗中我們將問題分成兩個部分：a) 如何找出中文詞和英文詞對應的翻譯，b) 如何解決英文的歧義。

本文接下來的章節組織如下。在第 2 節中說明所使用的資源。第 3 節中說明實驗的方法。第 4 節中說明實驗的結果。結論及未來的發展則在第 5 節中說明。

2. 使用資源

本研究使用了兩本詞典作爲語意抽取的對象：

a) 國立編譯館學術名詞詞庫 [NICT, 2004]。

b) 英漢詞典

其中國立編譯館所編輯的「學術名詞」詞庫的內容包含 63 個學科類別共 1,046,058 目詞。這些詞條中有 629,352 目詞是複合詞，佔總詞數的 60 %。英漢詞典共有 208,163 目詞，用來補足「學術名詞」之不足。此外我們使用 WordNet 2.0 做爲語意連結的對象。

由於中文的複合詞在詞和詞之間沒有空白分隔，不像英文詞間以空白字元做爲邊界，所以必需依賴自動斷詞程式將複合詞切分成一般詞。本實驗採用中央研究院詞庫小組所研發的自動斷詞

系統來切分複合詞。

3. 方法

我們將實驗分成兩個步驟：

1. 中英對應
2. 語意標記

第一個步驟目的是要找出中文詞和英文詞的對應翻譯。實驗的資料本身包含複合詞及單字詞，所以首先必需找出在英文複合詞及中文複合詞裡的組成成份對應的翻譯。例如：”water tank” 的翻譯爲 “水槽”，我們希望能對應成 “water”→”水”，”tank”→”槽”。第二個步驟的目的則是將詞連結到 WordNet 的 synset。例如：tank 一詞在 WordNet 中一共有五個語意：

tank-1 -- an enclosed armored military vehicle
tank-2 -- a large vessel for holding gases or liquids
tank-3 -- as much as a tank will hold
tank-4 -- a freight car that transports liquids or gases in bulk
tank-5 -- a cell for violent prisoners

要決定 “tank”→”槽” 連結到哪一個語意，必需要透過語意解歧，才能知道 tank 究竟是屬於哪一個語意。我們將在 3.1 節中說明中英對應的演算法，在 3.2 節中說明語意標記的演算法。

3.1 中英對應

所謂中英對應目的是要找出中文複合詞和英文複合詞的組成成份的對應翻譯。例如：”water tank” 及 “水 槽” 的對應爲 “water/水 tank/槽”，”supplementary education” 及 “補習 教育” 的對應爲 “supplementary/補習 education/教育” 等。 關於雙語對應的研究，有許多現成的文獻可參考，如[Brown et al., 1993][Och and Ney, 2000]等。本實驗中所要對應的是比較短的複合詞而非句子，因此我們只需要考慮詞對詞的翻譯機率，而不必考慮詞的先後順序的影響。這個策略分成兩個部分 1) 計算英文詞和中文詞翻譯機率，2) 搜尋詞和詞最佳對應的路徑。這兩個步驟分別在 3.1.1 節及 3.1.2 節中說明。在實驗前，中文的複合詞已經先用中央研究院詞庫小組的斷詞系統斷好詞。

3.1.1 計算詞和詞的對應機率

我們使用 EM 演算法[Dempster et al., 1977]計算中文詞和英文詞對應的機率。假設有一個平行的詞庫 S，是由許多不同的英文詞串e_i和對應中文詞串c_i所構成，即$S=\{(e_1,c_1),(e_2,c_2),\ldots,(e_n,c_n)\}$。這些詞串可以視爲一般對應演算法中的句子。要計算詞串中每一個英文詞 w_e 翻譯成中文詞w_c 的機率 $P(w_c|w_e)$ 的計算方法如下：

Initialization:

$$P_{(e_i,c_i)}(w_c \mid w_e) = \frac{1}{m}, m = |\{w \mid w \in c_i\}|$$

E-step:

$$Z(w_c, w_e) = \sum_{(e_i,c_i)\in S} P_{(e_i,c_i)}(w_c \mid w_e)$$

M-step:

$$P(w_c \mid w_e) = \frac{Z(w_c, w_e)}{\sum_{v\in \text{chinese words}} Z(v, w_e)}$$

$$P_{(e_i,c_i)}(w_c \mid w_e) = \frac{P(w_c \mid w_e)}{\sum_{v\in c_i} P(v \mid w_e)}$$

在上面的式子中 $P_{(e_i,c_i)}(w_c \mid w_e)$ 表示w_e和w_c在詞串 (e_i,c_i) 中對應的機率。在初始值的設定上，假設對e_i中的一個英文詞 w_e 而言，對應到c_i中的每個中文詞形的機率都是 1/m，其中m表示詞串c_i中的詞數。所以 $P_{(e_i,c_i)}(w_c \mid w_e)$ 的初始值設爲 1/m 。 E-step 的目的是計算出在S中，所有包含w_e的英文詞串集$\{e_i, e_j,\ldots,e_k\}$， w_c 出現在相對應的中文詞串$\{c_i,c_j,\ldots,c_k\}$次數的期望值。M-step 則是從期望值重新估算翻譯的機率，其中分母加總的部份是針對所有中文詞爲範圍。重覆EM-step 直到收斂停止。表 1 爲英文詞 “tank” 翻譯成中文詞的機率。

英文詞	中文詞	共現次數	機率
tank	槽	492	0.354159
tank	櫃	290	0.200845
tank	艙	157	0.101734
tank	箱	59	0.040555
tank	水槽	36	0.023986
tank	池	33	0.020965
tank	罐	28	0.018258
tank	油槽	23	0.014928

表 1. 英文詞“tank”翻譯成中文詞的機率

3.1.2 搜尋詞和詞對應的最佳路徑

在上一節中利用 EM 演算法得到了每個英文詞對應到中文詞的機率值，在本節中，將利用此機率值來找出中英詞串裡中文詞和英文詞最佳的對應。我們採用路徑搜尋的方式，來找最佳的中英文詞對應。此方法說明如下：

1. 從中文詞對應到英文詞：其目的爲將中文詞組合起來，以 ”cedar nut oil” 和 “雪 松 堅果 油” 的對應爲例，由於 “雪松” 是未知詞，在對應之前是分開來的。從中文詞對應到英文詞時 “雪” 和 “松” 會同時對應到 “cedar”，所以兩個中文詞會合成 “雪松” 對應到 “cedar” 如圖 1：

	雪	松	堅果	油
cedar	**0.064020**	**0.019535**	0.000047	0.008866
nut	$9.6x10^{-6}$	0.000096	**0.035525**	0.010841
oil	$6.9x10^{-11}$	0.001410	$8.3x10^{-10}$	**0.609328**

圖 1. 從 “雪 松 堅果 油” 對應到 “cedar nut oil” 的路徑，對應的結果爲 “cedar/雪松 nut/堅果 oil/油”。

2. 從英文詞對應到中文詞：將英文詞組合起來，以 “law of universal gravitation” 和 “萬有引力定律” 的對應爲例，從英文詞對應到中文詞時 “universal” “gravitation” 兩個詞會同時對應到 “萬有引力” ，所以對應的結果會將兩個英文詞合併成複合詞 ”universal gravitation” 再對應到 “萬有引力”，如圖 2：

	law	of	universal	gravitation
萬有引力	$1.2x10^{-6}$	--	**0.033920**	**0.372825**
定律	**0.603909**	--	0.001237	0.008153

圖 2. 從 “law of universal gravitation” 對應到 “萬有引力 定律” 的路徑對應的結果爲 “law/定律 universal_gravitation/萬有引力”。

在上列的 1,2 步驟中，每一個對應只需要挑機率值最高的對應即可。例如 Pr(“law”|”萬有引力”)=$1.2x10^{-6}$ < Pr(“law”|”定律”)=0.603909，所以可以決定 “law” 對應到 “定律”。大部份的情況都只需要挑最高機率值就能找到最佳對應，但是有一些例外的情況：

1. 交錯對應：例如，”external examination” 和 “校 外 考試” 的對應情況。在此例中，”校” 和 “external” 的機率低於 “examination”，所以 “校” 和 “考試” 同時對應到 “examination”，只有 “外” 對應到 “external”，這種交錯對應的情況不甚合理。如圖 3：

	校	外	考試
external	$1.4x10^{-7}$	**0.575537**	$5.3x10^{-9}$
examination	**$5.2x10^{-6}$**	5.2x10-6	**0.172751**

圖 3. 從 “校 外 考試” 對應到 “external examination” 的路徑，對應的結果爲 “external/外 examination/校 考試”，”校” 和 “考試” 雖間隔一個字卻對應到同一個英文 “examination”。

2. 功能詞爲複合詞的一部份：在對應時，功能詞原則上不對應，但是當功能詞爲複合詞的一部份時，功能詞不能夠捨棄，例如，”general theory of relativity” 和 “廣義 相對論” 的對應，”theory” 和 “relativity” 對應到 “相對論”，但完整的複合詞應該是 “theory of relativity”，此時，功能詞不能捨棄。

上述的例外情況 1 使得對應不能總是選機率最高的情況，而必需透過路徑搜尋的方式找到最佳的路徑。在路徑搜尋的演算法上，我們採取路徑成本的計分方式。假設有一個中文詞串 “$c_1\ c_2\ c_3 \ldots c_k$” 要對應到英文詞串 “$e_1\ e_2\ e_3 \ldots e_n$”，其中某條對應路徑可以表示爲 $path_i = (c_1, e_{i1}), (c_2, e_{i2}),\ldots, (c_k, e_{ik}), e_{ij} \in \{ e_1, e_2, e_3, \ldots, e_n \}$，欲計算 $path_i$ 路徑的成本，其成本函數(cost function)定義如下：

$$\cos t(path_i) = \begin{cases} \infty, & if\ cross\_alignment(path_i) = true \\ \sum_{j=1}^{k} -\log(p(c_j \mid e_{ij})), & else \end{cases}$$

在成本函數中必需偵測路徑中是否有交錯對應(cross alignment)的存在，若存在則該路徑應被捨棄，將其成本設爲 ∞ 。而交錯對應的的偵測方式如下：

$$cross\_alignment(path_i) = \begin{cases} true, & \exists (c_p, e_{ip}), (c_q, e_{iq}) \in path_i, e_{ip} = e_{iq} \text{ and } p - q > 1 \\ false, & else \end{cases}$$

其義意爲若在對應路徑 $path_i$中存在不相鄰的兩個詞 c_p, c_q， 對應到同一個英文詞，意即 $e_{ip}=e_{iq}$，則該路徑有交錯對應。選擇路徑時，只要選擇成本最低的路徑即可，選取規則如下：

$$best\_path = \arg\min_{path_i} \cos t(path_i)$$

而例外情況 2 比較單純，只要檢查功能詞所連接的詞是否對應到同一個詞，如果對應到同一個詞就保留，否則就將功能詞捨棄。表 2 爲詞和詞對應的一些實例。

英文複合詞	**中文複合詞**	**詞和詞對應**
evaporation tank	蒸發 槽	evaporation/蒸發 tank/槽
wind-wave tank	風浪 水槽	wind-wave/風浪 tank/水槽
wave tank	波浪 水槽	wave/波浪 tank/水槽
volumetric tank	量 水箱	volumetric/量 tank/水箱
curve of learning	學習 曲線	curve/曲線 of/ learning/學習
exchange of students	學生 交換	exchange/交換 of/ students/學生
practice teaching	教學 實習	practice/實習 teaching/教學
wall cloud	雲 牆	wall/牆 cloud/雲
gas mixture	混合 氣體	gas/氣體 mixture/混合
air choke valve	阻 氣 閥	air/氣 choke/阻 valve/閥

表 2. 詞和詞對應的實例

3.2 語意標記

在做語意標記時，如果英文詞本身沒有歧義，則可以直接將其 WordNet Synset 標記在詞庫的詞上。如果有歧義，則必需做語意解歧。在本實驗中我們參考[Atserias et al, 1997]的解歧方法，總共使用三種解歧方法：

1. 解歧法一：利用英文複合詞和其組成成份之關係

   假設英文複合詞 e為 n個詞$w_{e1}, w_{e2},\ldots,w_{en}$ 所組成，若其中有一個詞 w_{ei} 的某個語意s_j 為e的某個語意 s_k上位詞，則 w_{ei}的語意為s_j，且e的語意為s_k。

   例如， “water tank-1” 為 “tank-2” 的下位，則 “water tank” 之組成成份 “tank” 之語意可標為 “tank-2” 如圖 4：

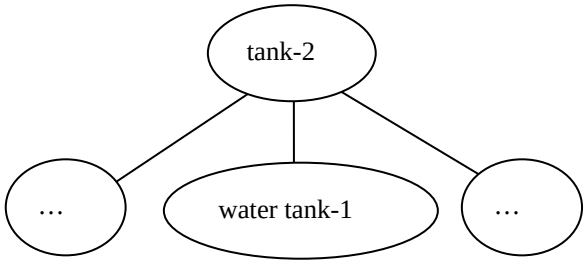


圖 4. “water tank” 和其組成成份 “tank” 的語意具有上下位關係

2. 解歧法二：英文詞之間的語意交集

   假設中文詞 w_c 可被翻譯成n個不同的英文詞 $w_{e1}, w_{e2},\ldots, w_{en}$，解歧的規則如下：

   a) 若$w_{e1}, w_{e2},\ldots, w_{en}$有一個共同的 synset s，則w_c被連結到 s。

   b) 若 t 為w_{ei} 的一個 synset ，且其餘的英文詞$w_{e1}, w_{e2},\ldots, w_{en}$都有一個 synset 落在 t 的下位，則 w_c 分別連結到 t 及這些下位。

   例如，”信號旗” 可翻譯為 “signal”, “signal flag” 及 “code flag”，而 “signal” 其中的一個 synset 為 “signal flag” 及 “code flag” 的上位，則 ”信號旗” 所標的語意可為 “signal-1”, “signal flag-1” 及 “code flag-1”。如圖 5：

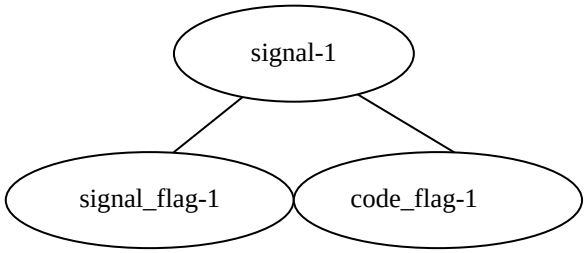


圖 5. “信號旗” 的英文翻譯 “signal”, “signal flag” 及 “code flag” 具有上下位關係

3. 解歧法三：標記沒有歧義的中文詞

利用前面實驗標記所得的結果，找出沒有歧義的中文詞，再利用沒有歧義的中文詞去標記。例如:”防波堤” 在前面實驗所得的結果，都對應到同一個 synset {breakwater-1, groin-2, groyne-1, mole-5, bulwark-3, seawall-1, jetty-1}，所以只要 “防波堤” 對應到的英文詞是該 synset 中的任一詞，就可以判別屬於該 synset。

4. 實驗結果

語意抽取的對象爲「學術名詞詞庫」和「英漢詞典」，總共有 1,254,221 個詞，其中所包含的複合詞有 645,200 佔總詞數的 51.44%。實驗的結果一共將 124,752 個中文詞連結到 42,589 個 WordNet synset，結果一共產生 165,775 個(中文詞, synset) 的連結組合。此結果平均一個中文詞落在 1.33 (由 165,775 / 124,752 得) 個 WordNet synset 中。在 4.1 節中將說明中英對應的實驗結果，4.2 節中說明解歧的結果。

4.1 中英對應的結果

由於只有複合詞需要對應，所以我們只針對複合詞做評估。評估的方法爲在對應好的詞庫中，隨機抽取 500 個詞條驗證，驗證的方式爲人工檢視。結果如表 3 所示，對應的正確率爲 95.19 %。

詞條抽樣數	對應正確數	正確率
500	476	95.19%

表 3. 中英詞與詞對應的正確率。

分析這些錯誤的對應約可規類成四種錯誤的類型，如表 4 所示。

錯語類型	錯誤的例子
中文詞切分點的錯誤	half-wave/半 length/波長 criterion/準則 spiral/螺旋 coal/煤機 cleaner/洗 american/西 ginseng/洋參 second/再 wind/生氣 microlen/微透鏡藕 coupler/合器 atomic/原子能 energy/階
音譯詞音節對應錯誤	san/聖胡 julian/連安
中英文翻譯不對稱	navigation/航行參考 star/星
英文爲縮寫	double/ III/托克馬克熱核反應器

表 4. 詞與詞對應錯誤的四種類型。

對應的結果一共得到 840,187 個中/英對譯項，其中包含了 445,830 個中文詞形和 318,048 個英文詞形。平均一個中文詞有 1.88 個英文翻譯，而一個英文詞有 2.64 個中文翻譯。

4.2 語意標記的結果

語意標記的結果評估分成兩部份探討，首先是針對語意解歧的正確性評估，其次是對整個實驗的覆蓋率評估。

4.2.1 語意解歧的正確性

在語意解歧的評估上，我們只針對有歧義的詞做評估。在抽樣的方法上採取隨機抽樣的方式，在兩個解歧實驗結果中隨機各取 200 個標記結果評估，評估的方式爲人工檢視。正確率的分析結果如表 5：

	取樣數	標記正確數	正確率
解歧法一	200	160	80.00 %
解歧法二	200	167	83.50 %
解歧法三	200	174	87.00 %

表 5. 語意解歧的正確率

4.2.2 標記的覆蓋率

在覆蓋率的分析上，我們分別針對 WordNet 2.0 語意對(word-sense pair)的覆蓋率及 synset 的覆蓋率評估。在 WordNet 2.0 中，總共有個 203,145 語意對，及 115,424 個 synset。分析的結果如表 6：

	tokens 個數	word-sense pair 個數	word-sense pair 覆蓋率	synset 個數	synset 覆蓋率
monosemous word	370991	48623	23.94 %	39953	34.61 %
解歧法一	29422	4211	2.07 %	3452	2.99 %
解歧法二	29311	2050	1.00 %	1685	1.46 %
解歧法三	81734	1931	0.95 %	1543	1.34 %
總共 (聯集)	484771	54654	26.9 %	42589	36.89 %

表 6. 抽出的語意對 WordNet 的覆蓋率

5. 結論及未來發展

在本文中我們提出一套利用「雙語學術名詞庫」來抽取中文語意的方法，本方法將問題分成兩個部份：一、先將複合詞作詞和詞對應，以得出中文和英文的翻譯，二、利用解歧的方法，將 WordNet 的語意標記在詞上。實驗的結果顯示，抽出的語意可以覆蓋 26.9 % 的 WordNet 語意對 (word-sense pair)，這些語意對函蓋了 36.89 % 的 synset。而三個解歧法的正確率分別可達 80 %, 83 % 及 87 % 的正確率。

使用學術名詞詞庫有許多好處，首先在詞庫中包含了大量的複合詞，同一個詞會搭配不同的詞一再地出現，並對應到不同的翻譯。這種翻譯多樣化的好處可以改善只使用一般「英漢詞典」翻譯用語過度典型化的缺點，同時多樣化的翻譯也有助於語意解歧。其次，由於複合詞的長度大部份只包含 2~3 詞，所以在詞和詞的對應上比句對句的對應單純且正確率也比較高。

在本實驗中可以發現，從大量的雙語學術名詞庫中的確可以抽取豐富的語意訊息。目前所使用的解歧方法只解開了一部分的歧義，未來可以嘗試不同的解歧方法，以抽取更多語意訊息。

參考文獻

A.P Dempster, N.M. Laird, and D.B. Rubin., “Maximum likelihood from incomplete data via the EM

algorithm," *Journal of the Royal Statistical Society*, B29:1-38, 1977.

CKIP, "The sense and semantic of Chinese Word," *Technical Report No. 03-01, 03-02*, 2003.

Franz Josef Och and Hermann Ney, "Improved Statistical Alignment Models," *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* , 2000.

Indrajit Bhattacharya, Lise Getoor, Yoshua Bengio, "Unsupervised Sense Disambiguation Using Bilingual Probabilistic Models, " *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004

Jason S. Chang, Tracy Lin, Geeng-Neng You, Thomas C. Chuang, Ching-Ting Hsieh, "Building A Chinese WordNet Via Class-Based Translation Model," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol 8, No.2 pp. 61-76, August 2003.

Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, Horacio Rodríguez, "Combining Multiple Methods for the Automatic Construction of Multilingual WordNets," *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 1997.

Mihalcea, R. and D. Moldovan, "A method for Word Sense Disambiguation of unrestricted text," *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999

Miller, G., "WordNet: An online lexical database," *International Journal of Lexicography*, 3(4), 1990.

Mona Diab and Philip Resnik, "An Unsupervised Method for Word Sense Tagging using Parallel Corpora, " *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

NICT, "學術名詞資訊網 (http://www.nict.gov.tw/tc/dic/index1.php), " *NICT*, 2004.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, "The Mathematics of Machine Translation: Parameter Estimation," *Computational Linguistics*, 19(2):263–311. 1993.

Philip Resnik, "Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation, " *Conference on Intelligent Text Processing and Computational Linguistics*, 2004.

利用向量支撐機辨識中文基底名詞組的初步研究

張席維 高照明 劉昭麟

台大資工系 台大外文系 政大資科系
b91083@csie.ntu.edu.tw zmgao@ntu.edu.tw chaolin@nccu.edu.tw

-

摘要

本文實做 Kudo and Matsumoto (2000, 2001)以向量支撐機（SVM）辨識基底名詞組（base NP）演算法。我們以中央研究院中文句結構樹資料庫 Sinica Treebank 3.0 的 80%作爲訓練語料,20%作爲測試語料,並比較以 Sinica Treebank 三種不同的詞性標記集訓練出來的 SVM 的辨識率（簡化標記，精簡標記，及簡化標記的大類）。實驗的結果顯示具備詳細次分類的簡化標記的辨識率最高，在封閉測試的 F-measure 爲 87.43%，初步小規模開放測試的 F-measure 爲 78.79%。詳細次分類的標記集的名詞組辨識率較高的原因是中文某些類別的動詞能夠修飾名詞，因此沒有詳細次分類的詞類標記集無法區別那些類別的動詞可以修飾名詞。與英文日文高達 94%以上的辨識率相比較，SVM 在中文基底名詞組辨識的效果並不理想，我們認爲中研院句法樹的表示法與中文本身的特性是造成辨識率不夠高的主要原因。

1. 前言

名詞組的辨識與標示（NP Chunking）是自然語言處理（NLP）的一個重要研究議題 (Ramshaw and Marcus (1995), Kudo and Matsumoto (2000, 2001))，無論是句法處理中的剖析(parsing)語意處理中的語意角色的標示（semantic role labeling）及篇章處理中的回指(co-reference)與連貫性（coherence），其它領域如資訊檢索(information retrieval)資訊擷取（information extraction）文件探勘（text mining）文件分類，與文件自動摘要都需要名詞組的辨識，例如在資訊檢索中最常被檢索的大都是名詞組（特別是人名，地名，組織名等所謂的 name entity），因此在文件或網頁中自動辨識名詞組並建立索引以方便檢索分類及自動摘要是智慧型資訊處理極爲重要的一環。

一般名詞組的辨識指的是基底名詞組（base NP），也就是將名詞組下面又包含名詞組的複雜名詞組（如關係子句及名詞組並列結構(NP conjunction)）排除在外。目前英文名詞組的辨識正確率可以達到 94%以上(Kudo and Matsumoto (2000, 2001))，但中文名詞組的辨識至今只有少數零星的研究。本文採用監督式機器學習（supervised learning）嘗試以向量支撐機(SVM, support vector machine)透過中研院句法樹庫實做 Kudo and Matsumoto (2000, 2001)所提出的演算法。本研究的主要目的在於（一）探討中文基底名詞組辨識的重要特徵（二）評估各種基底名詞組辨識的 SVM 表示法與其限制（三）從語言學的觀點分析影響中文基底名詞組辨識率的原因。

2. 文獻回顧

在大規模語法樹庫還沒有建立之前，名詞組辨識常將組成名詞組結構的規律透過有限狀態機（finite state machines）去找出符合名詞組的 pattern (Voutilainen (1993)) 或從標記好詞性的語料庫以統計的方式得到（Church (1988)），或結和語言規律和語料庫統計（Chen and Chen (1994)）。自從賓州大學大規模的英文語法樹庫(Penn Treebank)建構完成後 (Marcus, Santorini and Marcinkiewicz (1993)),絕大多數的名詞組辨識研究是以機器學習（machine learning）的方法透過語法樹庫裡面的語法結構及前後語境的特徵得到。運用機器學習辨識名詞組的方法大致可分爲 HMM (hidden Markov model)，transformation-based（Ramshaw and Marcus （1995））, memory-based （Veenstra （1998）, Tjong Kim Sang and Veenstra （1999） Argamon, Dagan and Krymolowski（1998））, maximum entropy （Skut and Brants（1998））, 及 SVM (Kudo and Matsumoto, 2000, 2001)等方法。上述幾種的方法都是監督式學習。HMM (hidden Markov model)使用統計的方法在 finite state machine 的 transition function 之上加上語料庫的統計結果。transformation-based learning 由現有的語料庫訓練出 transformational rules, 再利用這些規則對測試資料作 parse。HMM，transformation-based learning, memory-based learning 在自然語言處理中已被廣泛應用。 SVM 則是一種較新的 machine learning 技術,近幾年逐漸被應用到自然語言處理的各項研究議題,。

上述這些演算法針對英文 Wall Street Journal Corpus 訓練得到的結果顯示,精確率(precision)與召回率(recall)大都超過 90%, 其中以 SVM （Kudo and Matsumoto（2001））的效果最好，精確率(precision)與召回率(recall)都超過 94% （http://staff.science.uva.nl/~erikt/research/np-chunking.html）。

中文名詞組辨識的研究起步較晚，迄今只有零星的研究，還沒有針對同一個語料庫的大規模的測試與比較。例如中國大陸學者 Zhao and Huang (1998)提出以語料庫統計結合規律，利用 minimum description length principle (MDL)得到 quasi-dependency strength 加上規律來得到 base NP。這種採用非監督式機器學習（unsupervised learning）的方法，在封閉測試(close test)和開放測試(open test)中分別有 91.5% 和 88.7%的精確率。本研究使用 SVM 演算法來辨識中文基底名詞組（base NP），以便瞭解影響中文基底名詞組辨識的最重要特徵究竟有哪些，以及 SVM 在中文基底名詞組辨識的效果與限制。

3. 中文句法樹庫

由於 SVM 是監督式學習的演算法，我們必須擁有中文句法樹庫(treebank)的資料才能訓練出辨識名詞組的程式。目前我們擁有的兩個句法樹庫資料,一個是中央研究院中文句結構樹資料庫（Sinica Treebank 3.0）(http://www.aclclp.org.tw/use_stb_c.php)，另一個爲美國賓州大學中文句法樹庫（Penn Chinese Treebank 4.0）(http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T05)。兩者在語言，語料來源，語料庫大小，標記集，標記單位，標記訊息，及依據的語言學理論都不相同。下面是兩者的比較。

表（一）Sinica Treebank 3.0 與 Penn Chinese Treebank 4.0 的綜合比較

	語言	語料來源	語料大小	標記集	句法樹的性質	標記的訊息	所採用的語法理論
Sinica Chinese Treebank 3.0	繁體中文	台灣的報紙	290144 個詞，54902 棵結構樹	CKIP Tagset 包含未簡化標記，簡化標記及精簡標記三種	以標點符號分隔的詞組	語法及語意（包括詞組結構，中心語，修飾語，及語意角色等）	Information-based Case Grammar (ICG)
Penn Chinese Treebank 4.0	簡體中文	大部分爲中國大陸新華社新聞，少部分爲香港新聞及台灣光華雜誌社文章	404156 個詞，15162 棵結構樹	只有一個標記集。與英文 Penn Treebank 部分相同，另有部分標記是專爲中文設計	大部分爲完整的句子	語法結構與語法功能（主詞，受詞）	大致上採用 Chomsky 的 Government and Binding Theory 的詞組結構理論但額外加註語法功能的訊息。

表（二）Sinica Treebank 3.0 與 Penn Chinese Treebank 4.0 的部分例子

Sinica Treebank 的格式與例子	#PP(Head:P50:除了|DUMMY:GP(DUMMY:NP(Head:Nab:排鼓)|Head:Ng:以外))# #PP(Head:P21:在|DUMMY:GP(DUMMY:NP(Head:Nad:政治)|Head:Ng:上))# #NP(quantifier:NP(Head:Neqa:這些)|predication:VP・的(head:VP(Head:VL4:令 |goal:NP(Head:Nab:人)|theme:VP(Head:VK1:討厭))|Head:DE:的)|Head:Nac:戲)# #S(theme:VP(Head:VH11:這樣)|Head:VH11:好|particle:Ta:了)# #S(result:Cbca:而|theme:NP(quantifier:NP(Head:Nes:該)|Head:Ncb:校)|evaluation:Dbb:也 |Head:VK2:需要|goal:NP(quantifier:DM:一個|property:Nac:英文|Head:Nab:老師))#
Penn Chinese Treebank 的格式與例子	((IP (NP-PN-SBJ (NR 上海) (NR 浦东)) (VP (VP (LCP-TMP (NP (NT 近年)) (LC 来)) (VP (VCD (VV 颁布) (VV 实行)) (AS 了) (NP-OBJ (CP (WHNP-1 (-NONE- *OP*)) (CP (IP (NP-SBJ (-NONE- *T*-1)) (VP (VV 涉及) (NP-OBJ (NP-APP (NN 经济) (PU 、) (NN 贸易) (PU 、) (NN 建设) (PU 、) (NN 规划) (PU 、) (NN 科技) (PU 、) (NN 文教) (ETC 等)) (NP (NN 领域))))) (DEC 的))) (QP (CD 七十一) (CLP (M 件))) (NP (NN 法规性) (NN 文件))))) (PU ，) (VP (VV 确保) (AS 了) (NP-OBJ (DNP (NP (NP-PN (NR 浦东)) (NP (NN 开发))) (DEG 的)) (ADJP (JJ 有序)) (NP (NN 进行))))) (PU 。)))

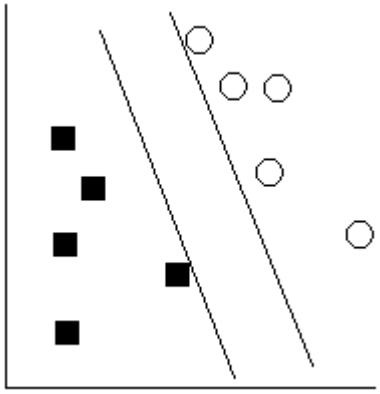
Sinica Treebank 與 Penn Chinese Treebank 最大的差別在於結構樹的語法單位不同。前者以標點符號作爲分隔不同結構樹的單位，因此一個結構樹很多時候只是一個詞組（如 PP, NP）而不是一個完整的句子。而後者除小部分結構樹是句子的片段(以 FRAG 標示)大部分的結構樹是完整的句子(sentence)(以 IP 標示)。另外 Sinica Treebank 語法結構採取中心語主導原則 (Head-Driven Principle)，註明中心語(Head)和其他成分（如附加語）的語法和語意訊息，表達出句子中詞和詞之間的語法結構和語意角色關係（http://godel.iis.sinica.edu.tw/CKIP/treebank.htm），而 Penn Chinese Treebank 並沒有中心語與語意角色的訊息，而是在詞組上加註如主詞 SBJ 受詞 OBJ 等語法功能的方式來取代。

我們選擇 Sinica Treebank 作爲訓練語料的主要原因在於做開放測試時我們需要一個與訓練語料採用同樣標記集的分詞與詞性標記程式。如果採用 Sinica Treebank，開放測時我們可以使用中研院的線上分詞與詞性標注程式 http://ckipsvr.iis.sinica.edu.tw/ 做爲我們 SVM 演算法的輸入資料。該線上程式的準確率相當高，特別是對於辭典未收錄詞的分詞與詞性標注的準確率比其它類似程式要高，採用該程式可以有效降低因爲一開始分詞與詞性標注錯誤而導致後面 SVM 演算法判斷錯誤的機率。另外由於 Sinica Treebank 有未簡化標記，簡化標記及精簡標記三種標記集，相較於 Penn Treebank 只有一種標記集，Sinica Treebank 的三種不同的標記集可以作爲不同的特徵。除此之外只有 Sinica Treebank 有標示語意角色的訊息，未來如果我們要以 SVM 來訓練標記語意角色，Sinica Treebank 顯然是較佳的選擇。

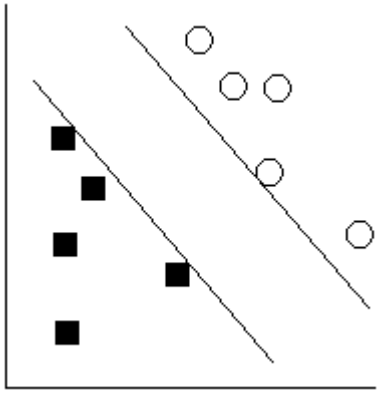
4. SVM 簡介

SVM 是較新的 machine learning 技術 (Boser, Guyon, and Vapnik (1992), Corts and Vapnik (1995)) 它使用一些策略來最大化具有不同特徵的資料中間的界限，並針對未知資料的特徵來判斷它屬於哪個類別。SVM 已在文件分類（Joachims （1998） Taira and Haruno （1999））以及名詞組標示 (Kudo and Matsumoto （2000, 20001）) 取得超越其它作法的準確性，而近幾年應用在自然語言處理的各個議題的研究更是方興未艾，如未知詞辨識（unknown word guessing）(Nakagawa, Kudo, and Matsumoto (2001)）詞性標注(part of speech tagging)Nakagawa, Kudo, and Matsumoto (2002)， Giménez J esús and Márquez Lluís (2004))句法依存關係辨識(dependency analysis)(Kudo and Matsumoto (2000))詞義辨別與標注(word sense disambiguation and sense tagging) (Cabezas, Resnik, and Stevens (2001))語意剖析(semantic parsing) (Pradhan et al. (2004) Sun and Jurafsky (2004))等都取得不錯的成果。

SVM 是一個分類用的 machine。請參照圖（一，二），



圖一



圖二

SVM 找出兩種資料 (黑色方形與白色圓形) 中間的界限, 圖一，圖二顯示出可能的兩種分割方式, 顯然的, 後者的切割方式是較佳的 (兩種資料的界線爲兩平行線之中線), 而 SVM 以滿足下面條件

$$\min\Phi(\omega)= (1/2)|\omega|^2$$

找出最佳平面 (即在線性可分的情況下, 可視爲解二次規畫的問題), 而此可由拉格朗日乘子法（Lagrange multiplier）求解.

由於很多的問題常常並不是線性可分的 (如我們的詞組切割), 這個時候 SVM 在比現有資料更高的向量空間 H 使用線性分類函數 $\Phi: R^d \rightarrow H$ 將 x 對應到高維空間, 便可在此以不破壞資料特徵亦不增加複雜度的方式對其進行分類.

在轉換的過程中, 我們會使用一 kernel function: $K(x_i,x_j) = \Phi(x_i)^T\Phi(x_j)$ 來實現非線性變換後的線性分類, 而使用不同的 kernel function 對不同的資料會有不同的效果.

以下爲一個簡單的 SVM 運作方式

給定一個訓練的資料集合:

(x_i, y_i) { i= 1, 2, ..., l; x_i 屬於 R^n; y_i 屬於 { 1, -1 }}

其中 l 爲訓練之資料數, x_i 爲一個 n 維向量, y_i 則是其類別 (分爲正類別 1 與負類別 -1) SVM 找到正類別與負類別中之最大的界限, 即解決下面的最佳化問題的解答

$\min_{w,b,e}(1/2)w^Tw + C\Sigma_{i=1}^{l}e_i$ 使得

$y_i(w^T\Phi(x_i) + b) >= 1- e_i, e_i >= 0$

x_i 經由 Φ 函數被對應到一個更高維的向量空間 H 之後 SVM 於此找到不同類別之間最大的界限; $K(x_i,x_j)$ 爲 Kernel function.

5. 詞性標記集與中文句法樹庫

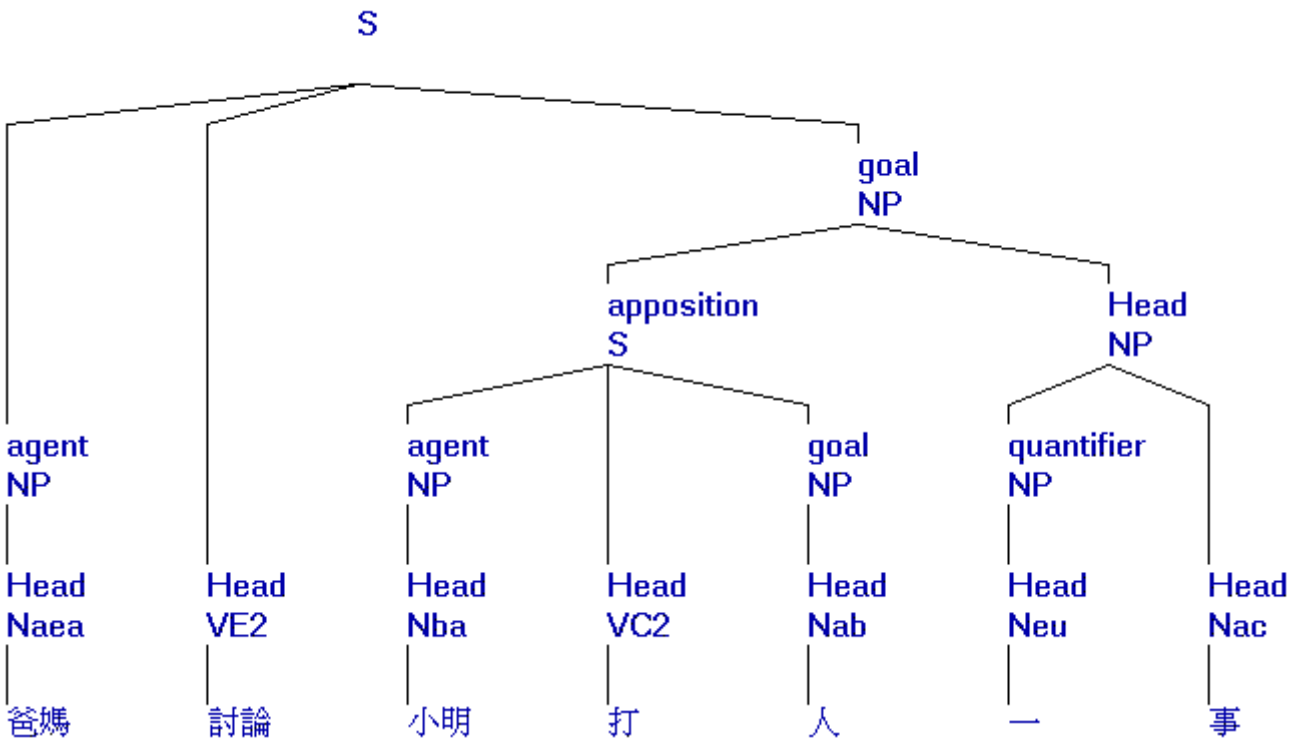
我們的實驗使用中研院語法樹庫 Sinica Treebank。中研院的詞性標記集及每個標記代表的語言學涵義如表(三)。（參考 http://www.sinica.edu.tw/SinicaCorpus/modern_c_wordtype.html）中研院詞知識庫小組所出版的「中文詞類分析」技術報告所提出的中文詞類的分類比表（三）的簡化詞類更細，但爲了顧及實用性中研院的漢語平衡語料庫所用的詞類標記爲已經經過合併的簡化詞類。我們可以看出即使是簡化詞類，連接詞，名詞，動詞，副詞每一項都有不少的次分類。以動詞爲例除了先分成動作與狀態兩大類之外，另外又根據動詞所帶的論元（argument）數目與種類各自分爲若干小類。中研院另外又將簡化詞類做進一步的合併形成所謂的精簡詞類。在簡化詞類裡面的動詞原先有 16 類但在精簡標記裡面只剩及物與不及物動詞 2 類。

表(三) 中研院的詞性標記集

簡化詞類	代表意義	精簡詞類	簡化詞類	代表意義	精簡詞類	簡化詞類	代表意義	精簡詞類
A	非謂形容詞	A	Nb	專有名稱	N	VB	動作類及物動詞	Vi
Caa	對等連接詞	C	Nc	地方詞	N	VC	動作及物動詞	Vt
Cab	連接詞，如：等等	POST	Ncd	位置詞	N	VCL	動作接地方賓語動詞	Vt
Cba	連接詞，如：的話	POST	Nd	時間詞	N	VD	雙賓動詞	Vt
Cbb	關聯連接詞	C	Nep	指代定詞	DET	VE	動作句賓動詞	Vt
D	副詞	ADV	Neqa	數量定詞	DET	VF	動作謂賓動詞	Vt
DE	的, 之, 得, 地	T	Neqb	後置數量定詞	POST	VG	分類動詞	Vt
Da	數量副詞	ADV	Nes	數量副詞	DET	VH	狀態不及物動詞	Vi
Dfa	動詞前程度副詞	ADV	Neu	數詞定詞	DET	VHC	狀態使動動詞	Vt
Dfb	動詞後程	ADV	Nf	量詞	M	VI	狀態類	Vi

	度副詞						及物動詞	
Di	時態標記	ASP	Ng	後置詞	POST	VJ	狀態及物動詞	Vt
Dk	句副詞	ADV	Nh	代名詞	N	VK	狀態句賓動詞	Vt
FW	外文標記	FW	SHI	外文標記	Vt	VL	狀態謂賓動詞	Vt
I	感嘆詞	T	T	語助詞	T	V_2	有	Vt
NAV	名謂詞	NAV	VA	動作不及物動詞	Vi			
Na	的, 之, 得, 地	N	VAC	動作使動動詞	Vi			

而 NP, VP 等詞組的判斷標準亦採用中研院句法樹庫的資料做爲我們測試的標準, （圖三）」是一個範例樹圖: (取自 http://godel.iis.sinica.edu.tw/CKIP/treebank/apposition.htm)



（圖三）中研院句法樹庫範例

如（圖三）所示，中研院的中文句法樹庫的 terminal node 是詞，詞上方有詞性標記和中心語（head）這類的語法訊息，構成詞組的結點(node)有詞組標記和語意角色等語意訊息。我們的焦點是 NP,也就是由 ”爸媽”, ”小明”, ”人”, ”一”, ”一事”組成的詞組。”小明打人一事”這類名詞組因爲包含其它的名詞組，不屬於基底名詞（base NP），

所以不在我們的討論之列。

6. 以 SVM 辨識中文名詞組的實作與實驗結果

訓練語料由於採取中研院的句法樹庫所以句子已經分詞並標注詞性。我們以 (Kudo and Matsumoto (2000, 2001) 的經驗做爲名詞組的辨識基礎。第一次實驗以 (I,O,B)三個標記分類:

這個方法以三個 class （I,O,B）表示一個詞在詞組中的位置:

I: 詞在詞組之中
O: 詞在詞組之外
B: 緊接著一個詞組之詞組的開頭

此種方法被 Tjong Kim Sang 稱爲 IOB1 表示法, 另外還有 IOB2, IOE1, IOE2,
在此不多加詳述.

Start/End
最初被用在日本語的作業上 (Uchimoto et al.（2000)）, 也就是 S，E，加上 I,O,B,共五個 class:

B: 多詞詞組的開頭
E: 多詞詞組的結尾
I: 詞在多詞詞組中
S: 單詞詞組
O: 詞在詞組之外

以下爲兩者之範例標記:

	Inside/Outside	Start/End
這	I	S
是	O	O
詞組	I	B
標記	I	I
範例	I	E
說明	B	S

一開始, 我們簡單的將測試資料排列成 7 維的向量, Wordi 是 i 位置的詞，POSi 是 i 位置詞的標記，加上前後各兩個詞的標記:

Wordi POS(i-2) POS(i-1) POSi POS(i+1) POS(i+2)

這裡根據詞，詞的標記, 和前面後面各兩個詞的標記來做分類。上面的範例向量表示如下:

I 1:這 2:0 3:0 4:N 5:S 6:N
O 1:是 2:0 3:N 4:S 5:N 6:V
I 1:詞組 2:N 3:S 4:N 5:V 6:N
I 1:標記 2:S 3:N 4:V 5:N 6:V
I 1:範例 2:N 3:V 4:N 5:V 6:0
B 1:說明 2:V 3:N 4:V 5:0 6:0

中研院中文句結構樹資料庫有 54,902 棵中文結構樹，290144 個詞。我們用樹庫的 80％做爲訓練的資料,20％做爲測試資料。由於不知道訓練語料對於 SVM 而言是否足夠，我們第一次實驗採用最少的特徵，採用的向量爲大類詞性 (即 N, V, P, ...)，也就是只看中研院詞類簡化標記的第一個字母所形成的大類。這比精簡標記的類別更少。我們使用現有的 SVM Tool: LIBSVM (Chang and Lin （2004)）作爲工具。SVM kernel function 爲 $K(x_i,x_j) = (\gamma\, x_i^T x_j + r)^d, \gamma > 0$ 以多項式趨近.實驗結果列於表（四）的第一列.

由表（四）可見辨識的成果並不好。 從語言學的角度來分析，中文名詞組的辨識比英文困難原因在於中文的動詞可以修飾名詞，例如投資大眾，建設公司,流浪教師等。這些詞沒有任何構詞上的特徵或證據可以視爲名物化（nominalization），因此詞性標記程式很難將這些詞判斷成名詞。由於中文的動詞可以修飾名詞使得自動辨識中文名詞組變得相當困難。不過我們仔細觀察後可以發現並不是所有的中文動詞都可修飾名詞，例如 VD(雙賓動詞), VK(狀態句賓動詞), VG(分類動詞)等這些類的動詞很少有修飾名詞的例子。由於我們採用的向量爲大類詞性 (即 N, V, P, ...), 動詞次分類這個重要特徵沒有考慮進去，因此實驗的結果非常不理想。如下面的例子:

可能(D) 代表(VK) 台灣(Nc) 人民(Na) 對(P) 朝野(Na) 政黨(Na) 傳達(VD) 訊息(Na)

程式抽取出來的 NP chunks 爲: “台灣人民”, “朝野政黨傳達訊息”; 顯然的 “傳達” 並不應該出現在 NP chunk 之中, 而就我們給予 SVM 的資料來看, 這邊並沒有明顯的訊息可以得知其不適用 (我們給予 SVM 的資料爲 “傳達(V)”), 而如 VH 等靜態動詞之類的動詞, 卻又常常出現在 NP 之中, 同樣標示爲 V。 由於我們採取簡化詞類標記的第一個字母的大類來表示，在缺乏動詞次分類訊息特徵的情形下使得實驗結果非常不理想.

因此, 我們保留將簡化標記動詞次分類的特徵，其它詞性則仍然使用大類, 結果如表（四）第二列所顯示，改良的方法在精確率上提升了 23%以上, 召回率也提升了 6%以上，雖然還不是非常好, 但顯示了詞性標記的選擇（有無動詞次分類的訊息）是影響 SVM 效果的重要的特徵.

表（四）動詞次分類訊息對 SVM 的影響

	Precision	Recall
(1)取簡化標記詞性第一個字母做大部分類	54.99%	53.17%
(2)動詞採用簡化標記細部分類其餘詞性取第一個字母大部分類	78.18%	59.33%

無論是精確率或召回率，我們實驗的結果與 Kudo and Matsumoto (2000,2001)發表的結果 (94%) 差了一大段距離; 可以改進的地方如下:

IOB tag, 我們的實驗只採取了 I/O 兩種 tag, 這在當兩個 chunk 緊連的時候會是一個致命的問題 (無法確認 chunk 的終結點)。修改 tag, 使用 IOB 與 Start/End 將可提升辨識率.

由目前的經驗得知, 好的詞性分類有助於準確度的提升。所謂好的詞性分類是指透過細部的詞性分類將能名詞組內部與外部兩種不同的特徵顯示出來，而將無助於此項辨識工作的詞性細部分類精簡成大類。如此透過 SVM 演算法可以提升名詞組的辨識精確率。

kernel function 與其微調的參數是影響 SVM 準確度的一大原因, 預期將會使用 linear, polynomial, radial basis function, sigmoid... 等等函數來做逼近, 並嘗試採用 cross validation 來尋找最佳參數.

目前面對的問題還有一點爲: 訓練的時間太久. 一個約 8,000 詞的訓練資料約需要花費 4 分鐘, SVM 之 time complexity 約爲 $O(n^2)$, 也就是說若有一 300,000 詞之訓練資料, 將需要花費約三天以上的時間訓練, 如此一來, 對於要使用 cross validation 將會是一大挑戰, 因此會嘗試使用 scaling 的方式來減少所需要訓練的時間.

YAMCHA (http://chasen.org/~taku/software/YamCha/)是 Taku Kudo 專門爲 NP Chunking 所設計的 SVM 工具，因此比一般性 SVM 工具（SVM Tool: LIBSVM (Chih-Chung Chang and Chih-Jen Lin, 2004)）方便實做。

YAMCHA 與 libsvm 的最大不同點在於:

a) Dynamic programming

b) Kernel Function

由於 libsvm 本身的限制, 我們很難能即時的將 chunking 的結果應用在下面一個未知 chunking 的判斷. 舉例而言, 之前的句子:

	Inside/Outside
這	I
是	O
詞組	I
標記	I

範例　　I

說明　　(B)

當 SVM 要判斷"說明"這個詞的 tag 時，它會去參考"標記"與"範例"的詞與詞性;原來的設計並未考慮到它們的 IOB tag，而由於中文（其實任何語言應該都一樣）有前後相依性，因此把 IOB tag 計算在內，會是一個適當而重要的特徵.

YAMCHA （Kudo and Matsumoto (2000,2001)）使用 IOB tag 代替 IO tag 方面，由於 B tag 表示了一個緊鄰之前 NP-chunk 的開頭，解決了兩個相鄰 NP-chunk 的分類問題.

另外 Kudo and Matsumoto (2000,2001)使用 voting 來提升辨識效果。voting 在很多應用中經常被使用。我們有許多不同的標記集，和不同方向的 parsing 方式（backward 即將所有的詞顛倒排列後做訓練與測試），藉著由不同標記集和不同的 parsing 方向訓練出來的 SVM 模型，可以採用其 Accuracy 之分數來統計未知詞組的得分. 這種方法可以避開某些詞性標記或者是 parsing 方向的盲點，以提升準確度。

另外從我們第一次的實驗結果得知動詞次分類訊息是一個影響 SVM 效果的重要的特徵。忽略動詞次分類的訊息會使辨識效果差很多。我們希望能從實驗數據中比較使用簡化詞類和精簡詞類是否會有很大的差別。

Kudo and Matsumoto (2000)以資訊檢索常用的 F measure 作爲評估系統的標準。F = (2* precision * recall) / (precision + recall)。由於 precision 高時則 recall 低，而 recall 高時則 precision 低，F measure 同時考慮 precision 與 recall，成爲評估時的綜合指標。

表（五）是我們利用 YAMCHA 實作 Base-NP chunking 所得到的結果。

表（五）不同的標記集和 parsing 方向的辨識率

	Precision	Recall	F measure
簡化詞類 (Forward)	86.48% (10360/11980)	88.41% (10360/11716)	87.43%
簡化詞類 (Backward)	86.29% (9983/11569)	85.21% (9983/11716)	85.74%
精簡詞類 (Forward)	87.34% (8789/10063)	75.02% (8789/11716)	80.71%
精簡詞類 (Backward)	84.88% (8651/10192)	73.84% (8651/11716)	78.98%
Vote using Accuracy Rate	88.71% (10048/11327)	85.76% (10048/11716)	87.21%

從表（五）可以觀察到 F measure 最高的是簡化詞類 forward parsing，使用 voting 並沒有提升 F measure，這是不是與訓練語料量不夠大有關，或其它因素造成，還是意味著中文只要 forward parsing 就能得到最好的效果不需要 backward parsing 和 voting，這些都有待進一步研究。值得注意的是在召回率（recall）方面簡化標記比精簡標記高 12 個百分點以上，原因是簡化標記具有 16 個動詞次分類而精簡標記動詞只有及物和不及物兩個次分類。由於精簡標記沒有足夠詳細的次分類的特徵，導致不少基底名詞組被誤判成動詞組。如果拿表（五）最好的結果與第一次的實驗結果表（四）比較，精確率提高了 10 個百分點，召回率則提高了 26 個百分點，這顯示

dynamic programming 和使用 IOB 與 Start/End 發揮了功用。雖然與英文的 95% F measure 仍有一大段差距，但是辨識效能已經大幅度的提升。

中研院的句法樹庫經過人工檢查，所以很少有錯誤。但開放測試時由於輸入的句子必須經過分詞和詞性標注（此部分透過中研院詞庫小組的線上分詞與詞性標注系統 (http://ckipsvr.iis.sinica.edu.tw/)），而分詞與詞性標注這兩個過程都有可能出錯，因此可以預期在開放測試時辨識的正確率會比封閉測試差，我們初步小規模的開放測試證實了這個預測。精確率與召回率分別爲 81.25%與 76.47%，F measure 則爲 78.79%。

7. 觀察到的問題

我們將訓練出來的模型實際使用在名詞組辨識時發現了幾個原來並未考慮到的問題。例如當我們實際測試如下的資料:

這(Nep) 是(SHI) 一(Neu) 個(Nf) 公平(VH) 的(DE) 審判(Na)

名詞組辨識的結果爲:
這(Nep)
一(Neu)

我們預期的結果:
這(Nep)
審判(Na)

我們發現了這個奇怪的結果之後，做了很多的測試，發現 Neu 類型詞皆會呈現單獨的 chunk，而 DE 之後則完全不會有 chunk. 這個問題在我們回頭檢查中研院句法樹庫時有了答案.

base-NP 就定義來看，爲 non-recursive NP chunk. 而在中研院句法樹庫中, DE 開頭的句子本身會成一個 (NP・的) 的 structure，而其餘接在 DE 詞 (如"的", "之") 之後的詞組皆不會成爲單獨的 NP chunk，也因此在上面我們所期望的"審判", 並沒有被抽取出來. 換句話說，在我們給予的訓練資料裡，就已經沒有將其標記爲 NP chunk 的例子了.

由於中研院句法樹庫特殊的標記方式，目前的情況是無法完全得到我們"看似"base-NP 的詞組分類結果，雖然 SVM 正確的分解出來它所判斷的詞組，但並不完全是我們想要的，這個問題目前似乎還沒有較好的解決方法.

檢視開放測試的資料，我們發現造成 SVM 判斷錯誤的主要有兩種情形。一種是 Nd 類的名詞不修飾名詞而當副詞，例如「以後(Nd) 經濟部(Nc) 」及「未來(Nd) 台灣(Nc) 水(Na) 」都被誤判成名詞組，另一種錯誤則是動詞修飾名詞例子，例如「重要(VH) 工作(Na)」這個名詞組並沒有被辨識出來。理論上中研院句法樹庫中 VH 類動詞修飾名詞的例子非常多，SVM 應該可以辨識出這樣的結構，實際上卻沒有辨識出來，造成此類錯誤的原因

還需要進一步研究。

8. 結論與未來的研究

我們的實驗中顯示動詞次分類的訊息對於提昇基底名詞組辨識的精確率與召回率而言是一個重要的特徵。我們的實驗間接證明中研院簡化標記中詳細動詞次分類訊息在中文自然語言處理上的優點。該分類系統先將動詞分成動作及狀態兩大類，再依據動詞的論元結構(argument structure)詳細分類。10 幾年前中研院詞知識庫小組院設計這套詞類標記系統的語言學家和計算語言學家或許只是單純從句法學與語意學的觀點提出這樣的分類系統。從事中文自然處理的研究人員對如此龐大詳細的詞類標記系統的必要性或許會質疑。但從我們的實驗中發現，中研院詳細的詞類次分類至少在動詞的次分類方面的確爲解決名詞組辨識的問題預先鋪路。機器學習演算法固然可以從訓練資料中自動學習，但是沒有專家的知識來分辨重要的特徵，仍然無法得到良好的效果。自然語言處理研究如果要有更進一步的發展，單靠機器學習演算法是不夠的，結合語言規律與知識是必然要走的路。

我們的實驗也顯示中研院句法樹庫某些結構表示法不利於我們辨識基底名詞組,加上中文的動詞可以修飾名詞造成同一個 SVM 演算法的辨識率比英文日文低許多。

近期我們會嘗試結合更多的語言知識及語言特徵，例如 Zhao and Huang (1998)以語料庫統計結合規律的方式來提升辨識率。我們相信類似的研究不僅有助於解決名詞組辨識的問題，對中文詞類標記集與句法樹庫的設計與修正也能提供回饋。

致謝

本研究得到國科會專題研究計畫 93-2411-H-002-013 「詞彙語意關係之自動標注─以中英平行語料庫爲基礎(3/3) 」94-2411-H-002-043 「中英平行句法樹庫的建立與英漢結構對應演算法的研究」及 94 年度國科會大專學生參與專題研究計畫「利用 SVM 標示中文名詞組的研究」經費補助，特此致謝。

參考文獻

Argamon, Shlomo, Dagan, Ido, and Krymolowski,Yuval (1998). A Memory-Based Approach to Learning Shallow Natural Language Patterns. In Proceedings of the 17th international conference on Computational linguistics, Vol. 1, pp. 67 - 73 , Montreal, Quebec, Canada.”

Brill, Eric and Ngai, Grace (1999), Man vs. Machine: A Case Study in Base Noun Phrase Learning. In Proceedings of ACL'99, pp. 65-72, University of Maryland, MD, USA.

Boser, E. Bernhard, Guyon, Isabelle, and Vapnik, Vladimir. (1992). A Training Algorithm for Optimal Margin Classifiers. COLT: pp. 144-152

Cabezas, Clara, Resnik, Philip, and Stevens,Jessica. (2001). Supervised Sense Tagging using Support Vector Machines. Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), Toulouse, France, 5-6 July 2001.

Cardie, Claire and Pierce, David (1998). Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification. In Proceedings of COLING-ACL'98, pp. 218-224, Montreal, Canada.

Chang, Chih-Chung and Lin, Chih-Jen. (2004) LIBSVM -- A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Chen, Kuang-hua and Chen, Hsin-Hsi (1994). Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation, In Proceedings of ACL-94, Las Cruses, NM, USA.

Church, K. (1988) A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Second Conference on Applied Natural Language Processing*, Austin , Texas , pp. 136-143.

Corte, Corinna, and Vapnik, Vladimir (1995). Support-Vector Networks. Machine Learning 20(3), pp. 273-297.

Giménez J esús and Márquez Lluís (2004). SVMTool: A general POS tagger generator based on Support Vector Machines Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. 2004 .

Joachims, Thorsten. (1998) *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* Proceedings of the European Conference on Machine Learning (ECML), Springer, 1998.

Hsu, Chih-Wei, Chang, Chih-Chung, and Lin, Chih-Jen. (2004). A Practical Guide to Support Vector Classification.

Kudo, Taku, and Matsumoto, Yuji. (2000). Use of Support Vector Learning for Chunk Identification. In Proceedings of CoNLL-2000, pp. 142-144.

Kudo, Taku, and Matsumoto, Yuji (2000). Japanese Dependency Analysis Based on Support Vector Machines, EMNLP/VLC 2000

Kudo, Taku, and Matsumoto, Yuji. (2001). Chunking with Support Vector Machine. In Proceedings of NAACL 2001, pp. 192-199.

Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann. (1993) Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics, 19:2. vol. 19, no. 2, pp. 313–330.

Pradhan, Sameer, Ward, Wayne, Hacioglu, Kadri, Martin, James H.and Jurafsky, Daniel. (2004). Shallow Semantic Parsing Using Support Vector Machines. In Proceedings of NAACL-HLT 2004,pp. 233-240..

Nakagawa, Tetsuji, Kudo, Taku, and Matsumoto, Yuji. (2001). Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines. NLPRS, pp. 325-331

Nakagawa, Tetsuji, Kudo, Taku, and Matsumoto, Yuji. (2002). Revision Learning and its Application to Part-of-Speech Tagging. In Proceedings of ACL 2002, pp. 497-504.

NP Chunking. http://staff.science.uva.nl/~erikt/research/np-chunking.html

Ramshaw, Lance A., and Marcus, Mitchell P.. (1995). Text Chunking Using Transformation-based Learning. In Proceedings of the Third ACL Workshop on Very Large Corpora, pp. 82-94, Cambridge MA, USA.

Skut, Wojciech and Brants, Thorsten. (1998) A Maximum-Entropy Partial Parser for Unrestricted Text. In Proceedings of the Sixth Workshop on Very Large Corpora, pp. 143-151, Montreal, Canada.

Sun, Honglin and Jurafsky, Daniel. 2004. Shallow Semantic Parsing of Chinese. In Proceedings of NAACL-HLT 2004, pp.192-199.

Taira, Hirotoshi, Haruno, Masahiko（1999）: Feature Selection in SVM Text Categorization. AAAI/IAAI 1999, pp. 480-486.

Tjong Kim Sang, Erik F. and Veenstra, Jorn (1999). Representing Text Chunks. In Proceedings of EACL'99, 173-179, Bergen, Norway.

Tjong Kim Sang, Erik F. (2002) Memory-Based Shallow Parsing. Journal of Machine Learning Research, Vol. 2, pp. 559-594.

Uchimoto, Kiyotaka, Ma, Qing, Murata, Masaki, Ozaku, Hiromi, Isahara, Hitoshi. (2000) Named entity extraction based on a maximum entropy model and transformation rules. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, pp. 326 – 335.

Veenstra, Jorn. (1998). Fast NP chunking using memory-based learning techniques, In F. Verdenius and W. van den Broek eds.,Proceedings of BENELEARN-98, pp. 71-79, Wageningen, The Netherlands.

Voutilainen, A. (1993) NPtool, a Detector of English Noun Phrase. In Proceedings of the First Annual Workshop on Very Large Corpora, pp. 48-57.

YamCha: Yet Another Multipurpose CHunk Annotator http://chasen.org/~taku/software/YamCha/

Zhao,Jun and Huang, Changning. (1998). A Quasi-Dependency Model for Structural Analysis of Chinese BaseNPs. In Proceedings of COLING-ACL 98, pp. 1-7 , Montreal, Canada.

中文詞類分析 (1988). 中央研究院詞知識庫小組技術報告,台北。

中研院詞知識庫小組中文斷詞系統(包含未知詞擷取與標記) http://ckipsvr.iis.sinica.edu.tw/

中文句結構樹資料庫」(Sinica Treebank Version 3.0). 中華民國計算語言學會
http://www.aclclp.org.tw/use_stb_c.php

史忠植 (2003). 知識發現, 清華大學出版社,北京.

Perceptual Factor Analysis for Speech Enhancement

Chuan-Wei Ting and Jen-Tzung Chien

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, ROC
{motor,chien}@chien.csie.ncku.edu.tw

Abstract

This paper presents a new speech enhancement approach originated from factor analysis (FA) framework. FA is a data analysis model where the relevant common factors can be extracted from observations. A factor loading matrix is found and a resulting model error is introduced for each observation. Interestingly, FA is a subspace approach properly representing the noisy speech. This approach partitions the space of noisy speech into a principal subspace containing clean speech and a complimentary (minor) subspace containing the residual speech and noise. We show that FA is a generalized data model compared to signal subspace approach. To perform FA speech enhancement, we present a perceptual optimization procedure that minimizes the signal distortion subject to the energies of residual speech and noise under a specified level. Importantly, we present a hypothesis testing approach to optimally perform subspace decomposition. In the experiments, we implement perceptual FA speech enhancement using Aurora2 corpus. We find that proposed approach achieves desirable speech recognition rates especially when signal-to-noise ratio is lower than 5 dB.

1. Introduction

Automatic speech recognition (ASR) systems have been employed to many real-world applications. However, ASR systems are always degraded in presence of different noises in practical situations. To provide good speech quality for ASR systems, the speech enhancement is an important preprocessing procedure for noisy speech recognition. In the past decade, the researchers on speech enhancement for robust ASR have been attracting many people working on this issue. Spectral subtraction algorithm [2] is one of the most popular methods for speech enhancement. This algorithm has the drawbacks of producing speech distortion and "musical noise". The method in [11] was proposed to overcome "musical noise" problem by using human auditory models where the perceptual effect of "musical noise" was reduced under predefined threshold. Below the masking threshold, the residual noise becomes inaudible by human ear. Other researchers presented subspace approaches to balance the trade off between speech distortion and residual noise [5] [8].

The general concept of subspace approaches is originated from that the noisy speech signal can be projected onto two subspaces; one is the signal subspace in which clean speech signal and few noises are included, and the other is the noise subspace that only contains noise information. In [4], Ephraim and Van Trees proposed signal subspace approach to find optimal estimator or filter by

minimizing the speech distortion subject to the constraint of residual noise kept under a threshold. This work decomposed the noisy signal into signal subspace and noise subspace by using Karhunen-Loève transform (KLT). The noisy speech signal was accordingly enhanced by using inverse KLT. Rezayee and Gazor [8] used a diagonal matrix instead of the identity matrix for finding the linear time domain constrained estimator of clean speech. Hu and Loizou [5] estimated the optimal filter by using common matrix diagonalizing the covariance matrices of the clean and noise signals.

In this paper, we are presenting a FA speech enhancement using the perceptual optimization procedure. In general, FA is a data analysis model, which is popular in societies of social science and machine learning. FA is highly related to principal component analysis (PCA) developed for feature dimension reduction. One major difference is that PCA represents the covariance or correlation matrix using singular value decomposition (SVD), whereas FA incorporates a prior structure of the residual terms. Also, the common factors extracted by FA model are useful to represent the correlation between different features [1]. The full covariance matrix can be properly modeled. Although FA generative model is new in the society of speech technology, some researchers have successfully combined FA model and hidden Markov model (HMM) for building ASR system [9]. In this paper, we present a new perceptual FA model and solution to speech enhancement. The noisy speech signal is decomposed into principal factors and minor factors, or correspondingly projected onto two subspaces. The first subspace represents the clean speech and the other subspace is a residual subspace containing noise and residual speech. The decomposition can be fulfilled via eigen-analysis for covariance matrix of speech signal. However, in conventional signal subspace approach, the smaller eigenvalues were assumed to be zero for speech enhancement. When considering FA modeling of noisy speech, the residual covariance matrix is assumed to be a diagonal matrix, which is practical for speech enhancement in presence of colored noise [8]. Furthermore, we exploit the hypothesis testing for finding the optimal FA subspace decomposition. Correspondingly, the noisy speech signal can be enhanced. Experiments on Aurora2 corpus show that the proposed FA speech enhancement approach attains good recognition performance for different cases of signal-to-noise ratio (SNR).

2. Subspace Approaches

2.1. Signal Subspace (SS)

Signal subspace is a popular speech enhancement approach using a linear model assuming that *K*-dimensional noisy observation vector $\mathbf{z}$ is corrupted in a form of

$$\mathbf{z} = W_{\mathrm{SS}} \cdot \mathbf{x}_{\mathrm{SS}} + \mathbf{n}_{\mathrm{SS}} = \mathbf{y} + \mathbf{n}_{\mathrm{SS}}, \tag{1}$$

where W_{SS} is a $K \times M$ matrix of rank M ($M < K$) with column vectors consisting of bases of a subspace of Euclidean space R^K. This is a subspace of clean speech $\mathbf{y}$. $\mathbf{x}_{\mathrm{SS}}$ denotes the

coordinate vector and $\mathbf{n}_{\mathrm{SS}}$ denotes the noise signal. This model is established assuming that noise signal is additive and uncorrelated with clean speech. The covariance matrix of $\mathbf{y}$ with rank M is given by

$$R_y = E\{\mathbf{y}\mathbf{y}^T\} = W_{\mathrm{SS}} R_{x_{\mathrm{ss}}} W_{\mathrm{SS}}^T = W_y \Lambda_y W_y^T . \tag{2}$$

Using eigen-decomposition, we obtain eigenvector matrix $W_y = [W_y^M \; W_y^{K-M}]$ and diagonal eigenvalue matrix Λ_y containing $K - M$ zero eigenvalues. The first M eigenvectors W_y^M span the same subspace as the clean speech subspace, i.e. $\mathrm{span}(W_{\mathrm{ss}}) = \mathrm{span}(W_y^M)$. To find the linear filtering for speech enhancement, it is popular to optimize the perceptually meaningful criterion, which is equivalent to minimize signal distortion while the residual noise energy is constrained under a predefined level. After solving a constrained optimization problem, we obtain the optimal solution to SS approach [4]

$$\hat{H}_{\mathrm{SS}} = R_y (R_y + \mu R_{n_{\mathrm{ss}}})^{-1} = W_y \Lambda_y (\Lambda_y + \mu W_y^T R_{n_{\mathrm{ss}}} W_y)^{-1} W_y , \tag{3}$$

where μ is the Lagrange parameter. In (3), we express the linear estimator $\hat{H}_{\mathrm{SS}}$ using eigen-decomposition of R_y.

2.2. Factor Analysis (FA)

On the other hand, FA is a general modeling approach to express an observed data vector [1]

$$\mathbf{z} = W_{\mathrm{FA}} \mathbf{x}_{\mathrm{FA}} + \mathbf{n}_{\mathrm{FA}} . \tag{4}$$

Here, the noisy speech signal $\mathbf{z}$ is considered with a preprocessing stage of mean removal. The basic idea of FA is to use a factor loading matrix W_{FA} and a common factor vector $\mathbf{x}_{\mathrm{FA}}$ to represent the observed data $\mathbf{z}$. Common factors are referred as the latent variables. The error term $\mathbf{n}_{\mathrm{FA}}$ is a specific factor representing the noise signal and/or residual speech signal. Different from principal component analysis (PCA) developed for dimension reduction, FA aims to extract the common factors for data modeling. Some properties have been specified to establish FA model. First, the observation, common factor and error term are assumed to be Gaussian distributed with zero mean $E[\mathbf{z}] = E[\mathbf{x}_{\mathrm{FA}}] = E[\mathbf{n}_{\mathrm{FA}}] = 0$. Also, common factor and error term are uncorrelated and their covariance matrices are diagonal, namely $E[\mathbf{x}_{\mathrm{FA}} \mathbf{n}_{\mathrm{FA}}^T] = 0$, $E[\mathbf{x}_{\mathrm{FA}} \mathbf{x}_{\mathrm{FA}}^T] = \mathbf{I}_M$ and

$E[\mathbf{n}_{\mathrm{FA}}\mathbf{n}_{\mathrm{FA}}^{T}]=\Psi$. For the case of isotropic noise, we have FA parameter $\Psi=\sigma^{2}\mathbf{I}_{K}$, where $\mathbf{I}_{K}$ is an $K\times K$ identity matrix. Typically, FA model in (4) is similar to the linear regression model. However, the estimation of FA and linear regression models is quite different. In linear regression model, only $\mathbf{x}_{\mathrm{LR}}$ is unknown (W_{LR} is known), whereas in FA model neither W_{FA} nor $\mathbf{x}_{\mathrm{FA}}$ are known. We should estimate FA parameters W_{FA}, $\mathbf{n}_{\mathrm{FA}}$ and later find $\mathbf{x}_{\mathrm{FA}}$. There are several approaches useful to estimate W_{FA}. One approach was derived from probabilistic PCA model [3] [10] using the maximum likelihood estimate. Nevertheless, W_{FA} can be estimated via eigen-decomposition of covariance matrix of $\mathbf{z}$

$$R_{z}=E[\mathbf{z}\mathbf{z}^{T}]=W_{\mathrm{FA}}W_{\mathrm{FA}}^{T}+\Psi=W_{z}\Lambda_{z}W_{z}^{T}= W_{z}^{M}\Lambda_{z}^{M\,1/2}\Lambda_{z}^{M\,1/2}W_{z}^{M\,T}+W_{z}^{K-M}\Lambda_{z}^{K-M}W_{z}^{K-M\,T}\quad. \tag{5}$$

where W_{z} and Λ_{z} are eigenvector and eigenvalue matrices, respectively. Through eigenvalue ordering, we obtain partitioned eigenvector matrix $W_{z}=[W_{z}^{M}\ W_{z}^{K-M}]$ and eigenvalue matrix $\Lambda_{z}=\mathrm{diag}[\Lambda_{z}^{M}\ \Lambda_{z}^{K-M}]$. Factor loading matrix W_{FA} is found using principal submatrix W_{z}^{M} and the preceding M eigenvalues in Λ_{z}. Or, we have $\mathrm{span}(W_{\mathrm{FA}})=\mathrm{span}(W_{z}^{M})$. The covariance matrix of error or noise term Ψ is generated using minor submatrix W_{z}^{K-M} and the last $K-M$ eigenvalues. Interestingly, FA parameters are estimated from two subspaces of $\mathbf{z}\in R^{K}$. FA can serve as SS approach. In what follows, we will explore the link between SS and FA for data modeling and find the solution to FA speech enhancement.

3. FA Speech Enhancement

3.1. Relation between SS and FA

Actually, the underlying concept of FA is similar to SS. Both methods decompose the signal space into two subspaces. Using FA model, the principal subspace $\mathrm{span}(W_{\mathrm{FA}})$ or $\mathrm{span}(W_{z}^{M})$ is used to represent all observed clean and noisy data. The minor subspace $\mathrm{span}(W_{z}^{K-M})$ contains the information of residual speech and noise. However, in SS approach, the signal subspace and noise subspace represent clean speech and noise signal, respectively. The linear models of SS in (1) and FA in (4) look similar. Typically, FA model is desirable for modeling full covariance or correlation matrix of observed data. After eigen-decomposition, the first M common factors have high energy. They are used for representing clean speech signal. The correlation between corresponding feature components is significant. But, the last $K-M$ common factors contain residual speech and noise signal with small energy. In SS model, the last $K-M$ eigenvectors span the noise subspace. This is the key difference between FA and SS models. To explain this property, let us use the same factor loading matrix W_{FA} and common factor $\mathbf{x}_{\mathrm{FA}}$ to express the corresponding clean speech

$$\mathbf{y}=W_{\mathrm{FA}}\mathbf{x}_{\mathrm{FA}}+\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}. \tag{6}$$

The term $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}$ means the error due to residual speech. Then, the observed noisy speech has the form

$$\mathbf{z} = W_{\mathrm{FA}}\mathbf{x}_{\mathrm{FA}} + \mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}} + \mathbf{n}_{\mathrm{FA}}^{\mathrm{n}}. \tag{7}$$

Here, the residual speech $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}$ and noise signal $\mathbf{n}_{\mathrm{FA}}^{\mathrm{n}}$ are summed up to denote the error term of noisy speech, i.e. $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}} + \mathbf{n}_{\mathrm{FA}}^{\mathrm{n}} = \mathbf{n}_{\mathrm{FA}}$. Accordingly, the covariance matrix of noisy speech turns out to be

$$\begin{aligned} R_z &= E[(W_{\mathrm{FA}}\mathbf{x}_{\mathrm{FA}} + \mathbf{n}_{\mathrm{FA}})(W_{\mathrm{FA}}\mathbf{x}_{\mathrm{FA}} + \mathbf{n}_{\mathrm{FA}})^T] \\ &= W_{\mathrm{FA}} R_{x_{\mathrm{FA}}} W_{\mathrm{FA}}^T + R_{\mathrm{rs}} + R_{\mathrm{n}} \end{aligned}. \tag{8}$$

Two covariance matrices R_{rs} and R_{n} corresponding to error variables $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}$ and $\mathbf{n}_{\mathrm{FA}}^{\mathrm{n}}$ are produced, respectively. Basically, FA is a generalized data modeling approach compared to SS.

3.2. Perceptual Criterion for Speech Enhancement

We have explained how FA is used to model noisy speech data. Under this data modeling framework, we would like to develop speech enhancement approach. Similar to SS speech enhancement, we should adopt an objective function to be optimized to estimate the clean speech signal $\hat{\mathbf{y}}$. A $K \times K$ matrix H_{FA} serves as a linear estimator or filter for speech enhancement $\hat{\mathbf{y}} = H_{\mathrm{FA}}\mathbf{z}$. The residual speech signal ε due to this estimation becomes

$$\varepsilon = \hat{\mathbf{y}} - \mathbf{y} = (H_{\mathrm{FA}} - \mathbf{I}_K)\mathbf{y} + H_{\mathrm{FA}}\mathbf{n}_{\mathrm{FA}}^{\mathrm{n}} = \varepsilon_y + \varepsilon_n, \tag{9}$$

where ε_y is the speech distortion and ε_n is the residual noise. The energies of signal distortion and residual noise are obtained by

$$\bar{\varepsilon}_y^2 = \mathrm{tr}E[\varepsilon_y^T \varepsilon_y] = \mathrm{tr}[(H_{\mathrm{FA}} - \mathbf{I}_K)R_y(H_{\mathrm{FA}} - \mathbf{I}_K)^T], \tag{10}$$

$$\bar{\varepsilon}_n^2 = \mathrm{tr}E[\varepsilon_n^T \varepsilon_n] = \mathrm{tr}[H_{\mathrm{FA}} R_{\mathrm{n}} H_{\mathrm{FA}}^T]. \tag{11}$$

Also, from (6), we calculate the covariance matrix of clean speech $\mathbf{y}$ as

$$R_y = W_{\mathrm{FA}} R_{x_{\mathrm{FA}}} W_{\mathrm{FA}}^T + R_{\mathrm{rs}}. \tag{12}$$

Notably, there is an additional term in FA model due to the residual speech. When finding FA speech enhancement solution, such generalized model should be better for estimation of clean speech. In this

study, we also take into account the auditory effects [6][7] while estimating the optimal filter. In human's auditory perception system, frequency masking is a phenomenon under which one sound can't be perceived if another sound close in frequency has a high enough level. Based on the masking effects, the residual noise is constrained to be smaller than a masking threshold rather than subtracting all noise in the noisy speech. Additionally, human is more sensitive to the distorted sound. There is a tradeoff between signal distortion and residual noise. Less residual noise will causes larger signal distortion, and the optimal filter will become an identity matrix if we enhance speech signal without distortion. According to these two properties, we adopt perceptual criterion for FA speech enhancement. Namely, we minimize the energy of speech distortion by considering the masking effect that the energy of residual noise should be controlled under a specific threshold. The objective function and constraint are given by

$$\min_{H_{\mathrm{FA}}} \bar{\varepsilon}_y^2$$

$$\text{subject to: } \bar{\varepsilon}_n^2 \le \gamma \sigma_n^2, \tag{13}$$

where $\gamma\sigma_n^2$ denotes the permissible residual noise level, σ_n^2 is a predefined noise energy, γ is an adjustable parameter which controls the masking level and that is restricted between the range of 0 and 1. The optimum linear estimator can be solved through Lagrange optimization procedure. By introducing the Lagrange multiplier μ, we can find the solution to FA speech enhancement

$$\hat{H}_{\mathrm{FA}} = (W_{\mathrm{FA}} R_{x_{\mathrm{FA}}} W_{\mathrm{FA}}^T + R_{\mathrm{rs}})(W_{\mathrm{FA}} R_{x_{\mathrm{FA}}} W_{\mathrm{FA}}^T + R_{\mathrm{rs}} + \mu R_{\mathrm{n}})^{-1}. \tag{14}$$

The key difference between SS solution in (3) and FA solution in (14) lies in the models of clean speech $\mathbf{y}$ in (1) and (6). Either $K \times M$ matrix W_{SS} or W_{FA} is not sufficient to represent clean speech signal. With an additional residual speech term $\mathbf{n}_{\mathrm{FA}}^{\mathrm{rs}}$, we are able to achieve precise data model contributed by the last $K - M$ eigenvectors. If we neglect the residual speech in FA, clean speech becomes $\mathbf{y} = W_{\mathrm{FA}} \mathbf{x}_{\mathrm{FA}}$. The covariance matrix R_{n} disappears in the solution. The FA solution is reduced to SS solution.

3.3. Optimal Subspace Decomposition

Using either FA or SS, it is critical to determine the partition of principal factors (or signal subspace) and minor factors (or noise subspace). This partition is controlled by the parameter of noise threshold σ_n^2. To significantly perform subspace decomposition, in this study, we employ hypothesis test principle to estimate optimal σ_n^2 instead of empirically assigning a value using SS approach.

Accordingly, we are not only able to determine the dimension of principal factors but also the parameters σ_n^2 without using the additional empirical parameter γ. We are testing the null hypothesis [1] that the last $K-M$ eigenvalues are equal $H_0 : \lambda_{M+1} = \lambda_{M+2} = \cdots = \lambda_K$ against the alternative hypothesis H_1 that at least two of the last $K-M$ eigenvalues are different. Assuming that eigenvalues are Gaussian distributed, we can represent the likelihood under null hypothesis as

$$L(H_0) = (2\pi)^{-\frac{N(K-M)}{2}} \left|\Lambda_2\right|^{-\frac{N}{2}} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} \Delta\mathbf{x}_i \Lambda_2^{-1} \Delta\mathbf{x}_i^T\right\}. \tag{15}$$

where $\Delta\mathbf{x}_i = \mathbf{x}_i - \overline{\mathbf{x}}$ denotes the *i*th row of $\Delta\mathbf{X}$ with $\sum_{i=1}^{N} \Delta\mathbf{x}_i \Delta\mathbf{x}_i^T = \mathrm{tr}\left[\Delta\mathbf{X}^T \Delta\mathbf{X}\right]$. Λ_2 is a diagonal matrix with its diagonal elements equal to the last $K-M$ eigenvalues and *N* is the number of training observations. $L(H_0)$ can be arranged as

$$\begin{aligned} L(H_0) &= (2\pi)^{-\frac{N(K-M)}{2}} \left(\prod_{k=M+1}^{K} \lambda_k\right)^{-\frac{N}{2}} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} (\Delta\mathbf{x}_i \Delta\mathbf{x}_i^T) \Lambda_2^{-1}\right\} \\ &= (2\pi)^{-\frac{N(K-M)}{2}} \left(\prod_{k=M+1}^{K} \lambda_k\right)^{-\frac{N}{2}} \cdot \exp\left\{-\frac{N}{2}\mathrm{tr}\left[\left(\frac{1}{N}\Delta\mathbf{X}^T \Delta\mathbf{X}\right)\Lambda_2^{-1}\right]\right\}. \\ &= (2\pi)^{-\frac{N(K-M)}{2}} \left[\left(\frac{1}{K-M}\sum_{k=M+1}^{K} \lambda_k\right)\right]^{-\frac{N}{2}} \cdot \exp\left\{-\frac{N}{2}\mathrm{tr}\left(\Lambda_2 \Lambda_2^{-1}\right)\right\} \end{aligned} \tag{16}$$

Similarly, the likelihood under alternative hypothesis is yielded by

$$\begin{aligned} L(H_1) &= (2\pi)^{-\frac{N(K-M)}{2}} \left|\Lambda_2\right|^{-\frac{N}{2}} \cdot \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\sum_{i=1}^{N} (\Delta\mathbf{x}_i \Delta\mathbf{x}_i^T) \Lambda_2^{-1}\right]\right\} \\ &= (2\pi)^{-\frac{N(K-M)}{2}} \left(\prod_{k=M+1}^{K} \lambda_k\right)^{-\frac{N}{2}} \cdot \exp\left\{-\frac{N}{2}\mathrm{tr}\left(\Lambda_2 \Lambda_2^{-1}\right)\right\} \end{aligned}. \tag{17}$$

We evaluate the likelihood ratio q of $L(H_0)$ to $L(H_1)$. The resulting test statistic q has the form of

$$q = \frac{L(H_0)}{L(H_1)}$$

$$= \frac{(2\pi)^{-\frac{N(K-M)}{2}} \left[\left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M} \right]^{-\frac{N}{2}} \cdot \exp\left\{ -\frac{N}{2} \mathrm{tr}\left(\Lambda_2 \Lambda_2^{-1} \right) \right\}}{(2\pi)^{-\frac{N(K-M)}{2}} \left(\prod_{k=M+1}^{K} \lambda_k \right)^{-\frac{N}{2}} \cdot \exp\left\{ -\frac{N}{2} \mathrm{tr}\left(\Lambda_2 \Lambda_2^{-1} \right) \right\}} . \tag{18}$$

$$= \left[\frac{\prod_{k=M+1}^{K} \lambda_k}{\left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M}} \right]^{\frac{N}{2}}$$

Then, the distribution of statistic $-2\log q$ turns out to be a χ^2 distribution

$$-2\log q = -2\log \left[\frac{\prod_{k=M+1}^{K} \lambda_k}{\left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M}} \right]^{\frac{N}{2}}$$

$$= -N \cdot \log \left[\frac{\prod_{k=M+1}^{K} \lambda_k}{\left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M}} \right]$$

$$= -N \cdot \left[\log \prod_{k=M+1}^{K} \lambda_k - \log \left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right)^{K-M} \right] . \tag{19}$$

$$= N \cdot \left[(K-M) \log \left(\frac{1}{K-M} \sum_{k=M+1}^{K} \lambda_k \right) - \sum_{k=M+1}^{K} \log \lambda_k \right]$$

$$= N \cdot \left[(K-M) \log \bar{\lambda} - \sum_{k=M+1}^{K} \log \lambda_k \right] \sim \chi^2_{(\nu)}$$

Finally, we find that null hypothesis H_0 is rejected at a significance level α if

$$N \cdot \left[(K-M) \log \bar{\lambda} - \sum_{k=M+1}^{K} \log \lambda_k \right] \geq \chi^2_{\nu;\alpha} . \tag{20}$$

In (20), $\bar{\lambda}$ is a sample mean of eigenvalues, and ν is the degree of freedom of χ^2 distribution.

4. Experiments

4.1. Speech Database and Experimental Setup

We performed speech recognition and SNR calculation using Aurora2 database for evaluating performance of proposed speech enhancement methods. Aurora2 database consisted of English digits and English alphabet-sequence in the presence of additive noise and linear convolutional distortion. There were three test sets in the corpus. Set A had four noise types (subway, babble, car and exhibition hall) that were similar to those in the training data, and set B contained four noise types (restaurant, street, airport and station noise) different from those in the training data. An additional convolutional channel was used in set C. All these three test sets consisted of six SNR conditions (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB) and clean condition. Acoustic models in clean and multi-conditional noise conditions were estimated for comparison. There were 8,440 clean training utterances. The multi-conditional training data consisted of the same utterances artificially added with four different noise types (subway, babble, car and exhibition hall) in five different environmental conditions (5 dB, 10 dB, 15 dB, 20 dB and clean).

Speech features consisted of 13 MFCC coefficients and energy along with the delta and acceleration coefficients. There were 39 features extracted for each frame. In this paper, we estimated continuous-density hidden Markov models (HMM's) and built the recognizer using HTK toolkit package [12]. Some parameters were used: 1) 16 states per word; 2) 3 mixture components of Gaussian density per state; 3) only the variances of all acoustic coefficients are used; 4) optimal subspace decomposition was done by performing hypothesis testing frame by frame and significance level α was set to be 0.05 in multi-condition training and 0.02 in clean training. In speech enhancement procedure, we used 40 sampling point for a frame. The filter of 40×40 matrix was estimated. When computing the covariance matrix, a window of 9 frames was used. The control parameter μ was dynamically specified according to the SNR measured in each frame. Larger μ corresponded to smaller residual noise and larger signal distortion. In the experiments, we preset the range $\mu = 0 \sim 4$.

4.2. Evaluation of SNR Performance

In this subsection, we collected test sets containing six SNR conditions in Aurora2 database for evaluation of SNR's when applying FA speech enhancement. Assuming clean speech and noise signals are independent, SNR formula is calculated by

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{t=1}^{T}\sum_{k=1}^{K} y_t^2(k)}{\sum_{t=1}^{T}\sum_{k=1}^{K}\left(z_t(k)-y_t(k)\right)^2}\times 100\%. \tag{21}$$

where y is clean speech signal and z is noisy speech signal. In Figure 1, the SNR's defined in Aurora2 are similar to those calculated by (21). When applying FA speech enhancement, we find that SNR's are significantly improved for different SNR conditions. The SNR evaluation shows that FA

approach does suppress the noise level. Such suppression does not assure small distortion of speech signal itself. Namely, over suppression of noise in noisy speech will cause the series distortion of speech signal at the same time. To verify the effectiveness of using FA approach, we further conduct experimental comparison for the application of noisy speech recognition using Aurora 2 database.

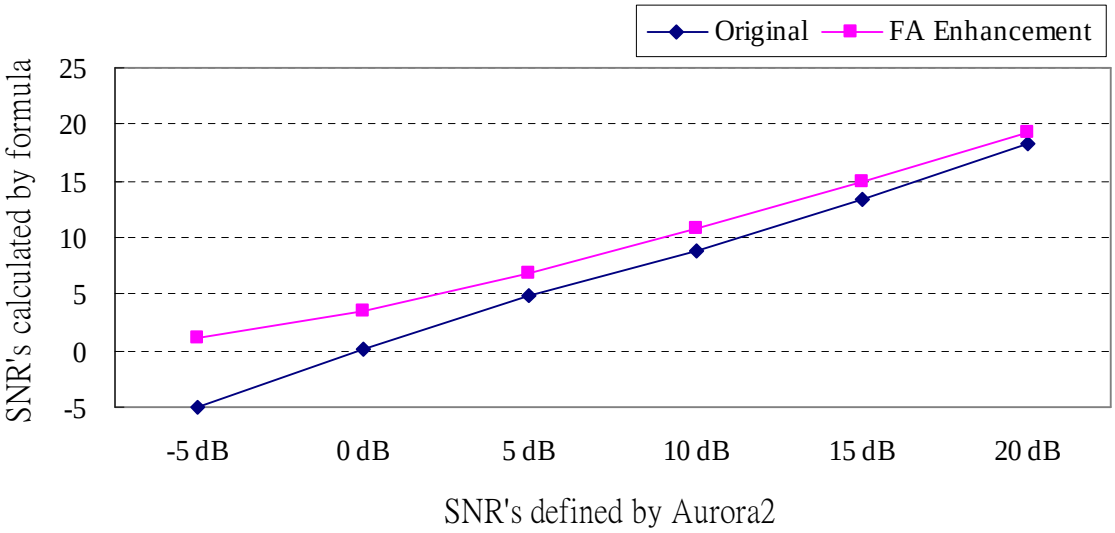


Figure 1. SNR improvement using FA speech enhancement

4.3. Evaluation of Speech Recognition Performance

Table 1 reports the baseline word accuracy (%) results of using clean training set in Aurora2, Tables 2 and 3 show the results after enhancement using signal subspace (SS) and factor analysis (FA) enhancement. On average, in Tables 1 and 3, the relative word error rate is improved by 14.88%, specifically in 0 dB (29.01%), -5 dB (24.88%), and 5 dB (18.18%). Error reduction is achieved for cases of Car (26.01%), Subway (19.47%), and Station (19.34%) environments. Figure 1 shows the performances for baseline system and two enhancement approaches in different environments.

Table 1. Baseline results for clean training

Aurora 2 Clean Training - Results (Baseline)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	98.83	98.61	98.78	99.01	98.81	99.29	99.24	99.25	99.48	99.32	99.42	99.00	99.21	99.09
20 dB	97.61	97.97	98.51	97.35	97.86	98.43	97.52	98.24	98.95	98.29	94.11	94.89	94.50	97.36
15 dB	95.09	94.20	95.71	94.82	94.96	95.36	94.35	95.32	95.68	95.18	87.96	89.45	88.71	93.79
10 dB	84.83	81.38	80.55	84.70	82.87	85.63	82.16	83.21	82.66	83.42	75.74	76.45	76.10	81.73
5 dB	63.77	59.37	48.97	56.09	57.05	63.31	54.23	60.72	54.89	58.29	54.59	51.72	53.16	56.77
0 dB	35.12	35.04	22.70	25.15	29.50	37.21	28.48	35.88	26.94	32.13	29.44	26.90	28.17	30.29
-5 dB	15.08	19.17	11.09	12.47	14.45	18.36	15.30	18.19	14.56	16.60	14.09	13.75	13.92	15.21
average	70.05	69.39	65.19	67.08	67.93	71.08	67.33	70.12	67.59	69.03	65.05	64.59	64.82	67.75

Table 2. Signal subspace enhancement for clean training

Aurora 2 Clean Training - Results (SS Enhancement)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	99.26	99.06	99.14	99.51	99.24	99.26	99.24	99.14	99.51	99.29	99.32	99.00	99.16	99.24
20 dB	98.28	97.40	98.09	97.81	97.90	97.97	97.67	97.55	97.55	97.69	97.42	96.55	96.99	97.63
15 dB	95.52	92.87	96.69	94.48	94.89	94.32	93.23	94.18	94.18	93.98	95.33	93.74	94.54	94.45
10 dB	87.69	81.89	90.34	85.34	86.32	84.07	83.92	84.46	84.46	84.23	87.81	83.65	85.73	85.36
5 dB	74.70	64.06	74.59	62.94	69.07	65.18	67.02	67.52	67.52	66.81	73.07	65.63	69.35	68.22
0 dB	52.16	39.84	45.45	38.29	43.94	41.23	39.02	43.66	43.66	41.89	46.05	39.21	42.63	42.86
-5 dB	28.31	21.49	20.34	16.11	21.56	17.93	16.84	21.47	21.47	19.43	19.62	16.14	17.88	19.97
average	76.56	70.94	74.95	70.64	73.27	71.42	70.99	72.57	72.62	71.90	74.09	70.56	72.32	72.53

Table 3. Factor analysis enhancement for clean training

Aurora 2 Clean Training - Results (FA Enhancement)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	99.32	99.03	99.19	99.48	99.26	99.32	99.21	99.22	99.54	99.32	99.39	98.94	99.17	99.26
20 dB	98.31	97.88	98.45	97.84	98.12	98.25	97.85	98.06	98.80	98.24	97.54	96.80	97.17	97.98
15 dB	95.95	93.86	96.72	95.62	95.54	95.15	94.77	95.14	95.99	95.26	95.30	93.86	94.58	95.24
10 dB	89.04	83.83	91.47	88.98	88.33	86.06	85.73	85.45	86.76	86.00	88.21	83.49	85.85	86.90
5 dB	77.04	66.08	76.53	68.22	71.97	68.25	69.80	69.79	72.72	70.14	74.74	66.81	70.78	71.00
0 dB	55.76	42.38	47.27	43.26	47.17	44.34	41.05	44.23	41.59	42.80	47.44	40.93	44.19	44.83
-5 dB	28.80	23.88	18.67	17.49	22.21	23.18	19.14	24.31	20.55	21.80	20.36	17.32	18.84	21.37
Average	77.75	72.42	75.47	72.98	74.66	73.51	72.51	73.74	73.71	73.37	74.71	71.16	72.94	73.80

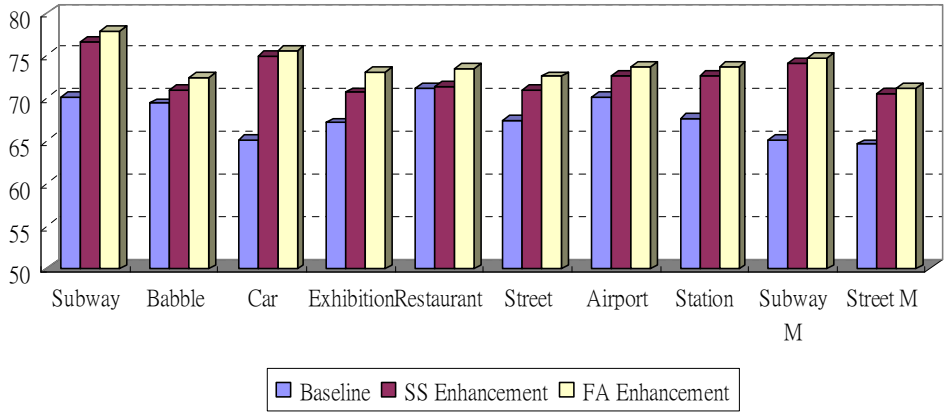


Figure 2. Comparison of different environments for clean training

In Figure 2, we find that the proposed FA enhancement performs better than SS enhancement especially in presence of larger background human voices, e.g. the noise conditions of exhibition and restaurant. Experimental results using multi-condition training for baseline system, signal subspace and factor analysis enhancement are also reported in Tables 4, 5 and 6, respectively.

Table 4. Baseline results for multi-condition training

Aurora 2 Multicondition Training - Results (Baseline)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	99.02	98.73	98.88	99.01	98.91	99.02	98.91	98.87	99.01	98.95	98.86	98.85	98.86	98.92
20 dB	98.43	98.04	98.27	97.99	98.18	98.71	98.46	98.63	98.83	98.66	97.76	97.64	97.70	98.28
15 dB	97.42	97.82	97.88	97.16	97.57	98.43	97.28	98.21	98.46	98.10	96.96	96.31	96.64	97.59
10 dB	95.00	96.43	95.97	94.17	95.39	96.90	95.89	96.96	97.04	96.70	94.20	93.68	93.94	95.62
5 dB	89.28	89.48	88.07	88.06	88.72	91.31	88.88	92.48	89.63	90.58	82.65	83.22	82.94	88.31
0 dB	67.39	65.69	53.50	64.89	62.87	71.26	66.60	72.08	64.49	68.61	47.44	56.65	52.05	63.00
-5 dB	25.15	30.14	19.59	24.56	24.86	37.95	30.59	35.73	27.21	32.87	18.36	26.00	22.18	27.53
average	81.67	82.33	78.88	80.83	80.93	84.80	82.37	84.71	82.10	83.49	76.60	78.91	77.76	81.32

Table 5. Signal subspace enhancement for multi-condition training

Aurora 2 Multicondition Training - Results (SS Enhancement)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	98.86	98.85	98.84	99.04	98.90	98.86	98.97	98.84	99.04	98.93	98.68	98.94	98.81	98.89
20 dB	98.77	98.25	98.60	98.09	98.43	98.80	98.43	98.72	98.95	98.73	98.00	98.04	98.02	98.47
15 dB	97.45	97.73	98.18	97.25	97.65	98.25	97.46	98.24	98.52	98.12	96.99	96.40	96.70	97.65
10 dB	96.28	95.37	96.48	95.34	95.87	95.43	95.77	96.60	96.82	96.16	95.52	94.07	94.80	95.77
5 dB	90.54	89.45	91.74	90.03	90.44	86.61	88.94	91.23	90.74	89.38	87.96	86.19	87.08	89.34
0 dB	71.72	66.75	74.89	67.23	70.15	70.03	73.43	76.44	78.46	74.59	72.46	60.37	66.42	71.18
-5 dB	49.19	35.07	50.40	47.82	45.62	41.93	43.77	44.94	45.82	44.12	41.20	36.67	38.94	43.68
average	86.12	83.07	87.02	84.97	85.29	84.27	85.25	86.43	86.91	85.72	84.40	81.53	82.96	85.00

Table 6. Factor Analysis enhancement for multi-condition training

Aurora 2 Multicondition Training - Results (FA Enhancement)														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall
Clean	98.99	98.70	98.87	99.04	98.90	98.99	99.03	98.87	99.07	98.99	98.86	98.97	98.92	98.94
20 dB	98.53	98.22	98.36	98.06	98.29	98.80	98.52	98.72	99.01	98.76	98.13	98.04	98.09	98.44
15 dB	97.85	97.52	98.30	97.19	97.72	98.40	97.52	98.33	98.58	98.21	96.93	96.70	96.82	97.73
10 dB	96.84	96.16	96.99	95.28	96.32	96.53	96.13	96.99	97.44	96.77	95.73	94.56	95.15	96.27
5 dB	91.00	90.60	92.54	89.95	91.02	89.59	90.57	92.25	92.22	91.16	88.30	87.30	87.80	90.43
0 dB	77.03	68.26	79.51	67.48	73.07	73.75	75.18	78.97	78.68	76.65	73.75	64.72	69.24	73.73
-5 dB	50.51	41.38	50.43	47.95	47.57	44.89	45.50	48.23	47.36	46.50	41.97	37.36	39.67	45.56
Average	87.25	84.41	87.86	84.99	86.13	85.85	86.06	87.48	87.48	86.72	84.81	82.52	83.67	85.87

When looking at Tables 4 and 6, we find that the relative word error rate is improved by 22.15%. The performances are improved in 5 dB (32.92%), 10 dB (28.30%), 20 dB (23.47%) and 15 dB (23.24%). The error reduction is obtained for Car (29.08%), Subway (28.72%), and Exhibition (23.85%) environments. The relative improvement percentage at -5 dB SNR in clean training (7.27%) is much less than that in multi-condition training (24.88%). This is because that well-trained clean models could not predict unknown noise influence. Performances for baseline and two enhancement approaches in different environments using multi-condition training are also shown in Figure 2.

From the experiments, we find that in noise environments of subway, car, and station, the speech recognition improvements were larger compared to other noise environments. This is because that machine noises are quite different from human voice noises. More human sound in background noise obtains fewer improvement using proposed enhancement methods. In this work, we evaluate the performances of FA enhancement using SNR measures and speech recognition rates. The experiments

show that FA enhancement is better than SS enhancement. This implies that a generalized model is desirable for representing signals. Moreover, the optimal subspace decomposition by hypothesis testing is used frame by frame to find the optimal filter with suitable dimensions of residual subspace. Such technique is also beneficial to obtain better performance than SS enhancement.

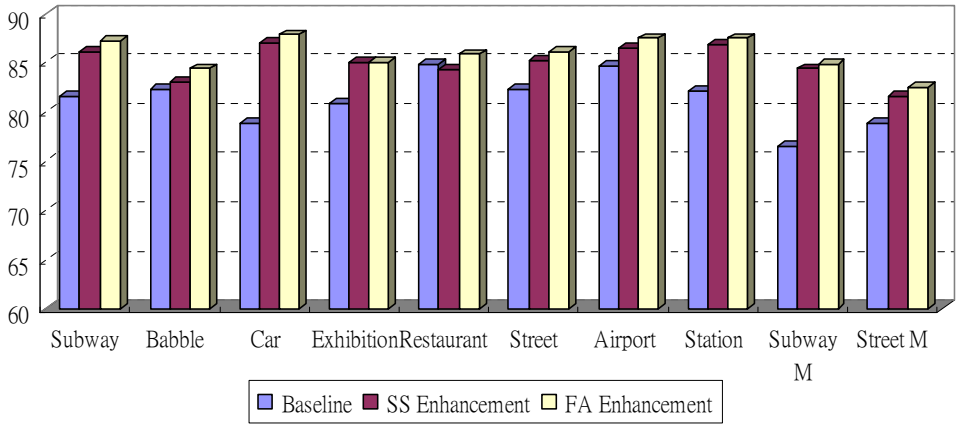


Figure 3. Comparisons for different environments (multi-condition training)

5. Conclusions

In this paper, we have presented FA enhancement of noisy speech signal for application of speech recognition. Interestingly, we built the bridge between FA and signal subspace approaches. Experimental results showed that the proposed approach improved the performance of ASR systems especially in low SNR environments. Compared to other subspace approaches, we presented a novel hypothesis testing approach to optimally perform subspace decomposition. In the future, we will extend this speech enhancement approach by considering the phase effect. Also, we will also improve FA framework through developing new estimation criteria for factor loading matrix. Additionally, we will also explore FA approaching to other speech related applications, e.g. speaker adaptation and acoustic modeling.

6. References

[1] B. Alexander, *Statistical factor analysis and related methods*, John Wiley & Sons, Inc., 1994.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[3] J.-T. Chien and C.-W. Ting, "Speaker identification using probabilistic PCA model selection", *Proc. of ICSLP*, vol. 3, pp. 1785-1788, 2004.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, 1995.

[5] Y. Hu, and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334-341, 2003.

[6] F. Jabloun and B. Champagne, "A Perceptual Signal Subspace Approach for Speech Enhancement in Colored Noise", *Proc. of ICASSP*, vol. 1, pp. 569-572, 2002.

[7] F. Jabloun and B. Champagne, "On the use of masking properties of the human ear in the signal subspace speech enhancement approach," *Proc. Int. Workshop Acoust. Echo Noise Control, Darmstadt, Germany*, pp. 199–202, 2001.

[8] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87-95, 2001.

[9] L. K. Saul, M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 115-125, 2000.

[10] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers", *Neural Computation*, vol. 11, pp. 443-482, 1999.

[11] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Transactions on Speech and Audio Processing*, vol. 7, no.2, pp. 126-137, 1999.

[12] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, The HTK Book, Cambridge University Speech Group, 2000.

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。

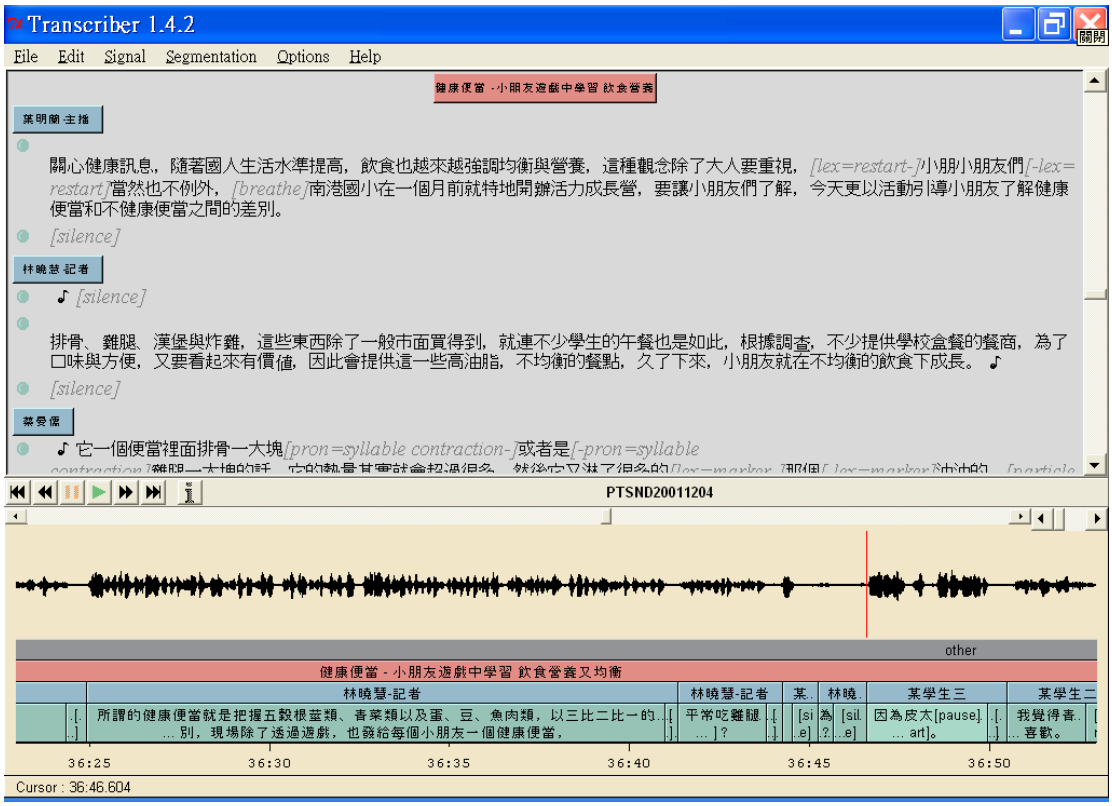
表一、MATBN 廣播語料庫之統計。

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。

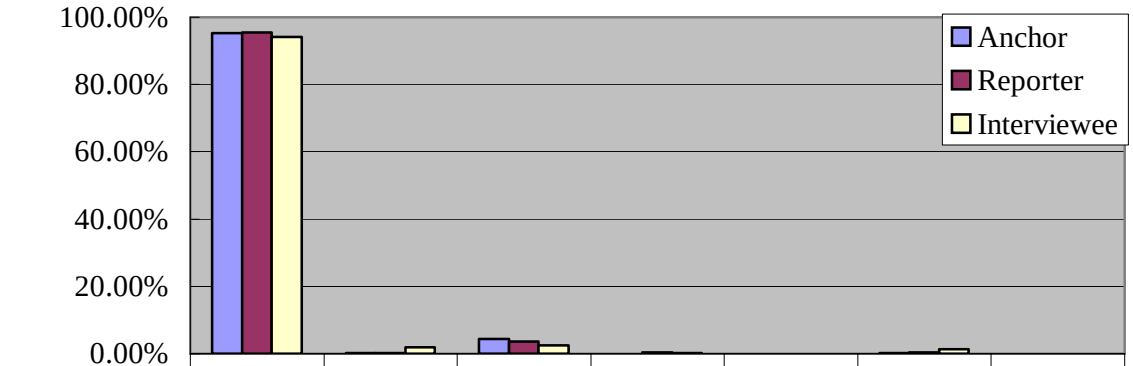
(3) Pronunciation Error

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%

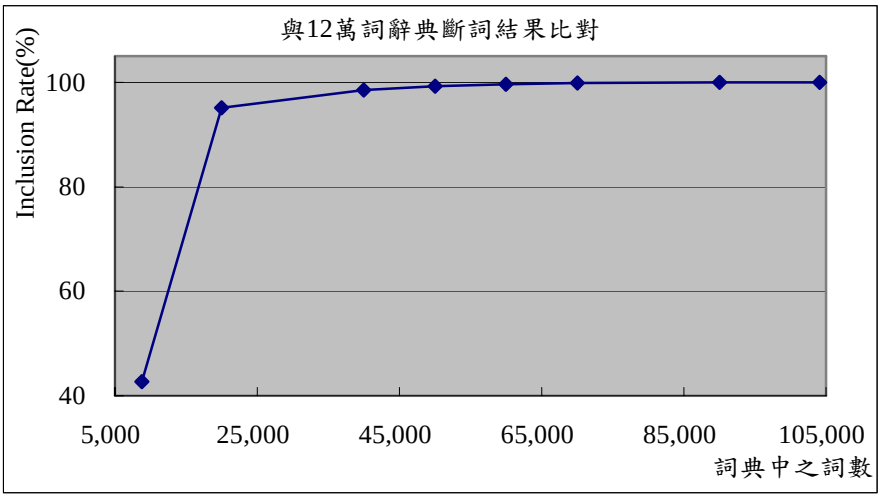
事實上，我們也做過未知環境之辨認實驗，對 anchor 及 reporter 之辨認率會下降 0.1%，而 interviewee 的之 411 音節辨認率還會提升 2%；而辨認爲正確環境的比例對 anchor/reporter/interviewee 環境分別爲 95%、93%及 83%。Interviewee 環境辨認率較低而音節辨認率會提高是因爲有些 interviewee 的錄音環境與 reporter 較爲相似，而使用 reporter 環境所建立之辨認器會得到較佳之辨認率。因爲未知環境之辨認所需之計算量非分之大，且辨認率相差不大，所以接著的實驗中均假設已知語者環境。

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。

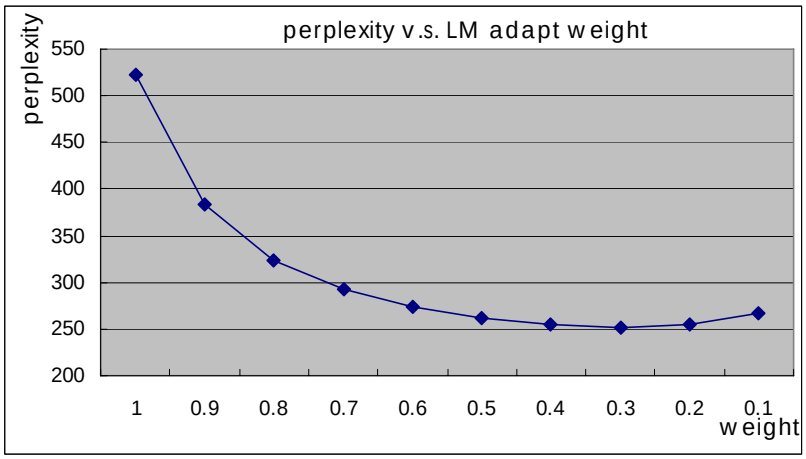
訓練語料	詞數（Word）	字數（Character）
光華雜誌	9,870,430	16,406,485
中文資訊檢索標竿測試集	124,442,861	206,847,107
平衡語料庫	4,796,163	7,972,113
合計	139,109,455	231,225,705

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：

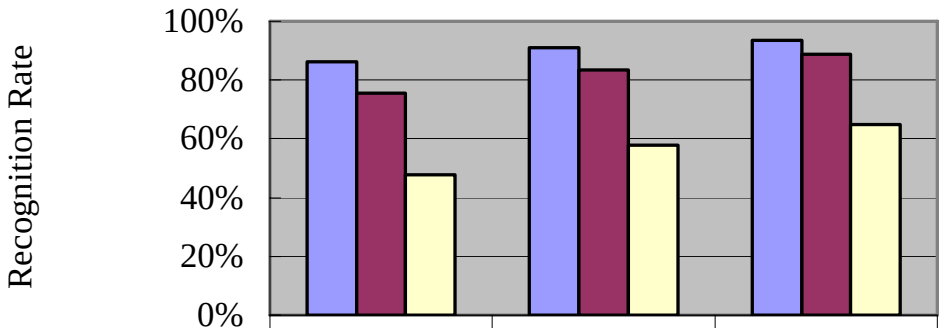
$$\log\left(P_{new}(w_i \mid w_{i-1} = Big5_j)\right)$$
$$= \log\left(P(w_i \mid w_{i-1} = Big5_j)\right) + \log\left(P(pinyin \mid big5_j)\right)$$

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。

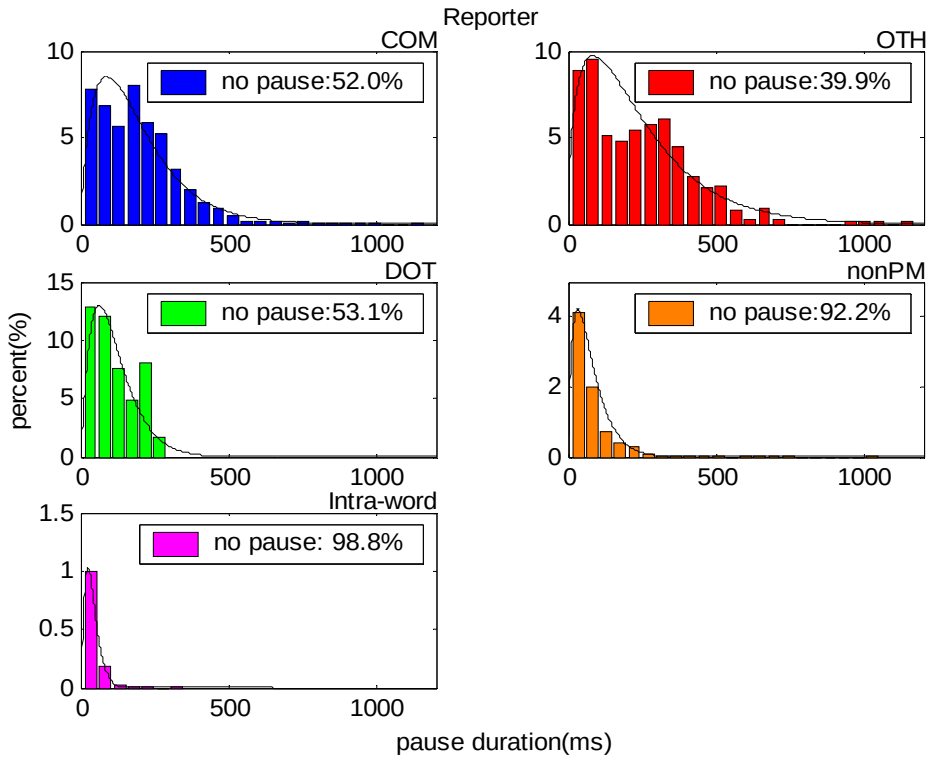
在論文中我們使用 Gamma distribution 來描述音節間靜音長度，但由表九結果得知，每種情形均有音節間靜音長度長度爲 0 情況，而且佔有相當程度的比重，所以最後使用下列機率來描述音節間靜音長度之分佈：

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

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

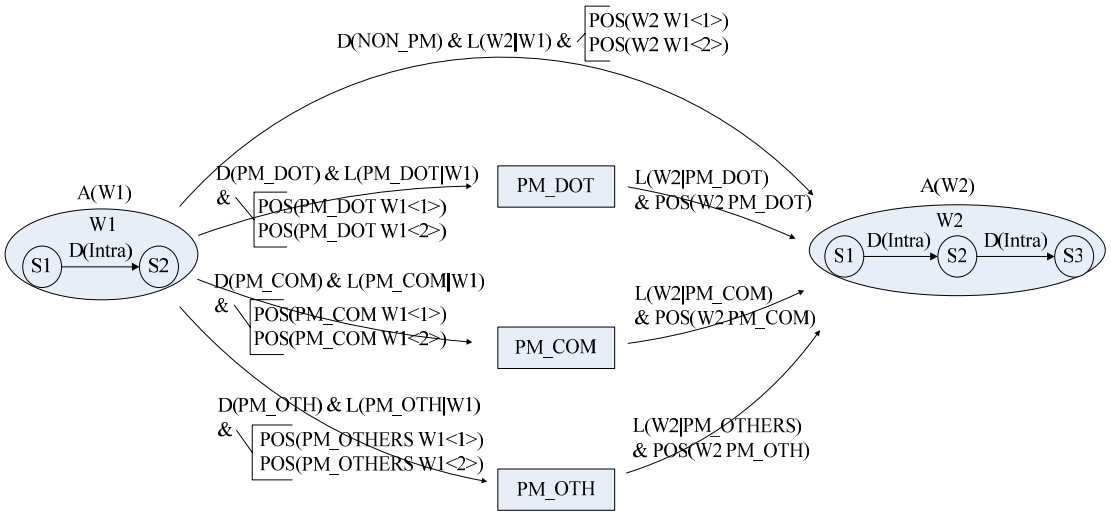
在此僅列出 reporter 環境下音節間靜音長度及其是否爲標點符號或詞邊界之關係，如圖五所

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)。

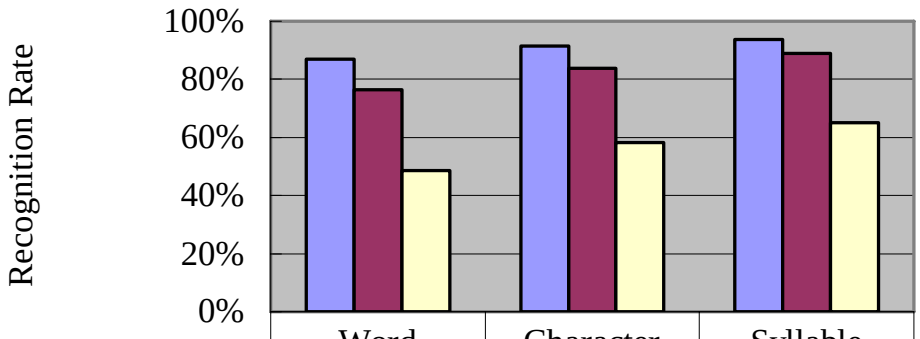
在加入音節間靜音長度之機率及標點符號之語言模式後，在辨認時 syllable 及 word 轉換時所要考慮的辨認分數可由圖五所示，在音節間轉移時須加上不同的音節間靜音長度機率分數，若是詞間轉移還需加上語言模型分數。



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.

An Approach of Using the Web as a Live Corpus for Spoken Transliteration Name Access

Ming-Shun Lin, Chia-Ping Chen, Hsin-Hsi Chen***

Department of Computer Science and Information Engineering
National Taiwan University, Taipei 106
TAIWAN
{d91022 , **hhchen}@csie.ntu.edu.tw

Department of Computer Science and Engineering
National Sun Yat-Sen University
70, Lien-Hai Road, Kaohsiung, Taiwan 804
*cpchen@cse.nsysu.edu.tw

Abstract

Recognizing transliteration names is challenging due to their flexible formulation and coverage of a lexicon. This paper employs the Web as a huge-scale corpus. The patterns extracted from the Web are considered as a live dictionary to correct speech recognition errors. In our approach, the plausible character strings recognized by ASR (Automated Speech Recognition) are regarded as query terms and submitted to Google. The top *n* returned web page summaries are entered into PAT trees. The terms of the highest scores are selected. Total 100 Chinese transliteration names, including 50 person names and 50 location names, are used as test data. In the ideal case, we input the correct syllable sequences, convert them to text strings and test the recovery capability of using Web corpus. The results show that both the recall rate and the MRR (Mean Reciprocal Rank) are 0.94. That is, the correct answers appear in the top 1 position in 94 cases. When a complete transliteration name recognition system is evaluated, the experiments show that ASR model with a recovery mechanism can achieve 3.82% performance increases compared to ASR only model on character level. Besides, the recovery capability improves the average ranks of correct transliteration names from the 18th to the 3rd positions on word level.

1. Introduction

Named entities [1], which denote persons, locations, organizations, etc., are common foci of searchers. Capturing named entities is challenging due to their flexible formulation and up-to-date use. The issues behind speech recognition make named entity recognition more challenging on spoken level than on written level. This paper emphasizes on a special kind of named entities, called transliteration names. They denote foreign people, places, etc. Spoken transliteration name recognition is useful for many applications. For example, cross language image retrieval via spoken query aims to employ spoken queries in one language to retrieve images with captions in another language [2].

In the past, Appelt and Martin [3] adapted TextPro system for processing text to processing transcripts generated by a speech recognizer. Miller et al [4] analyzed the effects of out-of-vocabulary errors and loss of punctuation in name finding of automatic speech recognition. Huang and Waibel [5] proposed an adaptive

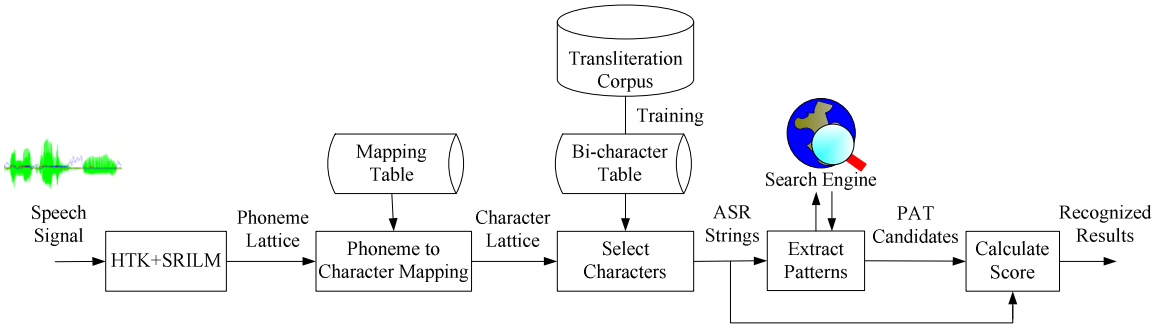


Figure 1. Flow of transliteration name recognition.

method of named entity extraction for meeting understanding. Chen [6] dealt with spoken cross-language access to image collection. The coverage of a lexicon is one of the major issues in spoken transliteration name access. Recently, researchers are interested in exploring the Web, which provides huge-collection of up-to-date data, as a corpus. Keller and Lapata [7] employed the Web to obtain frequencies for bigrams that are unseen in a given corpus.

In this paper, we consider the Web as a live dictionary for recognizing spoken transliteration names, and employ fuzzy search capability of Google to retrieve relevant web page summaries. Section 2 sketches the overall flow of our method. Section 3 employs PAT trees to learn patterns from the Web dynamically and correct the recognition errors. Section 4 shows the experiments with/without the uses of the Web. Section 5 concludes the remarks.

2. Flow of transliteration name recognition

A spoken transliteration name recognition system shown in Figure 1 accepts a speech signal denoting a foreign named entity, and converts it into a character string. It is composed of the following four major stages. Stages (1) and (2) are fundamental tasks of speech recognition. Stages (3) and (4) try to correct the speech-to-text errors by using the Web.

(1) At first, we employ the speech recognition models built by HTK (http://htk.eng.cam.ac.uk/) and SRILM (http://www.speech.sri.com/projects/srilm/) toolkits to get a syllable lattice of a speech signal.

(2) Then, the syllable lattice is mapped into a character lattice using a mapping table. Top-n character strings are selected from the character lattice using bi-character model trained from a transliteration name corpus. The character strings are called ASR strings hereafter.

(3) Next, each ASR string is regarded as a query, and is submitted to a web search engine like Google. From the top-m search result summaries of a query (i.e., an ASR string), the higher frequent patterns similar to the ASR string are considered as candidates. Because we employ PAT tree [8, 11] to extract patterns, the patterns are called PAT candidates hereafter. For PAT tree example, “湯姆漢克斯湯姆克魯斯喬治克魯尼” with MS950 encode, be shown in Figure2. The circle represents semi-infinite string number. The number located over the circle is length. The length indicates the first different bit of the character strings recorded in the sub-trees. In this example, the highest length patterns are “克魯” and “湯姆” on the nodes (7, 12) and (0, 5)

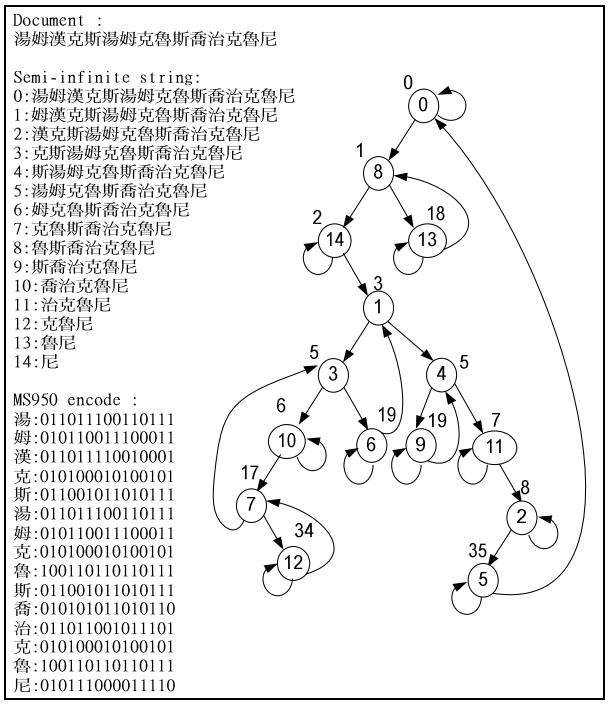


Figure 2. An example for extracting highest length patterns and its frequency.

with length 33 and 34 bits. The second highest length patterns are “克”,”魯”,”姆” and “斯” on nodes (3, 7, 12),(8, 13),(1, 6) and (4, 9) with length 16,17,18 and 18 bits. The pattern extraction task will be discussed in detail in Section 3.

(4) Finally, the PAT candidates of all ASR strings will be merged together, and ranked by their number of occurrences and similarity scores. Candidates of the best ranks will be regarded as recognition results of a spoken transliteration name.

Consider an example shown in Figure 3. The Chinese speech signal is a transliteration name “湯姆克魯斯” in Chinese denoting a famous movie star “Tom Cruise”. Syllable lattice illustrates different combinations of syllables. Each syllable corresponds to several Chinese characters. For example, ke is converted to “克”, “柯”, “科”, “可”, “喀”, “刻”, etc. ASR strings “塔莫克魯斯”, “塔門克魯斯”, “塔莫柯魯斯”, etc. are selected from character lattice. Through Google fuzzy search using query “塔莫克魯斯”, some summaries of Chinese web pages are reported in Figure 4. Although common transliteration of “Tom Cruise” in Chinese is “湯姆克魯斯”, which is different from the query “塔莫克魯斯”, fuzzy matching by Google can still identify the relevant summaries containing the correct transliteration. We call this operation recognition error recovery using the Web hereafter.

In the above examples, partial matching part is enclosed in rectangle symbol, e.g., “克魯斯”, and the correct transliteration name is underlined, e.g., “湯姆克魯斯”. Summaries (1), (4) and (5), mention a movie star “湯姆克魯斯” (Tom Cruise), and summaries (2) and (3) mention a football star “克魯斯” (Cruz). Figure 3 shows that the PAT patterns like “聖塔克魯斯”, “湯姆克魯斯”, “姆克魯斯演”, etc. are proposed. After merging and ranking, the possible recognition results in sequence are “湯姆克魯斯”, “洛普克魯茲”, “聖塔克魯茲”, etc.

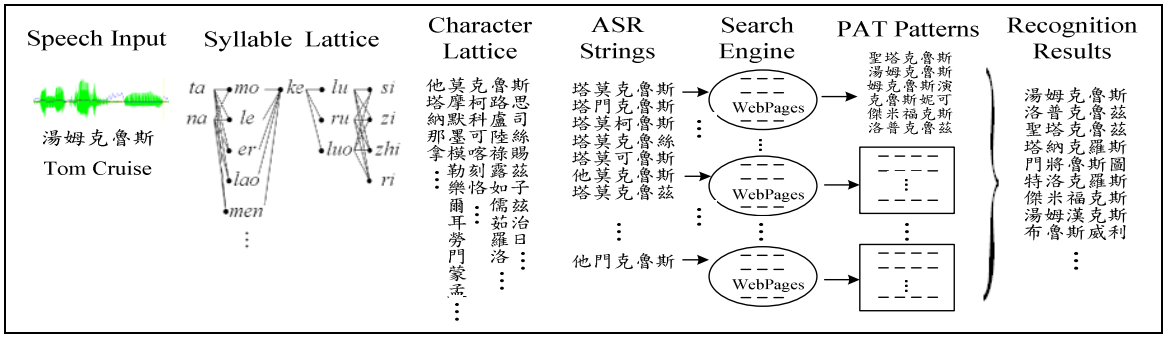


Figure 3. An example for recognizing transliteration name "湯姆克魯斯" ("Tom Cruise").

(1) … 至於贏家部份，則還是湯姆漢克斯、湯姆克魯斯、喬治克魯尼這些老面孔，…

(2) … 第 76 分鐘，克魯斯換下梅開二度的維埃裏。

(3) … 國際米蘭(4-4-2)：豐塔納/科爾多瓦，布爾迪索，馬特拉齊，法瓦利/斯坦科維奇，貝隆，扎內蒂，埃姆雷/克魯斯，馬丁斯。

(4) … 提起妮可即發火湯姆克魯斯"想殺死記者".

(5) … 電影節最具看點的明星當然非妮可基德曼與湯姆克魯斯有望在水城的戲劇性重逢莫屬。

Figure 4. Summaries of fuzzy search for query "塔莫克魯斯"

3. Recognition error recovery using the Web

The error recovery module tries to select the higher frequent pattern from the Web search results, and substitute the speech recognition results of Stages 1 and 2 (shown in Section 2) with the pattern. PAT tree [8,11], which was derived from Patricia tree, can be employed to extract word boundary and key phrases automatically. In this paper, the Web search results of an ASR string will be placed in a PAT tree and PAT candidates will be selected from the tree. Two issues are considered. A PAT candidate should occur more times in the PAT tree and should be similar to the ASR string.

The frequency Freq of a PAT candidate can be computed easily from PAT tree structure. The similarity of a PAT candidate and an ASR string is modeled by edit distance, which is minimum number of insertions, deletions and substitutions to transform one character string (ASR) into another string (PAT). The less the number is, the more similar they are. The similarity Sim of ASR and PAT strings is the length of string alignment minus the number of edit operations.

Finally, the ranking score of a PAT string relative to an ASR string is defined as follows.

$$Score(ASR, PAT) = \alpha \times Freq(PAT) + \beta \times Sim(ASR, PAT)$$

It is computed by weighted merging of the frequency of the PAT string, and the similarity of the ASR string and PAT string. This value determines if the ASR string will be replaced by the PAT string. In the above example, Freq(湯姆克魯斯)=43 and Sim(塔莫克魯斯, 湯姆克魯斯)=3.

4. Experimental results

The speech input to the transliteration name recognition system is Chinese utterance. We employed 51,114 transliteration names [9] to train the bi-character model specified in Section 2. In the experiments, the test data include 50 American state names, 29 movie star names from 31st annual people's choice awards (http://www.pcavote.com), and 21 NBA star names from NBA 2005 all star (http://www.nba.com/allstar2005/). The test set is different from the training set and it is open test. Because there may be more than one transliteration for a foreign named entity, the answer keys are manually prepared and checked with respect to the Web. For example, "Arizona" has four possible transliterations in Chinese – say, "亞利桑納", "亞歷桑納", "亞利桑那", and "亞歷桑那". On the average, there are 1.9 Chinese transliterations for a foreign name in our test set. In appendix A lists the name test set and its answer keys. As shown in Section 2, the transliteration name recognition system is composed of four major stages. Stages 1 and 2 perform the fundamental speech recognition task, and Stages 3 and 4 perform the error recovery task. To examine the effects of these two parts, we evaluate them separately and wholly in the following two subsections.

4.1 Performance of error recovery task

Assume correct syllables have been identified in speech recognition task. We simulate the cases by transforming all the characters in the answer keys back to syllables. Then, Stage 2 maps the syllable lattice into character lattice. Total 50 ASR strings are extracted from character lattice at stage 2, and submitted to Google. Finally, the best 10 PAT candidates are selected. We use recall rate and MRR (Mean Reciprocal Rank) [10] to evaluate the performance. Recall rate means how many transliteration names are correctly recognized. MRR defined below means the average ranks of the correctly identified transliteration names in the proposed 10 PAT candidates.

$$MRR = \frac{1}{M}\sum_{i=1}^{M} r_i \qquad (1)$$

, where $r_i = 1/rank_i$ if $rank_i > 0$; and r_i is 0 if no answer is found, and M is total number of test cases. The $rank_i$ is the rank of the first right answer of the i^{th} test case. That is, if the first right answer is rank 1, the score is 1/1; if it is at rank 2, the score is 1/2, and so on. The value of MRR is between 0 and 1. The inverse of MRR denotes the average position of the correct answer in the proposed candidate list. The higher the MRR is, the better the performance is.

Table 1. Performance of models wo/with error recovery

Models	Recall	MRR
ASR only	0.79	0.33
ASR + Web	0.94	0.94
ASR/Pre-Removed + Web	0.59	0.48

Table 2. Distribution before/after error recovery

	Before Error Recovery							After Error Recovery						
Length of NEs	Number of Matching Characters							Number of Matching Characters						
	0	1	2	3	4	5	6	0	1	2	3	4	5	6
2	11	23	0	-	-	-	-	13	21	0	-	-	-	-
3	6	29	76	0	-	-	-	6	39	64	2	-	-	-
4	6	25	90	184	0	-	-	19	52	66	62	106	-	-
5	9	10	12	77	193	0	-	11	23	36	41	53	137	-
6	0	0	1	8	20	39	0	0	3	19	12	7	5	22

Table 1 summarizes the experimental results of models without/with error recovery. In "ASR only" model, top 10 ASR strings produced at Stage 2 are regarded as answers. This model does not employ error recovery procedure. The recall rate is 0.79 and the MRR is 0.33. That is, total 79 of 100 transliteration names are recognized correctly and they appear in the first 3.03 (=1/0.33) positions. In contrast, "ASR + Web" model utilizes error recovery procedure. PAT candidates extracted from the Web are selected at Stage 4. The recall rate is 0.94 and the MRR is 0.94. Total 94 transliteration names are recognized correctly, and they appear in the first 1.06 (=1/0.94) positions on the average. In other words, when they are recognized correctly, they are always the top 1. Compared to the first model, the recall rate is increased 18.99%. In the third model, i.e., "ASR/Pre-Removed + Web" model, we try to evaluate the extreme power of error recovery. The correct transliteration names appearing in the set of ASR strings are removed before being submitted to search engine. That is, all the ASR strings submitted to search engine contain at least one wrong character. In such cases, the recall rate is 0.59 and the MRR is 0.48. That means 59 transliteration names are recovered, and they appear in the first 2.08 (=1/0.48) positions on the average. We further examine the number of errors in "ASR/Pre-Removed + Web" model to study the error tolerance of using the Web. Table 2 shows the analyses from the length of transliteration names (row part), and the number of matching characters (column part). For a transliteration name of length l, $0 \leq$ the number of matching characters $\leq l$. Each cell denotes how many strings belong to the specific category. For example, before error recovery, there are 6, 25, 90, 184, and 0 strings of length 4, which have 0, 1, 2, 3, and 4 matching characters with the corresponding answer keys, respectively. After error recovery, there are 19, 52, 66, 62, and 106 strings of length 4, which have 0, 1, 2, 3, and 4 matching characters with the answer keys, respectively. In other words, the recovery procedure corrects some wrong characters. The number of 1-character (2-character) errors is decreased from 184 (90) cases to 62 (66) cases, and total correct strings are increased from 0 to 106.

Table 3 shows the effects of the error positions (row part) and the string lengths (column part). Here only the cases of single errors are discussed. The cell denotes how many strings are recovered under the specific position and length. For example, total 37, 35, 20, and 17 single errors for strings of length 4 appearing at positions 1, 2, 3, and 4, respectively, can be recovered by the Web. From the length issue, the longer strings

Table 3. Effects of error positions and string lengths

Error Positions	Length=2	Length=3	Length=4	Length=5	Length=6	Total
Position 1	0	0	37	42	7	86
Position 2	0	2	35	42	4	83
Position 3	-	0	20	19	9	48
Position 4	-	-	17	24	3	44
Position 5	-	-	-	14	3	17
Position 6	-	-	-	-	1	1
Total	0	2	109	141	27	279

have better recovery capability than the shorter strings. In the experiments, 0% (=0/34), 1.80% (=2/111), 35.74% (=109/305), 46.84% (=141/301), and 39.71% (=27/68) of strings of lengths 2, 3, 4, 5, and 6 can be recovered, respectively. From the position issue, the errors appearing in the beginning are easier to be recovered than those appearing in the end. The experiments show that 30.82% (=86/279), 29.75% (=83/279), 17.20% (=48/279), 15.77% (=44/279), 6.09% (=17/279), and 0.36% (=1/279) of strings with wrong character appearing at positions 1, 2, 3, 4, 5 and 6 can be recovered, respectively.

4.2 Performance of speech recognition task

The set of 100 transliteration names in Section 4.1 are spoken by 2 males and 1 female, so that 300 transliteration names are recorded. We employ HTK and SRILM to get the best 100 syllable lattices (N Best, N=100). TCC-300 dataset for Mandarin is used to train the acoustic models. There are 417 HMM models and each has 39 feature vectors. The syllable accuracy is computed as: (M-I-D-S)/M * 100%, where M is the number of correct syllables, I, D, and S denote the number of insertion, deletion and substitution errors. The syllable accuracy is 76.57%. Easily, for estimating character recovery ability, we consider the exactly correct character number. The accuracy, ASR only and ASR+Web string character errors, are computed as:

$$\sum_{i=0}^{T} \max_{j=1 to K} \left(\frac{Sim(AnsKey_{ij}, ASR_i)}{Word_{Length}(TestName_i)} \right) \quad (2)$$

and

$$\sum_{i=0}^{T} \max_{j=1 to K} \left(\frac{Sim(AnsKey_{ij}, PAT_i)}{Word_{Length}(TestName_i)} \right) \quad (3)$$

,where T is total test number and K is answer keys number with a test name i. Table 4 shows character level results. The “ASR+Web” model has 3.82% performance increasing to the “ASR Only” model on the average. Table 5 shows word level results. The error recovery mechanism supported by the “ASR+Web” model improves the recall rate and the MRR of the “ASR Only” model from 0.20 and 0.054 to 0.37 and 0.290,

respectively. In other words, the average ranks of the correct transliteration names are moved from the 18th position (=1/0.054) to the 3rd position (=1/0.290) after error recovery.

5 Conclusions

This paper employs the web corpus to correct transliteration name recognition errors. Web fuzzy search proposes useful patterns for error recovery. Fault tolerance experiments show that longer transliteration names have stronger tolerance than shorter transliteration names, and the wrong characters appearing in the beginning of a transliteration name are relatively easier to be corrected than those appearing in the end. Thus, the improvement of character level accuracy will be helpful to the recovery mechanism, and vice versa. The ASR model integrated with the recovery mechanism by the Web search facilitates the spoken access to the Web directly.

Table 4. Performance on character level

ASR Only (Character Level Accuracy)					ASR + Web (Character Level Accuracy)				
Top 1	Top 2	Top 3	Top 4	Top 5	Top 1	Top 2	Top 3	Top 4	Top 5
38.74%	43.58%	46.97%	48.75%	50.04%	43.18%	47.78%	49.96%	52.30%	53.91%

Table 5. Performance on word level

ASR Only (Word Level)		ASR + Web (Word Level)	
Recall	MRR	Recall	MRR
0.20	0.054	0.37	0.290

Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC93-2752-E-001-001-PAE and NSC94-2752-E-001-001-PAE

6 References

[1] MUC *Message Understanding Competition*, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html, 1998.

[2] Lin, W.-C., Lin, M.-S. and Chen, H.-H., "Cross-Language Image Retrieval via Spoken Query", Proc. of RIAO 2004, 524-536, 2004.

[3] Appelt, D. E. and Martin, D., "Named Entity Extraction from Speech: Approach and Results Using the TextPro System", *Proc. of DARPA Broadcast News Workshop*, 1999, 51-54, 1999.

[4] Miller, D., Boisen, S., Schwartz, R. L., Stone, R., and Weischedel, R. M. "Named Entity Extraction from Noisy Input: Speech and OCR", Proc. of 6th Applied Natural Language Processing Conference, 316-324, 2000.

[5] Huang, F. and Waibel, A., "An Adaptive Approach of Name Entity Extraction for Meeting Application", Proc. of 2002 Human Language Technology Conference, 2002.

[6] Chen, H.-H., "Spoken Cross-Language Access to Image Collection via Captions", Proc. of 8th Eurospeech, 2749-2752, 2003.

[7] Keller, F. and Lapata, M., "Using the Web to Obtain Frequencies for Unseen Bigrams", Computational Linguistics, 29(3): 459-484, 2003.

[8] Gonnet, G. H., Baeza-Yates, R. A. and Snider, T., "New Indices for Text: PAT Trees and PAT Arrays", In Information Retrieval Data Structures Algorithms, Frakes and Baeza-Yates (eds.) Prentice Hall, 66-82, 1992.

[9] Chen, H.-H., Yang, C.-H., and Lin, Y. "Learning Formulation and Transformation Rules for Multilingual Named Entities", Proc. of ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, 1-8, 2003.

[10] Voorhees, E., "The TREC-8 Question Answering Track Evaluation", Proc. of the 8th TREC, 23-37, 1999.

[11] Lee-Feng Chien, "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval," ACM SIGIR Forum, 31, July 1997, 50-58.

Appendix A

Transliteration name	Answer keys list	Transliteration name	Answer keys list
科羅拉多	克羅拉多 柯羅拉多 科羅拉多	喬治克隆尼	喬治克隆尼 喬治克龍尼 喬治柯隆尼
加利福尼亞	加里福尼亞 加利福尼亞 加利弗尼亞	丹佐華盛頓	丹佐華聖頓 丹佐華盛頓
喬治亞	喬治亞 喬治雅	湯姆克魯斯	湯姆克魯斯
密西根	密希根 密西根	強尼戴普	強尼戴普
阿拉斯加	阿拉斯加	湯姆漢克斯	湯姆漢克斯
北卡羅萊納	北卡羅萊納 北卡羅萊那 北卡羅來納 北卡羅來那	*芮妮齊薇格	芮妮齊薇格 芮尼齊維格
康乃狄克	康乃迪克 康乃狄克 康迺迪克	莎莉賽隆	莎莉賽隆 莎莉塞隆 沙莉賽隆
德拉瓦	德拉瓦	妮可基嫚	妮可基嫚 尼可基曼 妮可基曼
佛羅里達	佛羅里達 佛羅理達	茱莉安摩爾	朱利安摩爾 茱莉安摩爾 朱莉安摩爾
南卡羅萊納	南卡羅萊納 南卡羅萊那 南卡羅來納 南卡羅來那	茱莉亞羅勃茲	茱莉亞羅勃茲 朱利亞羅伯茲 朱利亞羅勃茲 朱莉亞羅伯茲 茱莉亞羅伯茲
*夏威夷	夏威夷 夏威宜	威爾史密斯	威爾史密斯
愛荷華	艾荷華 愛何華 愛荷華	維果莫天森	維果莫天森 維果摩天森 維果墨天森
愛達荷	艾達荷 愛達荷	麥特戴蒙	麥特戴蒙
伊利諾	伊利諾 伊立諾 依利諾	休傑克曼	休傑克曼 休杰克曼
*印地安那	印地安那 印地安納 印弟安納	托貝馬奎爾	托貝馬奎爾
堪薩斯	坎薩斯 堪薩斯	鄔瑪舒嫚	烏瑪舒曼 鄔瑪舒曼
肯塔基	肯塔基	琪拉奈特莉	琪拉奈特莉 奇拉奈特莉
路易斯安那	路易斯安納	凱特貝琴薩	凱特貝琴薩
麻薩諸塞	麻薩諸塞 馬薩諸塞 麻塞諸塞	荷莉貝瑞	荷莉貝瑞 荷利貝瑞
緬因	緬因	安潔莉娜裘莉	安潔莉娜裘莉
馬里蘭	馬里蘭 馬利蘭	柴克巴爾夫	柴克巴爾夫
亞利桑那	亞利桑納 亞歷桑納 亞利桑那 亞歷桑那	布萊德彼特	布萊德比特 布萊德彼特 布萊得比特 布萊得彼特
明尼蘇達	明尼蘇達 明尼蘇答	金凱瑞	金凱瑞
密蘇里	米蘇里 密蘇里	柯林法洛	柯林法洛 科林法洛
密西西比	密西西比	裘德洛	裘德羅 裘德洛
蒙大拿	蒙大納 蒙大那 蒙大拿	娜塔莉波曼	娜塔利波曼 娜塔莉波曼

內布拉斯加	內布拉斯加	凱特溫絲蕾	凱特溫斯雷 凱特溫斯蕾 凱特溫絲雷 凱特溫絲蕾
阿拉巴馬	阿拉巴馬	*珍妮佛嘉納	珍妮佛嘉納
北達科他	北達科他 北達科塔	*茱兒芭莉摩	茱兒芭莉摩
新罕布夏	新漢布夏 新罕布夏	姚明	姚明
紐澤西	紐澤西	俠客歐尼爾	俠克歐尼爾 俠客歐尼爾
*新墨西哥	新墨西哥	凱文賈奈特	凱文加奈特 凱文賈奈特
內華達	內華達	*崔西麥格瑞迪	崔西麥格瑞迪 崔西麥葛瑞迪
紐約	紐約	柯比布萊恩	柯比布萊恩 科比布萊恩
俄亥俄	俄亥俄	文斯卡特	文斯卡特 文思卡特
奧克拉荷馬	奧克拉荷馬 奧克拉荷瑪 奧克拉河馬	提姆鄧肯	提姆鄧肯
奧勒岡	奧勒岡	葛蘭特希爾	格蘭特希爾 葛蘭特希爾
賓汐法尼亞	賓希法尼亞 賓西法尼亞 賓夕凡尼亞	勒布朗詹姆斯	勒布朗詹姆斯 勒布郎詹姆斯
*羅德島	羅德島	艾倫艾弗森	艾倫艾弗森 艾倫埃弗森 艾倫艾佛森
阿肯色	阿肯色 阿肯瑟 阿肯塞	*小歐尼爾	小歐尼爾
南達科他	南達科塔 南達科他 南達柯塔	拉希德華萊士	拉希德華萊士 拉西德華萊士
田納西	田納西 田那西	普林斯	普林斯
德克薩斯	德克薩斯 得克薩斯	賈米森	賈米森
*猶他	猶他	比盧普斯	比魯普斯 比盧普斯
佛蒙特	佛蒙特	斯托賈科維奇	斯托賈克維奇 斯托賈科維奇 斯托賈可維奇
維吉尼亞	維基尼亞 維吉尼亞 維吉尼雅	德克諾維茨基	德克諾維茨基 德克諾威茨基 德克諾維斯基
華盛頓	華聖頓 華盛頓 華勝頓	班華萊士	班華萊士 班華勒斯
西維吉尼亞	西維基尼亞 西維吉尼亞 西維吉尼雅	卡梅隆安東尼	卡梅隆安東尼 卡麥隆安東尼
威斯康辛	威斯康辛 威斯康新	斯塔德邁爾	斯塔德麥爾 斯塔德邁爾 斯塔達邁爾
懷俄明	懷俄明	基里連科	基里連科

***.** “印、島、兒、猶、小、芮” characters are not in training set and “尼、妮”, ” 辛、新” and “奇、琪” differentia of frequency is too high.

風險最小化準則在中文大詞彙連續語音辨識之研究

郭人瑋　劉士弘　陳柏琳
國立台灣師範大學資訊工程研究所
{rogerkuo, g93470185, berlin}@csie.ntnu.edu.tw

摘要

本論文探討風險最小化(Risk Minimization)準則在中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)之初步研究，內容包括了聲學模型訓練、非監督式聲學模型調適與搜尋演算法等方面。本論文以公視電視新聞語料庫作爲中文廣播新聞實驗題材。在聲學模模型訓練方面，我們使用了*最小化音素錯誤*(Minimum Phone Error, MPE)鑑別式訓練方法；實驗結果顯示，*最小化音素錯誤*訓練能較傳統最大相似度訓練相對地降低約 12%的字錯誤率。另一方面，在聲學模型調適上，我們則探討*最小化音素錯誤線性迴歸*(Minimum Phone Error Linear Regression, MPELR)調適法在非監督式聲學模型調適的使用；實驗結果顯示，*最小化音素錯誤線性迴歸*可以再進一步相對地降低約 5%的字錯誤率。最後，在搜尋演算法方面，本文探討詞錯誤最小化(Word Error Minimization, WEM)搜尋方法；實驗結果初步顯示，詞錯誤最小化搜尋方法較傳統最大化事後機率解碼方法來的稍佳。

1. 序論

隨著科技的快速演進，電腦早已融入每個家庭日常生活之中，而消費性的電子產品，更是改變了傳統的生活方式。然而，爲了攜帶上的便利，科技產品也愈做愈小，卻換來了輸入上的不便，人與機器的溝通，需要更簡便的方式才行。語音是人跟人之間最自然的溝通方式，自然地，我們也會希望人與機器之間最自然的溝通就是透過語音交談，因此自動語音辨識的研究也變得更加重要，特別是針對中文的輸入不便。另一方面，由於多媒體影音資訊迅速累積，例如廣播電視節目、語音信件、演講錄影和數位典藏等，這些多媒體資訊可以從網路上大量地取得，已成爲傳統文字資訊外社會大眾廣泛使用的資訊來源。是顯而易見的是，在上述的絕大部分多媒體資訊中，語音可以說是最具語意的主要內涵之一，當播出放多媒體的語音資訊或是顯示出對應的正確轉寫資訊，我們就可以大概地瞭解其中的主題或概念。因此，語音辨識技術對多媒體資訊的處理也扮演著相當重要的角色。

語音辨識可視爲一個分類的過程，在[1]中介紹了以最小化(分類)錯誤率(Minimum Error Rate, MER)來作爲分類的法則。在傳統的架構中，均以零壹函數(Zero-One Function)作爲減損函數(Loss Function)，藉此冀求分類過程能達到最小化分類錯誤。但在語音辨識中，每一文句代表一個類別，使用零壹函數作爲減損函數，可以最小化句錯誤率(Sentence Error Rate, SER)。但語音辨識在進行效能評估時，通常會以詞錯誤率(Word Error Rate, WER)，或在中文以字錯誤率(Character Error Rate, CER)作爲評估標準，使得最小化錯誤率法則與語音辨識實際上的評估方式產生了相當大的差異。換句話說，句錯誤愈小，不一定帶來較少的詞錯誤；而詞錯誤愈少，也不一定會有最少的

句錯誤。爲了克服此問題，近年來最常見的作法是以編輯距離(Levenshtein Edit Distance)[2]來取代零壹函數作爲減損函數，不論是在模型的訓練或是在搜尋演算法上，均有不錯的成果。本論文將在此架構下，探討風險最小化(Risk Minimization)準則在中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)之初步研究，內容包括了聲學模型訓練、非監督式聲學模型調適與搜尋演算法等方面。在聲學模模型訓練方面，我們使用了*最小化音素錯誤*(Minimum Phone Error, MPE)鑑別式訓練方法；在聲學模型調適上，我們則探討*最小化音素錯誤線性迴歸*(Minimum Phone Error Linear Regression, MPELR)調適法在非監督式聲學模型調適的使用；最後，在搜尋演算法方面，我們探討詞錯誤最小化(Word Error Minimization, WEM)搜尋方法。在以公視電視新聞語料庫作爲中文廣播新聞實驗題材下，初步驗證了上述這些方法在中文大詞彙連續語音辨識上均有不錯的成效。

本論文接下來的安排如下：第二章將介紹貝氏風險與全面風險；第三章則介紹*最小化音素錯誤*聲學模型訓練；第四章探討*最小化音素錯誤*爲基礎的線性轉換調適技術；第五章則探討詞錯誤最小化搜尋演算法；第六章爲實驗與討論；第七章爲結論與未來展望。

2. 貝氏風險與全面風險

若 O_r 爲一語句的聲學特徵向量序列，將 O_r 歸類至文句 s 時，可以用函數 $R(s \mid O_r)$ 代表此歸類行爲的風險(Risk)；而語音辨識則可視爲找出此風險最低的文句。將 O_r 歸類至 s 的風險 $R(s \mid O_r)$ 可定義如下[1]：

$$R(s \mid O_r) = \sum_{u \in \mathbf{W}_h} P(u \mid O_r) L(s,u), \tag{1}$$

其中 $\mathbf{W}_h$ 爲聲學特徵向量序列 O_r 所有可能對應的文句所成之集合； $P(u \mid O_r)$ 表示給定 O_r 時，文句 u 的事後機率(Posterior Probability)； $L(s,u)$ 爲一減損函數(Loss Function)，用以表示文句 s 與 u 之間差異所造成的損失(Loss)， $R(s \mid O_r)$ 爲將 O_r 歸類至 s 時的期望損失(Expected Loss)，又稱爲條件風險(Conditional Risk)。在語音辨識解碼上，可以最小化此條件風險來尋找最佳的文句 s^* ：

$$s^* = \arg\min_s R(s \mid O_r) = \arg\min_s \sum_{u \in \mathbf{W}_h} P(u \mid O_r) L(s,u), \tag{2}$$

而因此產生的風險即爲貝氏風險(Bayes Risk) R_{Bayes} [1]：

$$R_{Bayes} = \min_s R(s \mid O_r) = \min_s \sum_{u \in \mathbf{W}_h} P(u \mid O_r) L(s,u). \tag{3}$$

目前有許多語音辨識器根據貝氏決策定理(Bayesian Decision Rule)，即最小化此條件風險前提下來設計其搜尋演算法，如傳統的最大化事後機率(Maximum a Posteriori, MAP)解碼方法[3]、ROVER(Recognizer Output Voting Error Reduction)[4]、最小化貝氏風險(Minimum Bayes Risk, MBR)[5]及詞錯誤最小化(Word Error Minimization, WEM)[6] 等。然而，就模型訓練而言，則需

要最小化全面風險(Overall Risk) $R_{Overall}$ [1]：

$$R_{Overall} = \int R(s_r \mid O_r)P(O_r)dO_r, \tag{4}$$

其中 s_r 爲 O_r 對應之正確轉譯文句(Correct Transcription)， $P(O_r)$ 爲 O_r 的事前機率(Prior Probability)；全面風險 $R_{Overall}$ 是在語句空間上作積分，爲所有訓練語句的期望條件風險(Expected Conditional Risk)。由於訓練語料有限，故全面風險可簡化：

$$R_{Overall} = \sum_r R(s_r \mid O_r)P(O_r) = \sum_r \sum_{u \in \mathbf{W}_h} P(u \mid O_r)L(s_r, u)P(O_r). \tag{5}$$

若事後機率分佈 $P(u \mid O_r)$ 由聲學模型(Acoustic Model) λ 及語言模型(Language Model) Γ 所決定，記作 $P_{\lambda,\Gamma}(u \mid O_r)$ ，則全面風險可改寫成：

$$R_{Overall} = \sum_r \sum_{u \in \mathbf{W}_h} P_{\lambda,\Gamma}(u \mid O_r)L(s_r, u)P(O_r). \tag{6}$$

若再假設 $P(O_r)$ 對所有聲學特徵向量序列 O_r 均有一致(Uniform)的機率，且此項與模型參數 λ 及 Γ 的訓練無關，則可將此項省略：

$$R_{Overall} \approx \sum_r R(s_r \mid O_r) = \sum_r \sum_{u \in \mathbf{W}_h} P_{\lambda,\Gamma}(u \mid O_r)L(s_r, u). \tag{7}$$

在估測模型時，希望估測之模型 (λ, Γ) 能將全面風險降至最低：

$$(\lambda, \Gamma) = \arg\min_{\lambda', \Gamma'} \sum_r \sum_{u \in \mathbf{W}_h} P_{\lambda', \Gamma'}(u \mid O_r)L(s_r, u). \tag{8}$$

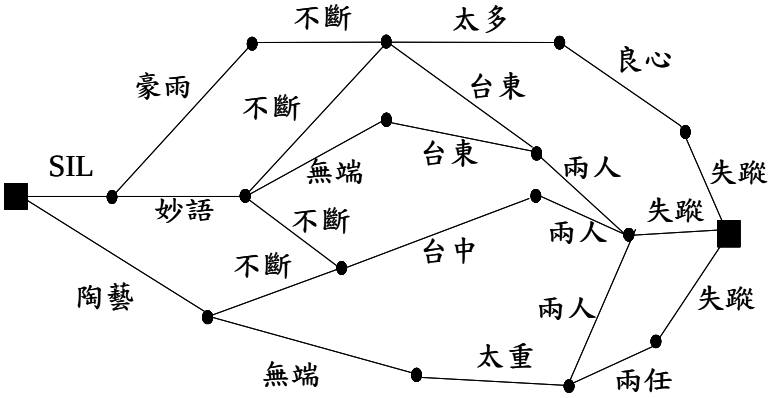
3. 最小化音素錯誤訓練

在估測模型時，我們希望估測之模型 (λ, Γ) 能將全面風險降至最低，式(8)因此可進一步表示成：

$$(\lambda, \Gamma) = \arg\min_{\lambda', \Gamma'} \sum_r \sum_{u \in \mathbf{W}_h} \frac{p_{\lambda'}(O_r \mid u)P_{\Gamma'}(u)}{\sum_{v \in \mathbf{W}_h} p_{\lambda'}(O_r \mid v)P_{\Gamma'}(v)} L(u, s_r), \tag{9}$$

其中 $p_{\lambda'}(O_r \mid u)$ 爲文句 u 對應的聲學模型產生聲學特徵向量序列 O_r 的機率分佈，如連續密度隱藏式馬可夫模型(Continuous Density HMM, CDHMM)； $P_{\Gamma'}(u)$ 爲文句 u 對應的語言模型機率分佈，如詞 *n*-連(詞雙連、詞三連)語言模型。

*全面風險法則估測*首先由 Na 等人所提出[7]，初步地使用零壹減損函數，來最小化訓練語料中的貝氏風險(Bayes Risk)，在獨立數字辨識(Isolated Digit Recognition)上可降低不少的錯誤率。Kaiser 等人在 2000 年時則以 Levenshtein 距離來取代零壹函數，以 *N*-最佳序列(*N*-Best List)作爲所有可能文句之近似，並使用延伸波氏重估(Extended Baum-Welch Re-estimation, EBW)演算法來爲模型參數進行最佳化[8][9]。



圖一、詞圖爲所有可能文句 $\mathbf{W}_h$ 的近似。

*最小化音素錯誤*訓練法則[10][11]類似全面風險法則的概念，同樣也是以最大化所有文句的期望辨識率爲目標。但*最小化音素錯誤*與*全面風險法則估測*有下列的差異：

1. 使用詞圖(Word Graph)來取代 *N*-最佳序列(*N*-best List)作爲所有可能文句之近似，如圖一所示。
2. 引入模型參數事前機率，來增加估測值的強健性。
3. 對於延伸波氏重估演算法中的控制參數，提出更佳的設定方式。
4. 強調音素層次的正確率而非詞正確率。

*最小化音素錯誤*訓練法則聲學模型訓練的目標函數 $F_{MPE}(\lambda)$ 爲：

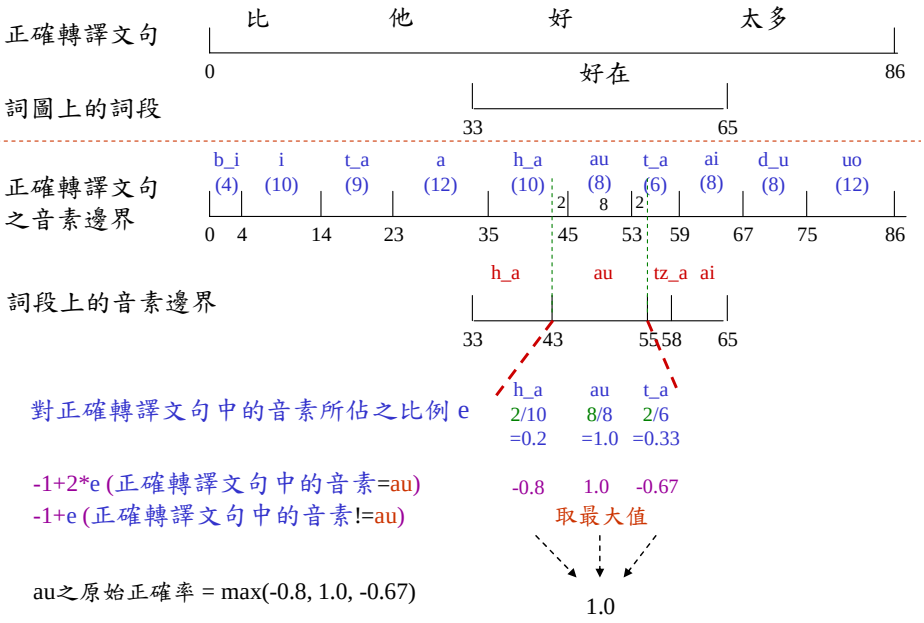
$$F_{MPE}(\lambda)=\sum_r \sum_{u\in \mathbf{W}_h} \frac{p_\lambda(O_r\mid u)P(s)}{\sum_{v\in \mathbf{W}_h} p_\lambda(O_r\mid v)P(v)} A(u,s_r). \tag{10}$$

在實作上，由於不可能對聲學特徵向量序列 O_r 所有可能對應的文句 $\mathbf{W}_h$ 作窮舉，與*全面風險法則估測*以 *N*-最佳路徑作爲所有可能文句之近似不同的是，*最小化音素錯誤*是以 $\mathbf{W}_{lat}^r$ 來近似，它是以第 r 句訓練語句辨識過後所產生的詞圖並加入正確轉譯文句 s_r 的詞分枝所形成之可能文句集合。另一方面， $A(v,s_r)$ 爲文句 v 相對於正確轉譯文句 s_r 的正確率，由於傳統以全域比對(Global Matching)結合編輯距離(Levenshtein Edit Distance)計算正確率並沒有考慮到時間上的相關性，無法提供充分正確的資訊供聲學模型訓練使用。因此 Povey 等人於 2002 年時對*最小音素錯誤訓練*提出一套音素之間計算正確率的方式[10]。如圖二所示，正確轉譯文句 s_r 爲「比-他-好-太多」，而以詞圖中某一詞段「好在」爲例，要計算詞圖上某一音素的正確率有三個步驟(在此以辨識文句 *v* 中的音素 au 爲例)：

Step 1. 在正確轉譯文句中找出與 au 有時間重疊之音素 h_a、au 與 t_a，分別的重疊長度爲 2、8、2 個音框(Frame)。

Step 2. 計算辨識文句中 au 對此三音素所重疊比例，如對 h_a 重疊 2 個音框，而 h_a 在正確轉譯文句中實際長度爲 10 個音框，所以所重疊的比例爲 0.2。同理可求得對轉譯文句中 au 的重疊比例爲 1.0、對 t_a 的重疊比例爲 0.33。

Step 3. 再來先計算辨識文句中 au 對此三音素的正確率，若音素相同，則計算方式爲-1+2*重疊比例，否則爲-1+重疊比例。對 h_a 來說，因爲 h_a ≠ au，所以對 h_a 之正確率爲-0.8，同理可得對 au 的正確率爲 1.0、對 t_a 的正確率爲-0.67。而最後 au 之正確率取上述三個值中的最大值 1.0。



圖二、音素原始正確率計算方式[10]。

爲了對目標函數 $F_{MPE}(\lambda)$ 進行最佳化，Povey 等人提出*最小化音素錯誤*的弱性(Weak-sense)輔助函數 $H_{MPE}(\lambda,\bar{\lambda})$ 爲[11]：

$$g_{MPE}(\lambda,\bar{\lambda})=\sum_r \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{r\,MPE}\gamma_{qm}^r(t)\log N(o_r(t);\mu_m,\Sigma_m) \tag{11}$$

其中 s_q 與 e_q 分別代表音素 q 的起始時間(Start Time)與結束時間(End Time)；$o_r(t)$ 是 O_r 的第 t 個語音特徵向量；$N(;\mu_m,\Sigma_m)$ 是音素 q 的第 m 個高斯分佈，μ_m 與 Σ_m 分別是它的平均值向量與共變異矩陣；$\gamma_{qm}^r(t)$ 則是第 r 句訓練語句中在時間 t 時音素上 q 的高斯分佈 m 的佔有機率；$\gamma_q^{r\,MPE}$ 可進一步表示成[11]：

$$\gamma_q^{r\,MPE}=\left.\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_r\mid q)}\right|_{\lambda=\bar{\lambda}}=\gamma_q^r\left(c_r(q)-c_{avg}^r\right) \tag{12}$$

其中 γ_q^r 是音素 q 在已知 O_r 情況下的事後機率；$c_r(q)$ 爲在詞圖 $\mathbf{W}_{lat}^r$ 中所有經過音素分枝 q 的文句對於 s_r 之期望正確率; c_{avg}^r 爲在詞圖 $\mathbf{W}_{lat}^r$ 中所有文句相對於 s_r 之期望正確率。γ_q^r 、$c_r(q)$ 與 c_{avg}^r 的統計量可在詞圖上使用波氏重估演算法來求得[11]。爲了增加模型估測的強健性，防止*最小化音素錯誤*過度的訓練並增進聲學模型的一般性(Generalization)，克服訓練與測試環境的不匹配，故在此引入以舊有模型參數爲超參數的平滑函數 $g_{EBW}^{smooth}(\lambda)$ 來加以輔助，由於弱性輔助函數加上平滑函數仍會滿足弱性輔助函數的性質，可提供較佳的參數估測。平滑函數 $g_{EBW}^{smooth}(\lambda)$ 則定義爲：

$$g_{EBW}^{smooth}(\lambda,\bar{\lambda})=\sum_m -\frac{D_m}{2}\left[\log(|\Sigma_m|)+(\mu_m-\bar{\mu}_m)^T\Sigma_m^{-1}(\mu_m-\bar{\mu}_m)+tr(\bar{\Sigma}_m\Sigma_m^{-1})\right] \tag{13}$$

其中 D_m 爲高斯分佈層次的平滑係數。此平滑函數以舊有之模型參數，如平均值向量 $\bar{\mu}_m$ 與共變

異矩陣 $\overline{\Sigma}_m$ 作爲超參數(Hyper-parameters)，使新估測之模型參數不致改變太大。加入此平滑函數後，輔助函數即成爲[11]：

$$g_{MPE}(\lambda,\overline{\lambda}) = \sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{r\,MPE}\gamma_{qm}^r(t)\log N(o_r(t);\mu_m,\Sigma_m) - \sum_m \frac{D_m}{2}\left[\log(|\Sigma_m|) + (\mu_m-\overline{\mu}_m)^T\Sigma_m^{-1}(\mu_m-\overline{\mu}_m) + tr(\overline{\Sigma}_m\Sigma_m^{-1})\right] \tag{14}$$

分別對平均值向量與共變異矩陣作偏微分並使式爲 $\vec{0}$ 向量及 **0** 矩陣，則可推導出用於平均值向量與共變異矩陣的延伸波氏重估公式[12]：

$$\mu_m = \frac{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t)o_r(t) + D_m\overline{\mu}_m}{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t) + D_m}, \tag{15}$$

$$\Sigma_m = \frac{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t)o_r(t)o_r(t)^T + D_m\left[\overline{\Sigma}_m + \overline{\mu}_m\overline{\mu}_m^T\right]}{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t) + D_m} - \mu_m\mu_m^T, \tag{16}$$

其中 D_m 必須確保估測值共變異矩陣(Σ_m)爲正定矩陣。I-平滑(I-smoothing)技術[10]同樣是爲輔助函數引入平滑函數 $g^{I-smooth}(\lambda)$，但此平滑函數是以*最大化相似度*(Maximum Likelihood, ML)估測之統計資訊作爲超參數(Hyper-parameters)，故此函數 $g^{I-smooth}(\lambda)$ 可定義爲[11]：

$$g^{I-smooth}(\lambda) = \sum_m -\frac{\tau_m}{2}\left[\log(|\Sigma_m|) + (\mu_m - \frac{\theta_m^{ML}(O)}{\gamma_m^{ML}})^T\Sigma_m^{-1}(\mu_m - \frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}) + tr\left(\left(\frac{\theta_m^{ML}(O^2)}{\gamma_m^{ML}} - \left(\frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}\right)\left(\frac{\theta_m^{ML}(O)}{\gamma_m^{ML}}\right)^T\right)\Sigma_m^{-1}\right)\right], \tag{17}$$

其中 τ_m 是一個高斯分佈層次的平滑係數，表示要由*最大化相似度*統計資訊加入的資料點數；$\gamma_m^{r\,ML}(t)$ 表示在第 r 句訓練語句中，時間 *t* 時，以*最大化相似度*法則所估測之高斯分佈 *m* 的佔有機率，可表示成 $\gamma_m^{ML} = \sum_r \sum_t \gamma_m^{r\,ML}(t)$；而 $\theta_m^{ML}(O)$ 與 $\theta_m^{ML}(O^2)$ 分別可表示成 $\theta_m^{ML}(O) = \sum_r \sum_t \gamma_m^{r\,ML}(t)o_r(t)$ 與 $\theta_m^{ML}(O^2) = \sum_r \sum_t \gamma_m^{r\,ML}(t)o_r(t)o_r(t)^T$。因此重估公式要修改爲[10]：

$$\mu_m = \frac{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t)o_r(t) + \frac{\tau_m}{\gamma_m^{ML}}\theta_m^{ML}(O) + D_m\overline{\mu}_m}{\sum_r \sum_{q\in\mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE}\gamma_{qm}^r(t) + \tau_m + D_m}, \tag{18}$$

$$\Sigma_m = \frac{\sum_r \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE} \gamma_{qm}^r(t) o_r(t) o_r(t)^T + \frac{\tau_m}{\gamma_m^{ML}} \theta_m^{ML}(O^2) + D_m \left[\overline{\Sigma}_m + \overline{\mu}_m \overline{\mu}_m^T\right]}{\sum_r \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{r\,MPE} \gamma_{qm}^r(t) + \tau_m + D_m} - \mu_m \mu_m^T, \tag{19}$$

其中 $\gamma_m^{r\,ML}(t)$ 表示在第 *r* 句訓練語句中，時間 *t* 時，以*最大化相似度*法則所估測之音素 *q* 的第 *m* 個高斯分佈的佔有機率。

4. 最小化音素錯誤爲基礎的線性轉換調適技術

在模型空間上的調適方法中，連續密度隱藏式馬可夫模型中的平均值向量與共變異矩陣分別使用不同的線性迴歸矩陣，本文的研究只針對平均值向量的調適。在平均值向量的調適中，假設調適前的高斯分佈 *m* 的平均值向量爲 $\overline{\mu}_m$ ，調適後的平均值向量爲 μ_m ，並希望透過一線性迴歸矩陣來調適此平均值向量：

$$\mu_m = A\overline{\mu}_m + b = W\overline{\xi}_m, \tag{20}$$

其中 $W = \left[b\ A\right]$ 爲調整平均值向量的線性迴歸矩陣，其中 A 爲旋轉矩陣，而 b 爲偏移向量；$\overline{\xi}_m = \left[1\ \overline{\mu}_m^T\right]^T$ 是調適前的延伸平均值向量(Extended Mean Vector)。由式(11)可得到輔助函數 $g_{MPE}(W, \overline{W})$ 爲[13]：

$$g_{MPE}(W, \overline{W}) = \sum_m \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) \log N(o(t); W\xi_m, \Sigma_m), \tag{21}$$

其中 $\overline{W}$ 爲舊有的轉換矩陣，γ_q^{MPE} 同樣可由式(12)求得，而估測 $\gamma_{qm}(t)$ 與 γ_q^{MPE} 所需的聲學模型，由舊有的轉換矩陣 $\overline{W}$ 所轉換。爲了增加線性迴歸矩陣估測的強健性，同式(14)，在此也加入了以舊有線性迴歸矩陣 $\overline{W}$ 爲超參數的平滑函數 $g_{EBW}^{smooth}(W, \overline{W})$ 來輔助最佳化，$g_{EBW}^{smooth}(W, \overline{W})$ 則定義爲[13]：

$$g_{EBW}^{smooth}(W, \overline{W}) = \sum_m -\frac{D_m}{2} (W\xi_m - \overline{W}\xi_m)^T \Sigma_m^{-1} (\overline{W}\xi_m - W\xi_m), \tag{22}$$

其中 D_m 爲高斯分佈層次的平滑係數，所以新的輔助函數爲[13]：

$$\begin{aligned} g_{MPE}(W, \overline{W}) = &\sum_m \sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) \log N(o(t); W\xi_m, \Sigma_m) \\ &- \frac{1}{2} \sum_m D_m \left[(\overline{W}\xi_m - W\xi_m)^T \Sigma_m^{-1} (\overline{W}\xi_m - W\xi_m)\right], \end{aligned} \tag{23}$$

我們若將式(23)對 W 作偏微分，並設之等於零，可得：

$$\mathbf{G}^{-1}\mathbf{w} = \mathbf{z}, \tag{24}$$

其中 $\mathbf{z} = vec(Z)$ ， $\mathbf{w} = vec(W)$ ，*vec*(・)是一個將矩陣轉換爲向量的函數，以列優先的排序方式；而 Z 表示成：

$$Z = \sum_m \Sigma_m^{-1} \left(\sum_{q\in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) o(t) + D_m \overline{W}\xi_m \right) \xi_m^T, \tag{25}$$

另一方面，$\mathbf{G}$ 可進一步表示成：

$$\mathbf{G} = \sum_{m} kron(V_m, R_m), \tag{26}$$

而其中 *kron*(・)則爲 kronecker 矩陣乘法[14]，V_m 與 R_m 分別爲：

$$V_m = \left(\sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) + D_m \right) \Sigma_m^{-1}, \tag{27}$$

$$R_m = \xi_m \xi_m^T. \tag{28}$$

W 可由線性方程組(Systems of Linear Equations)式(24)求出。爲了避免過度訓練，I-平滑技術同樣也可以使用在此。以*最大化相似度*之統計資訊爲超參數(Hyper-parameters)的平滑函數 $g^{I-smooth}(W)$ 可定義爲[13]：

$$g^{I-smooth}(W) = \sum_{m} -\frac{\tau_m}{2} \left[(o(t) - W\xi_m)^T \Sigma_m^{-1} (o(t) - W\xi_m) \right], \tag{29}$$

其中 τ_m 是一個係數，要表示要由*最大化相似度*統計資訊加入的資料點數，因此可將式(25)與式(27)修改爲[13]：

$$Z = \sum_{m} \Sigma_m^{-1} \left(\left(\sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) + \tau_m \right) o(t) + D_m \overline{W} \xi_m \right) \xi_m^T, \tag{30}$$

$$V_m = \left(\sum_{q \in \mathbf{W}_{lat}^r} \sum_{t=s_q}^{t=e_q} \gamma_q^{MPE} \gamma_{qm}(t) + \tau_m + D_m \right) \Sigma_m^{-1}, \tag{31}$$

5. 詞錯誤最小化搜尋演算法

本論文以詞錯誤最小化(Word Error Minimization, WEM)[6]搜尋演算法(或稱爲最小化貝氏風險搜尋(Minimum Bayes Risk, MBR)[5]作爲搜尋法則，在式(2)中，由於實際上不可能對所有可能的文句 $\mathbf{W}_h$ 作窮舉，所以我們根據[6]先對測試聲學特徵向量序列 O_l 先進行語音辨識產生詞圖 $\mathbf{W}_{lat}^l$，並從詞圖 $\mathbf{W}_{lat}^l$ 上找出 *N*-最佳序列(*N*-best List) $\mathbf{N}^l$，再接著估測 $\mathbf{N}^l$ 中每一路徑(或文句)的條件風險，進而選出擁有最小條件風險的文句。其詞錯誤最小化搜尋的標準法則爲：

$$s^* = \arg\min_{s} \sum_{u \in \mathbf{N}^l} P(u \mid O_l) L(s,u), \tag{32}$$

在上式中，每一條路徑 s 都要與其它所有的路徑 u 算編輯距離 $L(s,u)$ 並乘上 u 的事後機率，將之加總起來即爲一條路徑 s 的條件風險。我們可以基於此一條件風險從 $\mathbf{N}^l$ 找出最小條件風險的文句 s^* 作爲最後的語音辨識結果。由於在中文的斷詞上會有混淆的問題，詞錯誤率通常不是一個評估語音辨識效能的很好標準；因此，在本研究我們在計算編輯距離 $L(s,u)$ 時，實作上是以字作爲比對與計算的單位。

N-最佳路徑	正確答案：”昔日-執政黨-大-掌櫃”		
u (sentence)	$P(u \mid O_l)$	$\sum_{u \in \mathbf{N}^l} P(u \mid O_l) L(s,u)$	CER
“昔日-行政-長-大-掌櫃”	0.23	85	2
“七-日-執政黨-把-掌櫃”	0.21	90	2
“昔日-執政黨-把-掌櫃”	0.19	73	1
“昔日-指-政黨-大-掌櫃”	0.18	87	1
“昔日-行政黨-把-掌櫃”	0.17	101	2

表一、最小條件風險之例子，$\mathbf{N}^l$ 大小設爲 50。

性別	訓練語料		評估語料			語料重疊人數
	總長(分鐘)	人數(人)	總長(分鐘)	人數(人)	平均長度(字)	(人)
男生	766.69	≤66	21.69	9	89.8	9
女生	766.79	≤111	65.23	≤23		≥13

表二、語音實驗語料統計資訊。

如表一所示，正確答案爲“昔日-執政黨-大-掌櫃”，用傳統最大事後機率所找出的文句爲“昔日-行政-長-大-掌櫃”其 CER(字錯誤率)=2，若用詞錯誤最小搜尋法則，找到的最小條件風險文句爲“昔日-執政黨-把-掌櫃”其 CER=1。在此一例子可明顯地看出，利用詞(字)錯誤最小化搜尋法則可以降低字錯誤率，進而提高語音辨識效能。

6. 實驗與討論

6.1 實驗架構與設定

本論文所使用的大詞彙連續語音辨識器爲台灣師範大學資工所目前所發展的新聞語音辨識系統[15]，它基本上是一套大詞彙連續語音辨識系統，主要包括前端處理、詞彙樹複製搜尋(Tree-Copy Search)及詞圖搜尋(Word Graph Rescoring)[16]等部分。

在前端處理方面，本文使用梅爾倒頻譜特徵作爲語音訊號的特徵參數，並且使用倒頻譜平均消去法(Ceptral Mean Subtraction, CMS)以移除錄音時通道效應所造成的影響。在聲學模型我們總共使用了 151 個隱藏式馬可夫模型來作爲中文 INITIAL-FINAL 的統計模型，其中隱藏式馬可夫模型的每個狀態會依據其對應到的訓練語料多寡，以 2 到 128 個高斯分佈來表示，本論文總共使用到約 14,396 個高斯分佈。

另一方面，本論文所使用的詞典約含有七萬二千個一至十字詞，並以從中央通訊社(Central News Agency, CNA) 2001 與 2002 年所收集到的約一億七千萬(170M)個中文字語料作爲背景語言模型訓練時的訓練資料[17]。在本文中的語言模型使用了 Katz 語言模型平滑技術[18]，在訓練時是採用 SRL Language Modeling Toolkit (SRILM)[19]。在詞彙樹搜尋時，本系統採用詞雙連語言

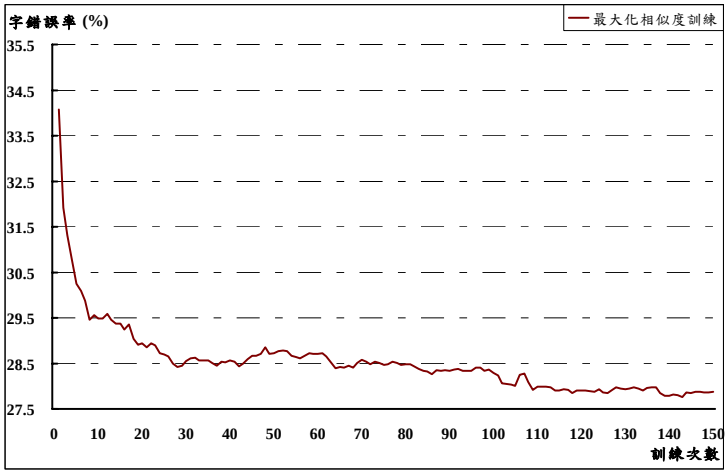
模型；在詞圖搜尋時，則採用詞三連語言模型。

6.2 實驗語料

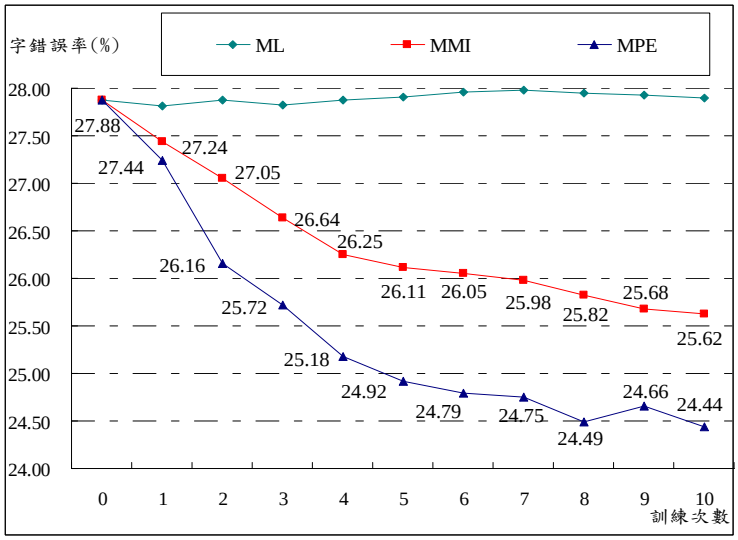
本論文實驗使用的訓練與測試語料爲 MATBN 電視新聞語料庫[20]，是由中央研究院資訊所口語小組[21]耗時三年與公共電視台[22]合作錄製完成。我們初步地選擇採訪記者語料作爲實驗語材，其中包含 25.5 小時的訓練集(5,774 句)，供聲學模型訓練之用，其中男女語料各半；1.5 小時的評估集(292 句)，供辨識評估之用。訓練集由 2001 及 2002 年的新聞語料所篩選出來的；評估集則均爲 2003 年的語料，由中研院的評估語料篩選出來，只選擇了採訪記者語料並濾掉了含有語助詞之語句。語料中語者資訊如表二所示。另外，評估集中每則新聞的平均長度爲 89.8 個字。

6.3 聲學模型訓練之實驗

本論文先進行 150 次迭代的*最大化相似度*訓練，字錯誤率曲線如圖三所示。然後分別再進行 10



圖三、150 次的最大化相似度聲學模型訓練之字錯誤率曲線。



圖四、10 次最小化音素錯誤訓練、最大化交互資訊訓練及最大化相似度訓練字錯誤率曲線。

	CER (%)
Baseline: MPE	24.44
MPE + MLLR	23.41
MPE + MMILR (τ_m=3)	23.52
MPE + MMILR (τ_m=10)	23.30
MPE + MMILR (τ_m=30)	23.20
MPE + MPELR (τ_m=3)	23.07
MPE + MPELR (τ_m=10)	23.11
MPE + MPELR (τ_m=30)	23.19

表三、鑑別式聲學模型調適之字錯誤率(%)。

	CER (%)
Baseline: MPE + MPELR	23.07
MPE + MPELR + WEM (N-best)	23.04
Word Graph Error Rate (GER)	12.28

表四、詞錯誤最小化搜尋之結果。

次迭代的鑑別式訓練(Large Scale Discriminative Training)，其中包含*最小化音素錯誤*訓練及*最大化交互資訊*訓練(Maximum Mutual Information, MMI)[23]。字錯誤率曲線如圖四所示，其中我們以經由額外 10 次迭代的*最大化相似度*訓練模型作爲對照組(所以總共經歷 160 次迭代的*最大化相似度*訓練)。在實驗的設定中，正確的轉譯文句要先經由強制對齊(Force Alignment)產生詞段，再加入詞圖中。I-平滑技術中的控制參數 $\tau_m = 10$ ，詞段的聲學分數使用維特比的分數再經過 1/12 次方來縮小和語言機率的比例，利用詞圖進行鑑別式聲學模型訓練時所用的語言限制則是使用詞單連(Unigram)語言模型。由實驗結果顯示，經過 150 次*最大化相似度*訓練之後，錯誤率的下降已趨於飽和，無法再藉由傳統*最大化相似度*訓練來降低錯誤率。需藉由鑑別式的方法做更進一步的訓練，如圖四，*最大化交互資訊*能相對降低 8.11%的字錯誤率，已有相當顯著的成效。而經過 10 次的*最小化音素錯誤*訓練之後，可相對降低 12.34%的字錯誤率，明顯優於*最大化交互資訊*訓練所帶來的成效，有兩個原因。第一、*最小化音素錯誤*使用 Levenshtein 函數的改良正確率計算方式作爲減損函數，較*最大化交互資訊*所使用的零壹函數更貼近於評估標準。第二、*最小化音素錯誤*直接最小化全面風險，*最大化交互資訊*卻是透過一個上界間接最小化全面風險。

6.4 非監督式聲學模型調適之實驗

在非監督式聲學模型調適的實驗中，以 10 次*最小化音素錯誤*訓練之模型作爲基礎(Baseline)，並將 14,396 個高斯分佈以聲母、韻母及靜音區分成 3 個迴歸群集(Regression Classes)對平均值向量做調適。實驗結果如表三所示，其中MLLR代表*最大化相似度線性迴歸*，MMILR代表*最大交互資訊線性迴歸*，MPELR代表*最小化音素錯誤線性迴歸*。在實驗設定方面，我們針對不同鑑別式聲學模型調適方法使用不同的τ_m最來作初步的探討，其他設定則與聲學模型訓練時維持一致。由

實驗結果可見，隨著τ_m 愈大，MMILR的表現愈來愈好，在τ_m=30 時有較佳的字錯誤率；而MPELR則剛好相反，在τ_m=3 時有較佳的字錯誤率。我們推測原因可能是MMILR的輔助函數與目標函數差異較大，需要引入較多的*最大化相似度*估測值，來得到較強健性的估測；而MPELR的輔助函數則較貼近於目標函數，使得愈多的*最大化相似度*估測值，反而會減緩估測的收斂速度。我們進一步分析可發現MPELR在τ_m=3 時能相對降低 5.61%的字錯誤率，相較於MMILR(τ_m=30)的 5.07%、MLLR的 4.21%，MPELR的確提供了較佳的聲學模型調適效能。

6.5 搜尋解碼之實驗

如表四的第三列所示，使用詞錯誤最小化(WEM)搜尋演算法後僅能將字錯誤率由 23.07%(表三最佳的字錯誤率)降低至 23.04%，效果並不如預期的好。經由觀察，我們發現在每一個詞圖所產生的前 1,000 名最佳序列中，即式(32)中的 $\mathbf{N}'$ 大小爲 1,000 時，大部分的序列都與第一名序列(即事後機率最大的序列)很相近，這樣我們便沒有辦法利用編輯距離的分數來降低某一條有可能是字錯誤最小序列的風險，進而將字錯誤最小序列取代第一名序列。另外，由於大部分序列都與第一名序列相近，若第一名序列的事後機率與其它序列的事後機率差不多的情況下時，利用詞錯誤最小化搜尋演算法有可能會將其他條序列取代第一名序列，造成辨識率下降；僅少部分的序列會與第一名序列較不相近而與其他前幾名的序列較相近，這樣情況下利用詞錯誤最小化搜尋演算法便可以找出字錯誤最小序列取代第一名序列(如表一的例子)，進而提高辨識率。

7. 結論與未來展望

由於傳統聲學模型訓練方式是以*最大化相似度法*則來訓練，使聲學模型在訓練語料上有最大的聲學相似度，但由於訓練環境與測試環境的差異(Mismatch)，過多的訓練次數，不僅無法帶來額外的辨識效能，反而會因此降低了辨識率，而*最小化音素錯誤*訓練法則，能在*最大化相似度*訓練之後，再對聲學模型施以鑑別式訓練及調適，使聲學模型更具鑑別力。本論文中，經由初步的實驗顯示，在聲學模型訓練上，*最小化音素錯誤*能有效的降低辨識錯誤率，相對於*最大化相似度*訓練能降低 15.52%的音節錯誤率、12.33%的字錯誤率及 10.02%的詞錯誤率。在聲學模型的調適上，也有小幅度的進步。在搜尋方面，由於詞圖的資訊提供了字正確率的上限爲 87.72%(如表四的第四列所示，詞圖的最佳字錯誤率爲 12.28%)，而在 *N*-最佳序列上的*詞錯誤最小化*搜尋僅能將字正確率提升至 76.95%(23.05%字錯誤率)，所以尚有 10.67%字錯誤率的空間可供改善。在寫作此論文的同時，我們正計畫使用 A*搜尋(A* Search)，以*詞錯誤最小化*準則直接在詞圖上找最佳的文句[5]，以冀求更佳的辨識率。

另外，我們也曾嘗試將*最小化音素錯誤*訓練法則應用在語言模型估測上，靠著估測語言模型之機率，來最大化訓練語料中詞圖的期望正確率，雖然在語音辨識率的提昇上不甚顯著，但是對語言模型估測卻提供了新的視野；而新近也有學者將*最小化音素錯誤*訓練法則應用在特徵空間轉換上[25]，用來求取特徵空間的轉換矩陣，研究結果顯示透過*最小化音素錯誤*訓練法則來求取轉換矩陣，除了能達到降低語音特徵向量維度與減少維度間相關性的功用外，對於語音辨識率的提昇也得到相當大的成效。

8. 參考文獻

[1] R. O. Duda, P. E. Hart and D. G. Stork. Pattern Classification, Second Edition. *New York: John & Wiley*, 2001.

[2] A. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions and Reversals,” *Soviet Physics Doklady*, Vol. 10, No. 8, pp.707-710, 1966.

[3] L. R. Bahl, F. Jelinek and R. L. Mercer, “A Maximum Likelihood Approach to Continuous Speech Recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. PAMI-5, No.2, pp.179-190, March 1983.

[4] J. Fiscus, “A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER),” in *Proc. ASRU 1997.*

[5] V. Goel and W. Byrne, “Minimum Bayes-risk Automatic Speech Recognition,” in Pattern Recognition in Speech and Language Processing, edited by W. Chou and B. H. Juang, *CRC Press*, 2003.

[6] L. Mangu, E. Brill and A. Stolcke, “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks,” *Computer Speech and Language,* Vol. 14, pp.373-400, 2000.

[7] K. Na, B. Jeon, D. Chang, S. Chae, and S. Ann, “Discriminative Training of Hidden Markov Models using Overall Risk Criterion and Reduced Gradient Method,” in *Proc. EUROSPEECH 1995.*

[8] J. Kaiser, B. Horvat, Z. Kacic, “A Novel Loss Function for the Overall Risk Criterion Based Discriminative Training of HMM Models,” in *Proc. ICSLP 2000.*

[9] J. Kaiser, B. Horvat, Z. Kacic (2002). “Overall Risk Criterion Estimation of Hidden Markov Model Parameters,” *Speech Communication*, Vol. 38, pp.383-398, 2002.

[10] D. Povey and P. C. Woodland, “Minimum Phone Error and I-smoothing for Improved Discriminative Training,” in *Proc. ICASSP 2002.*

[11] D. Povey. Discriminative Training for Large Vocabulary Speech Recognition. *Ph.D Dissertation, Peterhouse, University of Cambridge*, July 2004.

[12] Y. Normandin. Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem. *Ph.D Dissertation, McGill University, Montreal,* 1991.

[13] L. Wang and P. C. Woodland, “MPE-Based Discriminative Linear Transform for Speaker Adaptation,” in *Proc. ICASSP 2004*.

[14] R. D. Schafer. An Introduction to Nonassociative Algebras. *New York: Dover*, 1996.

[15]B. Chen, J. W. Kuo, W. H. Tsai, “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” in *Proc. ICASSP 2004.*

[16] S. Ortmanns, H. Ney, X Aubert, “A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language*, Vol. 11, pp.11-72, 1997.

[17] LDC: Linguistic Data Consortium. http://www.ldc.upenn.edu

[18] S. M. Katz, “Estimation of Probabilities from Sparse Data for Other Language Component of a Speech Recognizer,” *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 35, No.3, pp. 400-401, 1987.

[19]A. Stolcke, “SRI language Modeling Toolkit,” version 1.3.3, http://www.speech.sri.com/projects/srilm/.

[20] H. M. Wang, B. Chen, J.-W. Kuo, and S.S. Cheng. “MATBN: A Mandarin Chinese Broadcast News Corpus,” *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 10, No. 2, pp. 219-236, 2005.

[21] SLG: Spoken Language Group at Chinese Information Processing Laboratory, Institute of Information Science, Academia Sinica. http://sovideo.iis.sinica.edu.tw/SLG/index.htm

[22] PTS: Public Television Service Foundation. http://www.pts.org.tw

[23] Y. Normandin, R. Lacouture, R. Cardin, “MMIE Training for Large Vocabulary Continuous Speech Recognition,” in *Proc. ICSLP 1994*.

[24] J. W. Kuo and B. Chen, “Minimum Word Error Based Discriminative Training of Language Models,” in *Proc. EUROSPEECH 2005*.

[25] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, Geoffrey Zweig, “fMPE: Discriminatively Trained Features for Speech Recognition,” in *Proc. ICASSP 2005.*

基於統計與佚代的中英雙語詞及小句對應演算法

黃子桓　　高照明
臺灣大學資訊工程學系　臺灣大學外國語文學系
tzhuan@csie.org　　zmgao@ntu.edu.tw

Abstract

本文提出一基於段落對應的雙語語料中迭代進行詞對應及小句對應 (subsentential alignment) 之模型， 並提出可行的實作方式。和基於句對應的詞對應演算法相比， 本文提出的演算法不須經過句對應，在現實應用有更大的彈性。 而和基於字典的句對應演算法相比，本文提出的演算法不須額外的字典支援， 完全藉由本身的統計資訊進行詞條的蒐集和利用。 實驗結果顯示詞對應和 K-vec 相比有較佳的 precison 和 recall 值， 而小句對應結果顯示約有 77.74% 的小句對應是完全或部份正確。

1　導言

語言翻譯在資訊的傳遞上扮演十分重要的角色，在過去，語言翻譯的工作皆以人工翻譯爲主。 由於電腦科學的進步，運算能力大幅提高，各類相關的理論、演算法也相繼被提出， 如何利用電腦來進行自動翻譯工作成了重要的研究課題。 在許多自動翻譯的研究中，詞對應 (word alignment) 是不可或缺的重要步驟， 其正確率往往對翻譯的結果有關鍵性的影響。 傳統詞對應乃是由人工所建立，如雙語詞典即是人工建立的詞對應資料庫。 但人工建立不但費時費力，難以跟上新詞增加的速度， 且詞典有其極限，再完善的詞典皆不可能包含所有雙語詞彙對應。 加以現今網路上已有大量的雙語機讀資料，在研究資料充沛的情形下， 由電腦自動建立詞對應亦爲一重要的研究方向。

現有的電腦自動建立詞對應研究中，有許多是基於已正確句對應的研究， 並且取得不錯的研究成果。 然而要達到正確句對應並不容易，以人工標示費時費力， 且在現實環境中，並不保證有正確句對應。而以機器自動句對應的演算法， 基於語言的特性，不同性質的文章，其正確率的變動非常大 (McEnery and Oakes, 1996)， 與廣泛應用的水準尚有差距。 以詞對應的角度來看，正確的詞對應有助於句對應；而從句對應來看， 正確的句對應對詞對應也是十分正面的助益。 句對應和詞對應可說是雞生蛋、蛋生雞的問題。 本研究主要探討如何在同一語料中同時進行句對應及詞對應，並藉由彼此提高正確率。 我們選擇使用已正確段落對應的中、英語料庫，理由如下：

1. 基於翻譯的習慣，以句爲單位來看，往往會有增減的情形，但若以段落爲單位來看， 則較少有增減的情形。因此一般的翻譯文章或者已正確句對應， 或者經過極少的工作即可達到正確段落對應，對於現實的應用有很大的幫助。
2. 由於語言的結構關係，段落與段落間往往有利於機器處理的分隔符號存在， 因此機器自動分段可達 100% 正確。因此使用正確段落對應的語料庫， 在分段上幾乎不會有失敗的情形發生。

2 相關研究

2.1 詞對應演算法

Fung 與 Church (1994) 的 K-vec 演算法將雙語語料庫各切分爲相等的 K 區塊， 每一詞皆記錄該詞在 K 個區塊中出現與否， 組成一 K 維 vector $(v_1, v_2, \ldots, v_K)$，$v_i \in \{0, 1\}$。 對雙語的兩兩詞彙，皆透過彼此 vector 計算各自頻率及共同出現於相同區塊的頻率， 並以 MI (Mutual Information) 來計算兩詞彙的相依程度。 由於 MI 對於頻率甚少的詞會計算出極大值，嚴重影響可信度， 因此藉由 t-score 值修正，透過給定的常數值，忽略 t-score 值小於該常數值的結果，將大大提昇 MI 的可信度。

由於 K-vec 演算法需要切割雙語語料成 K 個區塊，錯誤的切割將使結果不如預期。 因此 Fung 與 McKeown (1994) 再提出 DK-vec 以解決此問題。 在 DK-vec 中每一詞彙皆記錄兩 vector， position vector 記錄該詞彙出現於雙語語料的所有位置， recency vector 則記錄兩兩位置的距離。 以 position vector 的資料爲橫座標值，recency vector 的資料爲縱座標值， 並連接相鄰之點，則可得一分佈於 2-D 座標系的函式分佈取樣圖。 利用 pattern matching 的 Dynamic Time Warping 的技術， 可計算兩兩詞彙函式分佈取樣的相似程度， 從相似程度的高低可得兩詞彙的相依值。

Melamed (1998) 的 Competitive linking algorithm 是基於已正確句對應的詞對應演算法。 對於雙語的兩兩詞彙，competitive linking algorithm 使用 LLR (Log-Likelihood-Ratio) 來評估其相依程度。 當一雙語對應句中兩兩詞彙的相依權值皆計算完畢， 將所有雙語詞彙對由相依權值大至小排序，依序取出雙語詞彙對，若兩詞彙皆未與其它詞彙連結， 則連結此兩詞彙，否則忽略並處理下一詞彙對，直到所有的詞彙對皆已處理完畢。

2.2 句對應演算法

雙語句對應的研究開始於 90 年代初期。 Gale 與 Church (1991) 及 Brown 等 (1991) 觀察到長句的翻譯對應句一般而言較長， 而短句的翻譯句通常較短。他們利用句長的關連性配合動態規劃或 EM 演算法得到 96% 以上的正確率。 Gale 與 Church (1991) 及 Brown 等 (1991) 兩者最大的差別是前者透過人工先得到先驗機率 (prior probability) 而後者利用 EM 演算法得到相關的參數。Wu (1994) 及 Xu and Tan (1996) 以句長爲主結合一個包含日期及數字等訊息小的辭典得到 96% 的正確率。 以句長爲基礎的統計方法的優點是不需要語言知識及辭典就可以運作。 缺點是如果語料中含有豐富的多對多的句對應關係， 或是翻譯的語料中有增添或刪減的現象發生就會造成正確率大幅下降。 前述幾項研究由於大都採用議會的紀錄， 例如 Gale 與 Church (1991) 及 Brown 等 (1991) 用加拿大國會 Hansard 英法平行語料， Wu (1994) 則利用香港立法局議會質詢與答詢的中英平行語料， 由於是口語紀錄所以句子較短，且不少是一對一對應。 Gale 與 Church (1991) 統計 Hansard 語料 80% 以上是一對一的對應關係， 罕有多對多的對應關係或增添或刪減的情形發生， 所以以句長爲主的統計方法得到很好的效果。 但 McEnery and Oakes (1996) 以 Gale 與 Church (1991) 的方法做實驗卻顯示此種演算法的正確率對不同的文類與語言會產生很大的差異。 例如波蘭文英文平行語料的正確率因文類不同介於於 100% 與 64.4%， 而他們所實驗的中英新聞平行語料更低於 55%， 這證明單純以句長關連性顯然無法得到高正確率。

另一個不需要辭典的方法是 Kay and Röscheisen (1993) 以詞彙的頻率 (去除低頻的詞及高頻

的詞) 及在文章中出現的分佈， 建立可能的詞對應表及句對應表並不斷的修正， 以 relaxation 方法達到收斂。 與 Gale 與Church (1991) 及 Brown 等 (1991) 方法一樣， Kay and Röscheisen (1993) 的方法只有在一對一的情形佔絕大多數時才會有好的效果。 此外此種方法過度重視詞頻，文章的長度太短會造成正確率的大幅下降。 這個演算法另一個實做上的問題是處理十分耗時，無法快速處理大量語料。

子句對應 (clause alignment) 則如 Kit et al. (2004) 以雙語法律條文的 glossary 和雙語辭典，再加上適當的標點符號轉換、 數字轉換 (如阿拉伯數字與羅馬數字)，再設計一估計函數來結合全部資訊而得相似程度， 以其評估字句對應，可達 94.6% 的正確率。 Kit et al. (2004) 以詞彙訊息得到非常高的小句對應正確率的主要原因是所用的語料爲法律雙語文件且使用法律術語的辭典， 且此類文件中代表法律條文的數字一再出現。 在我們之前的實驗 (林與高 (2004)) 顯示在一般的中英雙語文章使用雙語辭典、數字、 及標點訊息在大句的對應正確率尚且不到 90%， 小句的正確率必定無法達到 Kit et al. (2004) 的水準。

Wu et al. (2004) 則提出利用句長和標點符號進行小句對應 (subsentential alignment)， 加上雙語中的同源資訊 (如雙語中相同的數字部份)， 以香港立法局會議記錄爲實驗資料，可達 98% 的正確率。 Wu et al. (2004）所用的英漢對譯語料爲香港立法局的議會紀錄， 內容全是議員與官員之間一問一答的紀錄。此類議會語料多屬逐句翻譯， 且少有意譯的情形，由於採一問一答及逐句翻譯在句對應及小句對應比較容易。 如用文章之類的對譯語料該演算法勢必無法得到如此高的正確率。

3 段落對齊平行語料的詞對應暨小句對應演算法

3.1 段落對齊平行語料的詞對應演算法

在 Association-based binlingual word alignment 中， 詞彙的出現頻率扮演著關鍵的角色。 不論使用 MI、t-score 或者 LLR 來評估兩詞彙的相依程度， 皆利用頻率的資訊來計算。 而另一關鍵的角色則是文章的切分。 將文章切分成若干區塊，提供了一個強烈的假設及限制， 即該詞彙若有詞對應，必然出現於同一區塊中； 正確的切分方式，能使詞彙的相依程度提高，反之則會降低。 基於上述說明，我們設計一演算法， 在現有的區塊中，尋找一切分方式，可使總體相依值提高最多， 不斷重覆此一過程直到所有切分方式都無法再使總體相依值提高。

令 $E = B_1^e B_2^e \ldots$、$C = B_1^c B_2^c \ldots$， 其中 $B_i^{e|c}$ 表示一英文 (中文) 區塊， 稱此時的切分狀態爲 Ω。 令 $B_i^e = e_{i,1} e_{i,2} \ldots$，$B_j^c = c_{j,1} c_{j,2} \ldots$， 其中 $e_{i,k}$ ($c_{j,l}$) 表示一英文 (中文) 詞彙。 令 $asso(e, c)$ 表示詞彙 e 和詞彙 c 的相依權值大小 (此相依權值可視需要選用如 MI、t-score、LLR 等。 在本實驗中我們以 MI 爲主，搭配 t-score 以過濾詞頻低的對應)， 則 $ASSO(\Omega) = \sum_i \sum_j asso(e_i, c_j)$ 即爲在 Ω 切分狀態下的總體相依值。 令 $new(\Omega, i, start_e, end_e, start_c, end_c)$ 表示一種新的切分狀態， 其意義爲在 Ω 切分狀態中，第 i 區塊被切分了， 切分方式爲 $e_{i,start_e} e_{i,start_e+1} \cdots e_{i,end_e}$ 和 $c_{i,start_c} c_{i,start_c+1} \cdots c_{i,end_c}$ 爲一組對應區塊， 而 $e_{i,1} e_{i,2} \ldots e_{i,start_e-1} e_{i,end_e+1} \cdots e_{i,|E_i|}$ 和 $c_{i,1} c_{i,2} \cdots c_{i,start_c-1} c_{i,end_c+1} \cdots c_{i,|C_i|}$ 爲另一組對應區塊。 因此，對 Ω 狀態而言，計算

$$value = \max_{\substack{1 \le start_e \le |E_i| \\ 1 \le start_c \le |C_i| \\ start_e \le end_e \le |E_i| \\ start_c \le end_c \le |C_i|}} ASSO(new(\Omega, i, start_e, end_e, start_c, end_c)) \quad \forall i = 1, 2, \ldots, |\Omega|$$

若 $value > ASSO(\Omega)$，即表示該切分方式能夠提高總體相依值， 依此時之 $start_e$、 end_e、 $start_c$、 end_c 進行切分， 可得一新的切分狀態 Ω'。若 $value \le ASSO(\Omega)$， 表示所有的切分方

式都無法再提高總體相依權值，因此該區塊沒有再被切分的必要。 重覆此一步驟，則切分之區塊數將會不斷增加， 直到所有的區塊都無法再被切分。

演算法如下：

1. 以雙語語料的段落對應作爲初始切分狀態。
2. 在目前切分狀態 Ω 中，對每一區塊進行切分嘗試，並記錄新的切分方式於 Ω'。
3. 如果 $|\Omega| = |\Omega'|$ 則結束，否則回到 2.。
4. 利用目前切分狀態求出詞彙間的相依權值並輸出結果。

3.2 段落對齊平行語料的詞對應暨小句對應演算法

在上述演算法中，如果加上特殊的限制條件，則可使切分區塊的自由度降低，形成特定的區塊。 例如限制 $start_{e|c}$ 的前一個詞必須是分句符號 (如句號、問號、驚嘆號等)， $end_{e|c}$ 後一詞也必須是分句符號，則所得的區塊對將成爲句對應或多句對應形式。 亦即此演算法爲一詞對應暨句對應之演算法。

3.3 加速與實作

在上述演算法中，由於要對所有可能的切分方式計算 $ASSO$ 值， 亦即對於所有可能的切分方式都要執行一次類似 K-vec 的演算過程， 則此演算法的計算複雜度將會十分地高， 在實作上雖然並不困難，但計算時間將會十分地久。 而若是加上對 $start_{e|c}$ 及 $end_{e|c}$ 的限制，將會有效減少可能切分方式的總數。 然而計算時間仍然相當長，因此難以取得廣泛應用。 在此我們提出一個加速的作法。

考慮上述的理論架構，對於每個可能的切分方式都要重新計算 $ASSO$ 值， 顯然付出太大的代價。 重新計算 $ASSO$ 值的理由在於這是一個足夠好、可信賴的評估方式， 可有效評估現行分割方式的優劣。 因此加速的關鍵即在於使用新的評估方式， 新的評估方式需滿足下列條件：

1. 和 $ASSO$ 相比同樣可被信賴。
2. 計算複雜度要低。

我們所提出的新評估方式說明如下：

在一個區塊中，我們可以指定的標點符號 (例如句號、問號等) 將區塊再切分爲較小的區塊 (可能包含一或多個句子)，稱爲子區塊。令 $B^e = S_1^e S_2^e \dots S_m^e$ 和 $B^c = S_1^c S_2^c \dots S_n^c$，其中 B^e 和 B^c 爲雙語語料中對應的其中一區塊； 而 $S_i^{c|e}$ 表示子區塊。令 $W_i^{e|c} = \{w_{i,1}^{e|c}, w_{i,2}^{e|c}, \dots\}$ 表示在 $S_i^{c|e}$ 子區塊中，所有相異詞彙所成的集合。 定義

$$score(S_i^e, S_j^c) = \sum_{e \in W_i^e} \sum_{c \in W_j^c} asso(e, c)$$

其中 $asso(e, c)$ 的定義如前所述。 則藉由求出

$$(max_e, max_c) = arg \max_{\substack{1 \le i \le |B^e| \\ 1 \le j \le |B^c|}} score(S_i^e, S_j^c)$$

可知，在現行條件下，$S^e_{max_e}$ 對應 $S^c_{max_c}$ 是最可信賴的。 因此，將 $S^e_{max_e}$ 與 $S^c_{max_c}$ 取出使成新的區塊。 對於每一區塊，重覆這個步驟，直到區塊的總數不再變動。

$asso(e, c)$ 乃根據目前爲止的詞對應相依權值來計算； 而接著詞對應乃根據新的切分狀態來求其相依權值。交互迭代後將會收斂， 亦即兩者皆不再變動。由於我們的切分是以標點符號切分的子區塊爲單位， 因此若目前處理區塊只包含一個子區塊，則不可再被區分， 因此該演算法保證會收斂。 而初始的段落對應則用於提供最初的相依權值計算， 此外也保證初始的區塊對應是完全正確的。

4 實驗材料

本研究使用的中英對譯文章取自光華雜誌 (http://www.sinorama.com.tw/ch/)，統計資料如下：

	段落數	總詞數	相異詞數
中文	59	3291	1192
英文	59	3908	1082

英文分詞以空白和標點符號爲主，搭配常見縮寫詞以減少分詞錯誤。 計算英文相異詞時則以一般變化規則 (-s -ing -ed 等) 加上不規則動詞變化表來還原各詞類原形。 中文分詞則以中央研究院中文斷詞系統 (http://ckipsvr.iis.sinica.edu.tw/) 來進行分詞。 該系統提供線上使用，爲現階段中文分詞正確率最高的系統之一。

5 實驗結果與討論

我們實作了我們所提出的演算法，並實作 K-vec 演算法以進行比較。 在 K-vec 演算法中，由於作者建議切分區塊數 $K = \sqrt{\text{total word number}}$ 會有較好的結果， 而在我們的實驗資料中，段落數的平方 ($59 \times 59 = 3481$) 恰約等於總詞數 (中文 3291、英文 3908)， 因此我們以段落作爲 K-vec 的區塊。 相依權值使用 MI 及 t-score 判別， MI 及 t-score 的參數比照原作者的建議：以 t-score 值爲篩選器， 只考慮 t-score ≥ 1.65 的詞對應。MI 則做爲主要相依權值的依據， 輸出時以 MI 的值由大到小排序，並捨棄 MI < 1.0 的結果。 在這個條件下的輸出如下表所示：

演算法	t-score 篩選值	MI 最小值	詞對應數	正確數	precision
K-vec	1.65	1.0	28	12	0.42
ours	1.65	1.0	79	32	0.40

雖然在 MI ≥ 1.0 的輸出條件下，我們的演算法 precision 較低， 但由 Figure1 和 Figure 2 可看出在同樣個數的輸出 (輸出皆以 MI 的權值大小爲序)下，我們的演算法有較好的表現。 在輸出前 10 條詞對應時，我們的演算法和 K-vec 差異不大， 但從第 10 條詞對應之後的輸出結果，明顯我們的演算法有更高的正確率， 到前 30 項輸出仍維持 0.6 以上的 precision， 而在前 30 項輸出時 K-vec 的 precision 僅約 0.45。

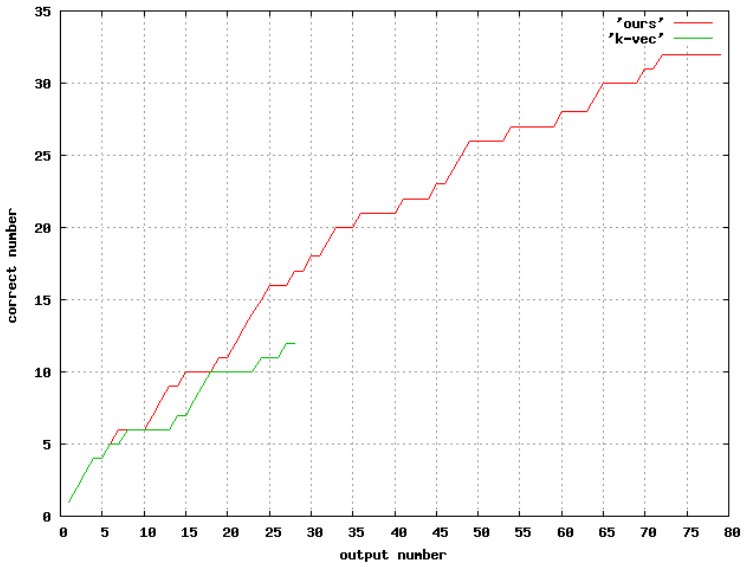


Figure 1: 輸出結果數與正確數關係圖

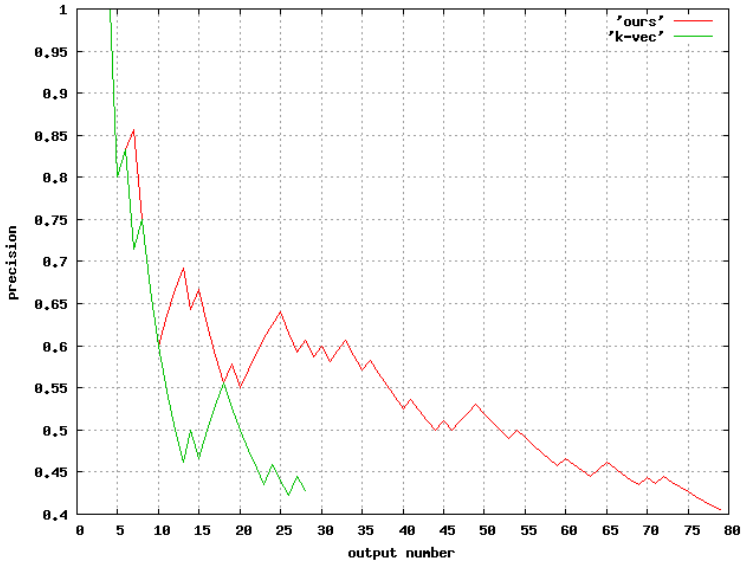


Figure 2: 輸出結果數與 precision 關係圖

由下面列舉的詞對應結果可以看出，部份正確的詞對應佔了相當的比例， 如果加上這些複合詞的擷取演算法，則可望大幅提升 precision。 此外，由於基於統計的相依權值計算相當依賴詞頻，過多或過少都會影響其信心。 以本實驗爲例，詞頻甚低的正確詞對應有可能無法通過 *t*-score 的篩選門檻； 而詞頻不夠高的功能詞對應也可能仍有相當高的 MI 值以致於未被過濾。 詞頻太低是所有基於統計的詞對應都會面臨的困難問題， 因爲過低的詞頻並沒有辦法分辨其爲偶然或是正確； 而功能詞的部份可用預先建立的功能詞列表來解決。

在 recall 方面，由於 recall 的計算需要以人工找出雙語語料中所有正確的詞對應， 在此實驗資料中，共有 1192 個相異中文詞、1082 個相異英文詞， 基於時間及人力的關係無法以人工標記此實驗資料的所有正確詞對應， 然仍可知在分母相同的條件下，我們的演算法有較高的 recall 值。

由於翻譯的關係，雙語對譯的用詞可能相當靈活， 例如同一個動詞卻在不同位置用不同的詞彙翻譯， 以致於所有的對譯詞頻都不高，而無法找出正確詞對應。 由於此原因，許多詞彙無法找出正確詞對應， 而最利於找出的則是有固定翻譯及適當詞頻的專有名詞。

以下試列舉出前 18 條由我們提出的演算法所得到的詞對應結果， 其正確對應與否於附註中說明，若部份正確則在附註中顯示正確之詞對應。

asso	英文詞	中文詞	附註
6.363	table	桌	正確
5.948	Kuo	郭慧明	正確
5.626	hope	希望	正確
5.141	Jen-an	人安	正確
4.948	each	天	each day 每天
4.877	volunteers	義工	正確
4.778	day	天	正確
4.778	goal	努力	錯誤
4.725	fund	經費	正確
4.626	Jen-an	基金會	The Jen-an Foundation 人安基金會
4.626	each	每	正確
4.626	month	月	正確
4.533	even	甚至	正確
4.488	but	長期	錯誤
4.404	welfare	社福	social welfare 社福
4.247	social	社福	social welfare 社福
4.141	elderly	失	elderly people 三失老人
4.041	day	每	each day 每天

在分句演算法方面，我們的原始語料共有 59 段落， 在實作中，我們指定這些符號 . , ; ! ? 。 ， ； ！ ？ 作爲切分區塊的標點符號限制。 經過我們的演算法，最後收斂時共輸出 265 個對應區塊。 下表是經由我們的演算法的分句結果與原始文章以 . ; ! ? 。 ； ！ ？ 分句的比較結果， 中文部份我們用兩組標點符號來分句，其中一組包含逗號，另一組不包含：

		句數	句平均詞數	標準差
原始文章以標點分句	中文 (使用逗號)	396	8.300	4.166
	中文	104	31.615	18.074
	英文	165	23.666	12.524
以本演算法分句	中文	265	12.384	9.683
	英文	265	14.709	9.172

在利用 regular expression 或其它方法解決英文縮寫點 (如：Mr. 或 I.B.M.) 的問題之後， 英文基本上可以靠句號，問號，驚嘆號，分號當作分隔句子的界限。 中文的句子無法像英文一樣靠標點符號來判斷。 原因是逗點在中文使用的非常的鬆散， 逗號和句號的使用是作者風格的問題而非文法的問題。 如果用句號、問號、驚嘆號、分號來分的話， 很多是比句子更大的言談單位 (discourse) ， 如果加上逗號的話又會造成許多只是詞組而不是句子。 這就是爲什麼當我們用 。 ！ ？ ；來分割句子時， 中文句數比英文句子少很多， 而中文加逗點作爲分隔句子界限之後又比英文句子多很多的原因。 從以上的討論，我們可以看出經過我們的分句演算法所得到的是比句子還要小的區塊， 可視爲一種小句對應的結果。

由於我們的演算法並不保證按順序對應， 因此輸出結果並不按原始文章的順序。另外基於我們演算法的特性， 不相鄰的區塊有被合而爲一的可能。因爲上述原因， 要對輸出的對應區塊分析其正確分句程度極爲困難。 因此我們採用較簡單的估計方式， 以人工標記 265 句中完全正確、部份正確及完全錯誤的小句對應， 完全正確表示該對應是最小可能的切分方式，例如「 E_iE_{i+1} 正確對應 $C_jC_{j+1}C_{j+2}$ 」即表示不論是 E_iE_{i+1} 或 $C_jC_{j+1}C_{j+2}$ 皆無法再切割以得到更小的正確對應。部份正確對應以上述正確對應爲例， 任意 $\{E_i, E_{i+1}\}$ 的子集合對應任意 $\{C_j, C_{j+1}, C_{j+2}\}$ 的子集合都可視爲部份對應。若非上述兩種情況，則稱爲錯誤對應。 實驗的結果統計如下：

	對應數	所佔全體比例
總對應數	265	-
完全正確	59	22.26%
部份正確	147	55.47%
完全錯誤	59	22.26%

由於我們的小句對應是基於「完全對應」，即任一小區塊皆必對應於某區塊， 且僅對應於該區塊。因此任一區塊若爲部份正確， 則必然會影響另一區塊爲部份正確或完全錯誤， 因此部份正確數佔了極大比例是可預期的結果。

以下試舉部份小句對應結果：

英文區塊	中文區塊
完全正確	
even into his old age	甚至 在 遲暮 之 年
whenever cswf has needed them	在 創世 有 需要 時
the service hua-shan offer the elderly are of two variety	華山 照顧 老人 的 方式 有 兩 種

部份正確	
it is also renowned as a “master fundraiser” and admire by other social welfare organization for operate at a surplus year after year	還被喻爲「募款高手」
not only is cswf well known for its service how do they do it	他們是怎麼做到的
we finally reach the pvs hospice	來到植物人安養中心：創世的發源地
完全錯誤	
at the start	幫幾個家庭喘口氣而已
cswf has open branch hospice around the country	創世的目標是全省２３個縣市都有植物人安養院
thus far they have complete 13	籌備中的有４個

6 結論與未來研究方向

我們的研究展示了一個不必依賴正確句對應，也不必依賴字典的迭代詞與小句對應演算法。 相較於依賴句對應的詞對應演算法，我們的演算法不必經過人工或機器的句對應， 可有效減少工作量，並且避免了由錯誤句對應所引發的錯誤。而相較於依賴字典的句對應演算法， 我們的演算法如同一邊分句一邊建立小型字典，除了不需額外資料庫外， 對於字典沒有的新字我們的演算法仍能透過統計的方式得到相依關係， 因此擁有較大的彈性可適應不同種類的文章。

在實驗結果裡，和同樣不需已句對應的 K-vec 演算法相比，我們提出的方法有較佳的 precision 值， 且在同樣的條件下能找出更多正確的詞對應，亦即有較佳的 recall 值。 而在句對應中，結果顯示有許多輸出是詞組和子句的對應， 換言之我們的演算法能得到小句對應， 這是目前大部分基於統計演算法不容易做到的。

未來的研究方向擇要列舉如下：

1. 目前我們的模型僅標示出一對一的詞對應， 實際上詞對應有很大的機會是多對多，尤其是具有特定翻譯的專業詞彙。 對於這些複合詞，若能在迭代過程中取出，則可增加詞對應的信心， 進而對句對應有正面的助益。 因此如何利用此模型來運用複合詞資訊，將是未來研究的方向之一。
2. 由於演算法的特性，在切分的情形下，會將原本不連續的區塊合併。 對詞對應而言，這個合併的動作並不會造成太大的影響， 但對句對應而言此動作並不恰當。而這個問題可透過修正的切法方法來解決，例如， 當找到最有信心的子區塊對應時，將該區塊切分成三組新對應而不是兩組， 可避免合併的動作。
3. 本研究基於兩前提：正確段落對應及正確分詞。 由於現實語料的支援，正確段落對應可視爲合理的假設，分詞對英文而言也有很高的正確率， 然而分詞對中文而言遠較英文困難，正確率也遠不及英文。 錯誤的分詞結果將改變詞頻，對詞對應的結果有很大的影響。 如何降低對分詞正確性的依賴是我們未來研究的課題。

4. 本研究的理論模型乃「不可迴溯式」，如果在過程中發生錯誤的切分， 則該錯誤會永久保留，甚至可能會擴散。雖然過程中每個步驟都儘可能選取最有信心的切分方式， 但不可避免一定有發生錯誤的可能。如果能在現有模型上加入可事後補救的機制， 將可使穩定度更爲提升。

5. 除了經統計所得的訊息外，在一般的雙語語料中常常還有其它的訊息可供利用， 例如數字、未翻譯的人名、地名、專業詞彙等，這些訊息比統計所的的詞對應更爲可靠， 因此在我們提出的演算法中結合這類訊息的使用，我們預期能得到更好的結果。

致謝

本研究得到國科會 NSC93-2815-C-002-063H 「從中英平行語料庫自動擷取雙語詞組知識」及 93-2411-H-002-013 「 詞彙語意關係之自動標注—以中英平行語料庫爲基礎(3/3)」經費補助，特此致謝。

參考資料

[1] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In *Meeting of the Association for Computational Linguistics*, pages 169–176, 1991.

[2] R. Catizone, G. Russell, and S. Warwick. Deriving translation data from bilingual texts. *Proceedings of the First Lexical Acquisition Workshop*, 1989.

[3] B. Chang, P. Danielsson, and W. Teubert. Extraction of translation unit from chinese-english parallel corpora. *COLING-02: The First SIGHAN Workshop on Chinese*, 2002.

[4] P. Fung and K. Church. K-vec: A new approach for aligning parallel texts. *COLING-94: 15th International Conference on Computational Linguistics*, pages 1096–1102, Aug 1994.

[5] P. Fung and K. McKeown. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. *AMTA-94, Association for Machine Translation in the Americas*, pages 81–88, 1994.

[6] W. Gale and K. Church. A program for aligning sentences in bilingual corpora. *Proceedings of the Annual Conference of the Association for Computational Linguistics*, pages 177–184, 1991.

[7] M. Kay and M. Röscheisen. Text-translation alignment. *Computational linguistics*, (1):121–142, 1993.

[8] C. Kit, J. J. Webster, K-K. Sin, H. Pan, and H. Li. Clause alignment for hong kong legal texts: A lexical-based approach. *International Journal of Corpus Linguistics*, (1):29–51, 2004.

[9] I. D. Melamed. Models of co-occurrence. *IRCS Technical Report*, 1998.

[10] I. D. Melamed. Models of translational equivalence. *Computational Linguistics*, pages 221–249, 2000.

[11] Robert C. Moore. Association-based bilingual word alignment. *Proceedings, Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, Ann Arbor, Michigan*, pages 1–8, 2005.

[12] D. Wu. Aligning a parallel english-chinese corpus statistically with lexical criteria. *Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics*, 1994.

[13] Jian-Cheng Wu, Thomas C. Chuang, Wen-Chi Shei, and Jason S. Chang. Subsentential translation memory for computer assisted writing and translation. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 106–109, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[14] D. Xu and C. L. Tan. Automatic alignment of english-chinese bilingual texts of cns news. *Computational Linguistic Archive*, 1996.

[15] 林語君 and 高照明. 結合統計與語言訊息的混合式中英雙語句對應演算法. *ROCLING*, 2004.

電視新聞語料場景的自動切割與分類

姜柏亘 [1]，呂仁園 [1]，楊博厚 [2 3]，謝鴻文 [2]

1. 長庚大學資訊工程研究所
2. 長庚大學電機工程研究所
3. 中央研究院資訊所

E-mail: rylyu@mail.cgu.edu.tw, TEL:886-3-2218800ext5967

摘要

在本篇論文中，我們提出場景自動切割與分類的演算法，我們將一小時新聞分成爲四種場景：新聞主播報導（Anchor Reporting）、現場採訪報導（Live Reporting）、氣象主播報導(Weather Anchor Reporting)與廣告（Commercials）。我們擷取了時域與頻域的特徵值用以描述場景的特性，並使用高斯混合模型(Gaussian Mixture Model)當作場景分類器。場景切割的策略有兩種：(1)每秒移動策略、(2)快速策略。每秒移動策略，是利用每次移動一秒，並觀察3秒的聲音去決定場景的轉換點，效能評估方面，其Deletion Rate爲5.56%，Insertion Ration爲5.56%。由於上述的方法耗費計算的時間較久，因此我們也開發了一套快速策略，其Deletion Rate爲2.27%，Insertion Ration爲5.4%。在場景分類方面，我們使用了MFCC、LSTER、HZCRR、SF與MFS去將經過真實轉換點切割出的一段段聲音去作分類，可以達到92.5%的平均正確率。

1. 簡介

隨著網際網路的蓬勃發展，越來越多的新聞資訊可以直接從網路下載。而新聞資訊裡富含語音、音樂、文字、顏色樣型及影像圖形。雖然人類可以快速的透過觀察來解釋這些內容的含意，但是透過電腦分析去瞭解其內容還是處於初步的階段。新聞資訊的檢索、分析應該也要像我們人類的頭腦一樣去處理，換句話說就是在作處理前應先透過電腦先分析及瞭解其內容。假設我們要搜尋某一主題的新聞片段，我們必須把有關這個主題的整個聲音片段及文字資訊列舉出來，然而傳統的語音辨認系統並無法藉由文字資訊來切割出主題式的片段，因此考慮到場景轉換的語音切割與分類方法便是需要且直觀地，而瞭解場景內容對於以內容爲基礎的新聞資料庫索引與檢索是相當重要的。近幾年越來越多的研究在這領域努力。

一般來說，研究場景的切割與分類可以使用Model-based segmentation及Metric-based segmentation，其中Model-based segmentation的方法是將不同的場景聲學群組(acoustic class)建立不同的模型，例如高斯混合模型或隱藏式馬可夫模型等。舉例來說，若我們要切割電視新聞的話，我們便可以爲棚內主播、外場記者、外場受訪者、氣象主播等建立個別的模型，之後測試的聲音透過Model Testing便可以依照既有模型去算出此分析音框的Maximum Likelihood，進而可決定轉換點。另一方面，Metric-based segmentation的方法是利用距離量測的概念，選擇某一相異度量測公式，計算相鄰兩個frame的相異度，並決定一個門檻值去決定轉換點，而常用的相異度量測公式有KL distance、Common Component GMM-based Divergence[2]、Delta BIC[1]等。

Hsin-min Wang [1]收集了公共電視新聞語料，並利用Bayesian Information Criterion定義一個相異度量測的方法去偵測環境或是語者的轉換點。Yih-Ru Wang [2]則是使用GMM來描述相異兩個

聲音片段的統計特性，利用共用的mixture component來減少估計混合權重的計算量，以估計出權重向量來代表聲音片段的特性，進而量測相鄰聲音片段間的相異度，決定可能的轉換點。Tong Zhang[8]則是使用四種特徵值：平均過零率、能量、基礎頻率與頻譜鋒的追蹤(spectral peak tracks)，並設計一套有規則的策略將電視audio訊號分成語音、音樂、環境聲音、含音樂背景的語音、含音樂背景的環境聲音與靜音等，正確率可達到90%以上。Lekha Chaisorn [9]使用多個特徵值與技術將影片分析成一個個shot與場景，在shot階段，配置一個選擇樹去分類shot到13種與先定義的類別其中一種。Zhu Liu利用12種音訊特徵值，並結合神經網路分類器（neural network classifier）[4]與隱藏式馬可夫模型(HMM)[3]將電視節目場景分爲廣告、籃球賽、足球賽、新聞及氣象報告。Lie Lu [6]使用梅爾倒頻譜參數、過零率、能量、亮度與頻寬（Brightness and bandwidth）、頻譜流量、頻帶週期性（Band periodicity）、噪音音框比率（Noise frame ratio）並結合支援向量機（Support vector machine）將聲音串流切割分類爲靜音、音樂、背景聲音、純語音、含有音樂的語音。

論文接下來的架構：第 2 節描述了特徵值擷取與分析，第 3 節研究整個系統的切割與分類演算法流程，第 4 節爲實驗的效能評估與分析，第 5 節爲結論與未來展望。

2. 特徵值擷取

2.1 梅爾倒頻譜參數(Mel-frequency cepstral coefficients)

$$Cn = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} (\log S_K) \cos[n(k-0.5)\pi / K] \quad n = 1,2,3...L \tag{2.1}$$

2.2 短時距低能量比率(Low Short Time Energy Ratio)

我們使用能量的變化率當成我們的特徵向量的成分，而不是準確的短時距能量值。我們使用短時距低能量比率(LSTER)去表示短時距能量的變化率。

$$LSTER = \frac{1}{2} \sum_{n=0}^{N-1} [\mathrm{sgn}(0.5 avEng - STE(n) + 1)] \tag{2.2}$$

其中 n 代表音框索引，N 代表一秒內的音框總數，sgn[.]是符號函式，以及 Eng(n)代表在第 n 個音框的能量，avgEng 是一秒內的能量平均值。

LSTER 是一個很有效的特徵，特別是在區分語音與音樂。通常，在語音中有許多的靜音，所以測量 LSTER 的值會高於音樂。下圖 2.5 代表 LSTER 的機率分佈曲線：(a)代表語音 (b)代表音樂。

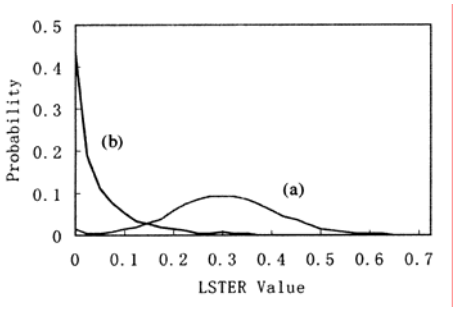


圖 2.1 LSTER 的機率分佈曲線

2.3 高過零率比率(High Zero Crossing Rate Ratio)

過零率在特徵化不同類型的音訊上被證明是非常有用的，他被使用在很多先前的語音與音

樂的分類演算法上。在我們的實驗中，我們發現過零率的的變化比原先的過零率的值更有辨識性，所以我們利用高過零率比率(HZCRR)當成演算法中的特徵值，如下定義：

$$HZCRR = \frac{1}{2}\sum_{n=0}^{N-1}\left[\mathrm{sgn}\left(ZCR(n) - 1.5avZCR\right) + 1\right] \qquad (2.3)$$

其中 n 代表音框索引，N 代表一秒內的音框總數，sgn[.]是符號函式，以及 ZCR(n)代表在第 n 個音框的過零率。通常語音訊號是由交替的 voice 聲音與 unvoice 聲音所組成，另一方面，音樂並沒有這種組成結構。因此，對於語音而言，它的過零率變化將會大於音樂。

下圖 2.7 代表 HZCRR 的機率分佈曲線：(a)代表語音 (b)代表音樂

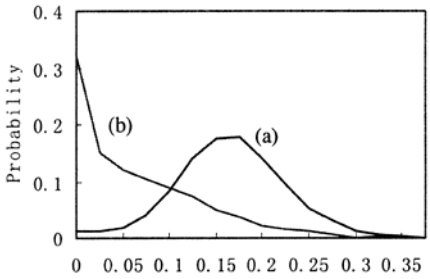


圖 2.2 HZCRR 的機率分佈曲線

2.4 頻譜流量(Spectrum Flux)

頻譜流量被定義成一秒內相鄰兩個音框的平均變化率的值，公式如下：

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}\left[\log\left(A(n,k)+\delta\right) - \log\left(A(n-1,k)+\delta\right)\right]^2 \qquad (2.4)$$

其中 A(n,k)是輸入信號第 n 個音框的離散傅利葉轉換(Discrete Fourier Transform)：

$$A(n,k) = \left|\sum_{m=-\infty}^{\infty} x(m)w(nL - me^{-j(2\pi/L)km})\right| \qquad (2.5)$$

x(m)是原始輸入的訊號，w(m)是窗函式，L 代表窗的長度，K 是 DFT 的階數，N 則是音框的總數，以及 δ 爲一個極小的數值避免計算時的溢位。

在我們實驗中，我們發現通常場外記者的 SF 值高於廣告，因爲場外的主要成分是爲語音或是環境聲，而廣告大部分都是由音樂組成。

下圖 2.3 代表 spectrum flux 特徵值的曲線：0~200 代表語音，200~350 代表音樂，350~450 代表環境聲音。

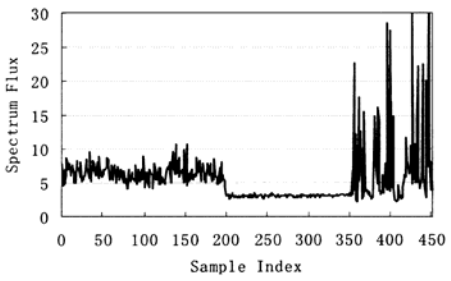


圖 2.3 spectrum flux 特徵值的曲線

2.5 梅爾頻率頻譜(Mel Frequency Spectrum)

在頻譜圖(spectragram)中，其實就可以明顯的看出語音與音樂的不同，但是若是直接取 FFT 完後的 512 或 1024 的值又太多，而且我們觀察頻譜時也不是全部觀察，而是去看它的密度比較深的地方。所以我們希望在取 MFCC 時，不要作最後的離散餘弦轉換，而是在做完 FFT 經過梅爾濾波組後的 26 個值當作一個特徵向量，稱之爲 MFS。

3.場景切割與分類

我們的系統架構同如圖3.1，首先我們會訓練出四種場景，輸入爲一小時的測試新聞，輸出爲四種場景的切割與分類。第一階段爲場景轉換點的偵測，第二階段爲場景的分類，我們會再接下來詳細介紹。

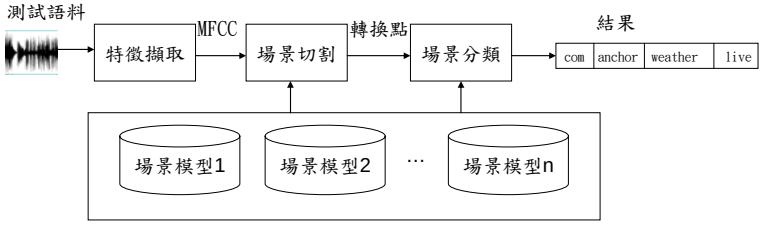


圖 3.1 系統架構圖

3.1 場景轉換點的偵測

在公共電視新聞語料中，我們可以發現研究中要切割與分類的場景，可以由新聞主播報導的場景將所有的場景類型切割出來。如圖3.2所示，我們將新聞主播報導場景的開始時間點與結束時間點找出來的話，這樣就可以順利的找到所有場景轉換點。

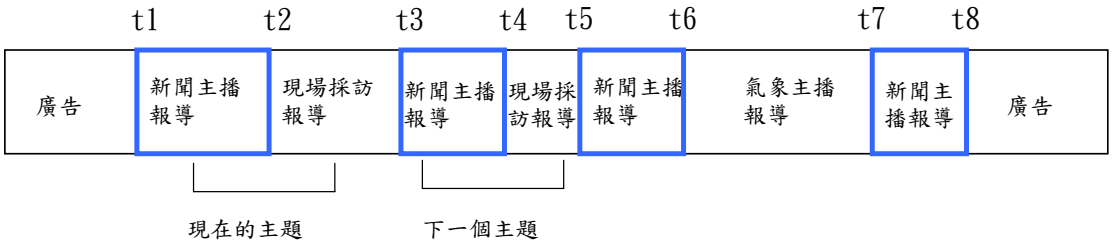


圖3.2 場景轉換時間點

t1代表第一個新聞主播報導的開始時間點，同時也是廣告場景轉換到新聞主播報導場景的轉換時間點，t2代表第一個新聞主播報導的結束時間點，同時也是新聞主播報導場景轉換到現場採訪報導場景的轉換時間點，t3、t4…以此類推。新聞主播報導場景的主要是由單一主播的語音構成，而攝影棚內的背景環境很安靜，並沒有背景環境聲音干擾。在許多語者辨識的系統中都是利用梅爾倒頻譜參數當成重要的特徵，將不同的語者區分出來，所以分類新聞主播報導方面，很適合用梅爾倒頻譜參數來描述主播的口腔組成，進而達到分類的效果。而語音訊號中富含了許多重要的因素讓我們來辨識新聞主播報導。

3.1.1 模型訓練

在研究中，如圖 3.3 所示，首先，從公共電視新聞語料中隨機選出三小時的測試語料，並把這三小時的語料，分出新聞主播報導、現場採訪報導、氣象主播報導、廣告等四種場景類型，

再對每一種場景透過特徵擷取子系統(Feature Extraction Subsystem)後，以特徵向量的形式儲存 39 維梅爾倒頻譜參數，之後利用模型訓練子系統(Model Training Subsystem)後，以模型高斯混合模型參數的型態儲存下來，訓練出四種不同的模型參數。

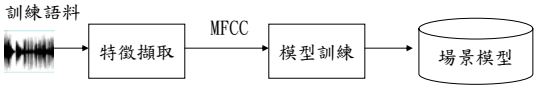


圖 3.3 場景模型訓練流程圖

3.1.2 模型測試

接著隨機選取出一小時的測試語料，經過特徵擷取子系統後，儲存 39 維的梅爾倒頻譜參數向量，利用模型測試子系統(Model Testing Subsystem)來對之前訓練出的模型參數找出最大事後機率（Maximum A Posteriori，MAP）的高斯混合模型，以辨識出場景的種類，如圖 3.4 所示：

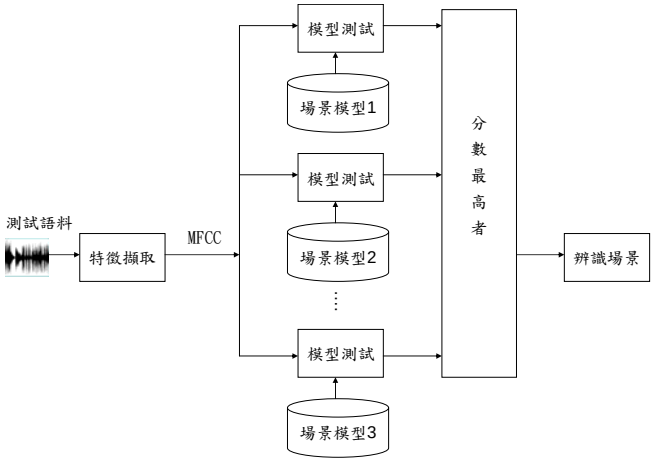


圖 3.4 模型測試流程圖

測試過程中，我們並不是拿整整一小時的新聞所擷取出的 MFCC 去跟每一個模型作比對，而是一段段的聲音片段去觀察與辨識場景，其中觀察的聲音片段爲 3 秒，並移動 1 秒計算一次 MAP，如圖 3.5 所示：

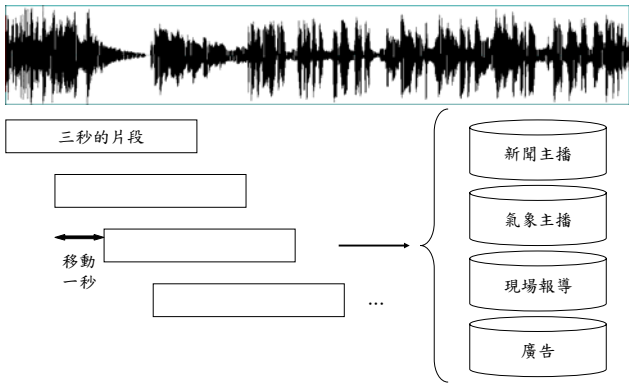


圖 3.5 一小時新聞測試的示意圖

換句話說，就是將測試的語音片段(3 秒)分別與新聞主播報導、氣象主播報導、現場採訪報導及廣告場景的高斯混合模型參數計算出可能的機率，之後選出機率最大者，然後我們判定此測

試語音的片段屬於此參數模型所對應的場景。由於我們是每移動一秒去計算是屬於哪個場景，所以當下一秒計算出的場景與現在不同時，我們便標示此時間點爲一個場景的轉換時間點。

3.1.3 後處理分析

另一方面，在每秒計算出的場景中，偶爾會出現 1 到 7 秒的錯誤判斷，我們稱之爲一個 error。對於此現象，我們利用 Median Filter 將其同化成相鄰的場景，根據實際聽取場景維持的秒數，平均一個場景的片段在 8 秒以上，所以我們針對 1 到 7 秒的 error 進行同化，以提升場景轉換時間點的正確率，另一方面透過實驗，我們也發現當同化到 6 秒以上正確率並不會再提升，反而有時還會下降，是因爲會同化到正確的轉換時間點，因此我們最多同化到 5 秒。

由以上 3 個步驟的步驟我們可以找出場景間轉換的時間點。

3.2 場景分類

在 3.1 節中，我們找出了所有場景轉換時間點，透過這些轉換時間點，便可以將一小時的新聞語料切割出一段段的場景片段，但是 MFCC 並不能正確的描述廣告的特性，因爲廣告與現場採訪報導最大的不同在於廣告含有音樂成分，因此在這部分，我們將非主播部分去擷取它的 LSTER、HZCRR、SF 與 MFS。將這一段段的非主播場景片段當作測試語料，並透過 GMM 分類器將這一整段的場景去作分類，並計算出是屬於現場採訪報導還是廣告場景。

3.3 加速策略

若是利用一秒一秒的去判斷屬於哪個場景，這樣一小時的新聞總共需要判斷 3600~3800 次，這樣所需的計算時間大致上要 10 分鐘，所以這部分我們開發了一套新的加速策略，希望可以將判斷切割與分類場景的計算時間縮短。發現當轉換點發生時，大致上都會有一段 0.2 到 1.0 秒的 silence，而在這 silence 的左右兩部分的特徵值分佈也不盡相同，如圖所示：

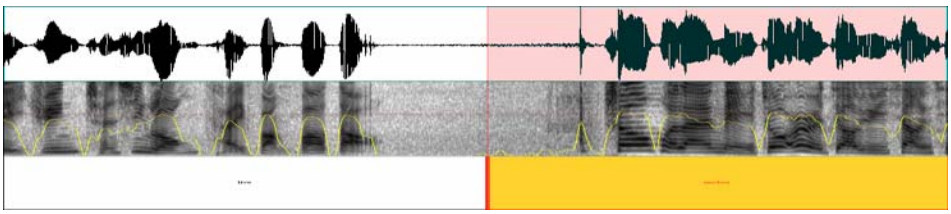


圖 3.6 現場採訪報導<->新聞主播報導

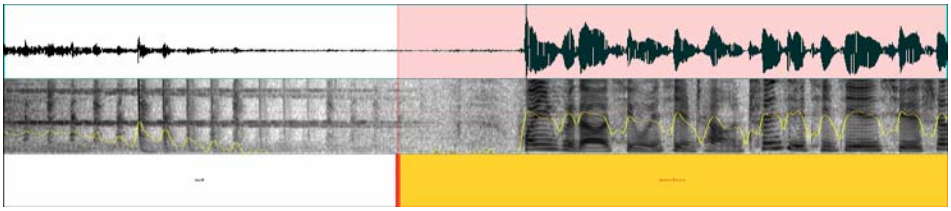


圖 3.7 廣告<->新聞主播報導

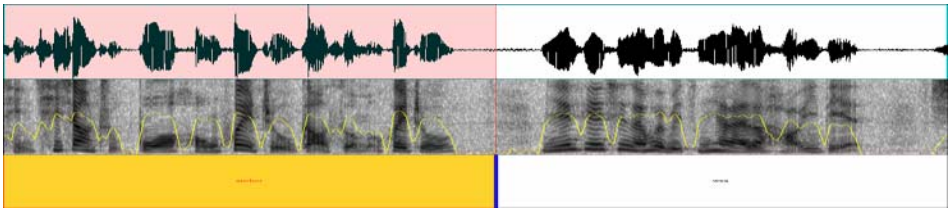


圖 3.8 新聞主播報導<->氣象主播報導

因此我們設計了一套快速判斷場景演算法去判斷這四大場景的轉換點，計算出的暫時轉換點(temp change point)大約有 600 到 700 個，而真實轉換點是這些暫時轉換點的子集合，相較於之前方法的 3600 個到 3800 個，我們大概可以省下大約 1/2 的時間，下面爲快速策略的演算法：

Step1：計算一小時新聞的能量

Step2：if(能量維持一段 0.1 秒長的 silence)

此時取這一段 silence 開始與結尾的 1/2 的時間點當成一個暫時轉換點。

Step3：將暫時轉換點的左右各 3 秒，總共 6 秒的聲音送進場景辨識器辨識。若左右兩段的場景不同，則此暫時轉換點爲真實轉換點。若左右兩段場景相同，則刪除此暫時轉換點。

Step4：合併出兩個真實轉換點間的場景類型。

圖 3.9 中，綠色曲線代表 step1 計算出的能量，紅色箭頭代表此時的靜音長在 0.1 秒以上，此時就設定藍色直線爲暫時轉換點，若暫時轉換點的左右場景不同則設爲真實轉換點，相同則刪除。

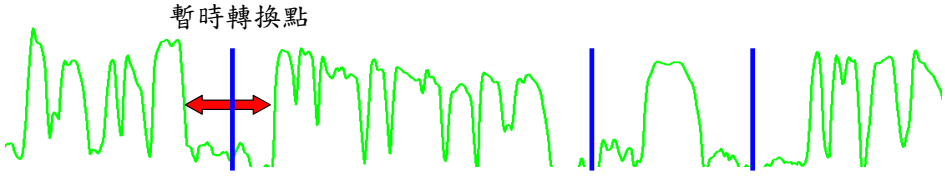


圖 3.9 快速策略

接下來，我們再利用另一個圖來解釋演算法，圖 3.10 代表一小時的新聞，其中 t1，t2，t3…代表在 Step2 之後計算出的暫時轉換點。在 t1 的左邊爲廣告，右邊爲新聞主播報導，所以我們變標示此點爲真實轉換點(實線)。另一方面，由於在 t3 的左邊爲現場採訪報導，同時右邊也是現場採訪報導，因此我們會刪除此暫時轉換點(虛線)，以下以此類推。因此 t1、t2、t4、t5 與 t7 爲真實轉換點。同時我們也將不同的場景類型分類了出來。

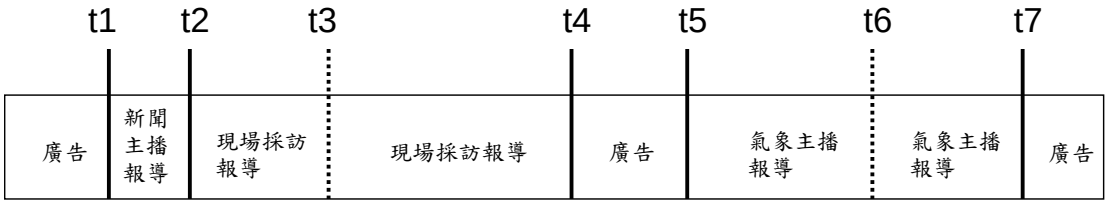


圖 3.10 快速策略

4. 實驗設計與實驗結果

4.1 公共電視新聞語料簡介

本論文所使用的語音資料庫爲公共電視新聞語料庫(Public Television Service News Database, PTSND)，是由中研院王新民教授以及助理研究團隊所整理規劃的中文電視新聞語料，收集了西元2001~2003年共220小時的新聞wave檔；其錄音的參數爲44.1kHz的取樣率，16-bit的解析度，而每段節目長約60分鐘，由數位錄音機(DAT recorder)直接由公視新聞的主控台所錄製而成，因考量檔案傳輸及讀取速度的問題，所以每個DAT都經由人爲處理成16kHz 16-bit單聲道的WAV檔。接下來我們簡述一下PTSND語料庫的一些統計特性，如表4.1所示；首先若我們以語者類別來區分的話，因爲外場記者及受訪者有相似的背景聲音，所以我們把兩者合併爲一類，稱之爲現場採訪報導，而氣象主播因爲其背景大多爲音樂，因此獨立出來統計；此外，新聞主播報導無背景聲音，故自成一類。

表4.1 PTSND 基本統計特性

Scene types	Percentage(in time)
新聞主播報導	17.68%
氣象主播報導	15.12%
現場採訪報導	59.20%
廣告	8.00%

4.2 系統效能評估

實驗的效能評估是利用插入率（insertion rate）以及刪除率（deletion rate）來評估我們的方法。

如圖 4.1 所示，Reference Boundaries 是指經由人工標示出的正確場景轉換點，Testing Boundaries 是指電腦經過我們設計出的策略後，計算出的場景轉換點。Insertion 代表電腦有計算出來的轉換點，但是人工並沒有標示此轉換點。Deletion 則是相反情況，及人工有標示此轉換點，但是電腦沒計算出此點。Matching 是指電腦計算出的轉換點跟人工標示的正確轉換點差距在 2 秒內。

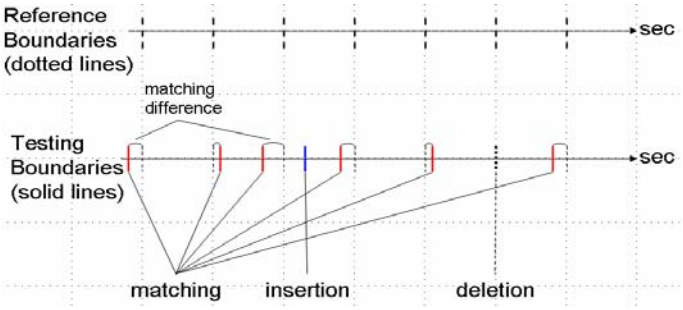


圖 4.1 效能評估

而它們之間的關係如下式所示：

$$N_{matching} + N_{insertion} = N_{Testing} \quad ,$$

$$N_{matching} + N_{deletion} = N_{\text{Re}\,ference}\ ,$$

其中$N_{matching}$ 代表Matching轉換點的數量， $N_{insertion}$ 代表Insertion轉換點的數量， $N_{deletion}$ 代表Deletion轉換點的數量。 $N_{\text{Re}\,ference}$ 和 $N_{Testing}$ 各代表人工標示與電腦計算的轉換點數量。接著我們定義了插入率(insertion rate)以及刪除率(deletion rate)的公式：

$$Insertion\_Rate = \frac{N_{insertion}}{N_{Testing}} \times 100\%$$

$$Deletion\_Rate = \frac{N_{insertion}}{N_{\text{Re}\,ference}} \times 100\%$$

4.3 實驗參數設定與結果

4.3.1 不同特徵向量維度對系統效能的影響

首先我們先評估不同特徵向量維度對轉換時間點辨識率的影響，如表所示：我們可以發現當特徵向量維度提高時，Insertion Rate 與 Deletion Rate 都會下降，這代表越高的特徵向量維度去描述場景的機率分佈與特性會越好。當特徵向量維度增加時，計算量會大幅升高，而導致耗費計算很多的時間，所以兩者是一種 trade off。最後，我們也發現當 mixture 增加時，Insertion Rate 與 Deletion Rate 會漸漸降低，但是 mixture 到了 64 有些不降反升，這是由於特徵值的分佈用了太多高斯分佈去描述了。

表 4.3 不同特徵向量維度的效能

mixture	condition / Evaluation	MFCC(13 dims)	MFCC+delta (26 dims)	MFCC+delta+ Deltadelta (39 dims)
4	Insertion Rate(%)	39.58	23.81	25
	Deletion Rate(%)	19.44	11.11	16.67
8	Insertion Rate(%)	25	20	20
	Deletion Rate(%)	16.67	11.11	11.11
16	Insertion Rate(%)	27.27	13.16	8.33
	Deletion Rate(%)	11.11	8.33	8.33
32	Insertion Rate(%)	26.19	15	8.33
	Deletion Rate(%)	13.89	5.56	8.33
64	Insertion Rate(%)	23.80	17.5	5.56
	Deletion Rate(%)	11.11	8.33	5.56

4.3.2 不同觀察片段時間長對系統效能的影響

接下來我們評估觀察片段時間長對轉換時間點辨識率的影響，如表所示，當觀察片段增加到 4 秒的時候，使由於可能包含到下一場景的特徵，導致影響這個時候的判斷。而觀察片段爲 2 秒時，判別場景的特徵不夠充分，所以 Insertion Rate 與 Deletion Rate 也會上升一些。因此我們設定觀察片段時間長 3 秒爲我們研究中的評估。

表 4.4 不同觀察片段時間長的效能比較

mixture	condition Evaluation	2秒	3秒	4秒
4	Insertion Rate(%)	23.68	25	22.5
	Deletion Rate(%)	19.44	16.67	13.89
8	Insertion Rate(%)	21.05	20	23.8
	Deletion Rate(%)	16.67	11.11	11.11
16	Insertion Rate(%)	11.11	8.33	18.42
	Deletion Rate(%)	11.11	8.33	13.88
32	Insertion Rate(%)	15.78	8.33	13.16
	Deletion Rate(%)	11.11	8.33	8.33
64	Insertion Rate(%)	8.33	5.56	8.33
	Deletion Rate(%)	8.33	5.56	8.33

4.3.4 每秒移動策略與快速策略速度比較

在研究中，主要影響計算時間的地方在模型測試子系統，因此我們針對兩種不同的策略，將它們的模型測試的時間作比較。如表 4.6，我們發現快速策略比每秒移動策略快了大約兩倍時間，主要是因爲每秒移動策略總共要測試模型 3600~3800 次，但是經由快速策略後的測試模型只要 600~700 次，但是快速策略要測試左右 3 秒的聲音各一次，判斷是屬於哪種場景，所以大致上也要測到 1200~1400 次。所以，快速策略比每秒移動的速度省了 1/2 以上。

表 4.6 每秒移動與快速策略的比較

method feature	每秒移動	快速策略
MFCC	153 sec	79 sec
MFCC+delta	230 sec	120 sec
MFCC+ delta+ delta_delta	294 sec	163 sec

4.3.5 每秒移動與快速策略的效能比較

我們發現快速策略的 Insertion Rate 與 Deletion Rate 比每秒移動的策略低。快速策略幾乎所有的轉換點都會去判斷，但是若轉換點出現在不是 silence 並維持 0.1 秒以上的話，快速策略就無法找出來。另一方面，由於每秒移動策略是每移動一秒就去判斷場景，這樣對於偵測場景轉換來說太細微了，也就是說一點點的差異就容易被誤判爲錯誤的場景，而這部分的錯誤無法利用同化場景的方法更正。而快速策略是利用場景轉換間會存在一段 silence，這是一個關鍵，而這種方法可以偵測出幾乎全部的場景轉換點，也因此快速策略的效能表現都比每秒移動策略佳。

表 4.7 每秒移動與快速策略比較

feature	condition Evaluation	每秒移動	快速策略
MFCC	Insertion Rate(%)	27.27	8.33
	Deletion Rate(%)	16.67	8.33
MFCC+delta	Insertion Rate(%)	20	5.40
	Deletion Rate(%)	11.11	2.77
MFCC+ delta+ delta_delta	Insertion Rate(%)	8.33	5.40
	Deletion Rate(%)	8.33	2.77

4.3.6 維持靜音長對快速策略效能影響

實驗中，靜音維持 0.1 秒，可以找出 862 個暫時轉換點，靜音維持 0.2 秒時可以找出 598 個暫時轉換點，0.3 秒可以找出 335 個。其中維持 0.1 秒可以找出較多的暫時轉換點，所以真實轉換點不容易遺漏，但也因此要判斷更多次場景的轉換。而有幾個轉換點，其之間的轉換就是小於 0.2 秒，因此若設爲維持 0.2 秒以上的話，此類的暫時轉換點就無法找出。

表 4.8 維持靜音長對快速策略效能比較

mixture	condition Evaluation	0.1秒	0.2秒	0.3秒
MFCC	Insertion Rate(%)	8.33	13.51	6.45
	Deletion Rate(%)	8.33	11.11	19.4
MFCC+delta	Insertion Rate(%)	5.40	11.11	6.66
	Deletion Rate(%)	2.77	11.11	22.22
MFCC+ delta+ delta_delta	Insertion Rate(%)	5.40	11.11	9.67
	Deletion Rate(%)	2.77	11.11	22.22

4.3. 分類結果的評估

我們評估場景分類的正確率，分類正確率的定義如下：

$$\text{分類正確率} = \frac{\text{正確辨別的測試檔案數}}{\text{所有的測試檔案數}} \times 100\%$$

每一種測試類型都包含 20 個測試檔案，總共 80 個檔案。接著將每個測試檔案透過場景分類器去作分類，例如第一行，代表新聞主播共 20 個檔案被分類到新聞主播的比率爲 100%。另一方面，我們發現在廣告的部份還是容易出錯，是由於廣告的組成相當複雜，其中主要成分是語音的測試檔案很容易出錯，在這種檔案類型中，音樂的聲音幾乎都被語音覆蓋掉，因此會容易辨識成現場採訪報導，我們用人耳去聽的確不是很清楚的可以辨別出其差異性。另外一方面，就分類結果而

言，我們的平均正確率可達 92.5%。

表 4.9 分類結果評估

分類類型 測試類型	新聞主播報導	現場採訪報導	氣象主播報導	廣告
新聞主播報導	100%	0%	0%	0%
現場採訪報導	5%	90%	5%	0%
氣象主播報導	0%	0%	100%	0%
廣告	0%	20%	0%	80%

5. 結論與未來展望

首先我們要感謝中研院王新民教授，他們研究團隊所開發的公共電視新聞語料(PTSD)，提供我們有關新聞中很多詳細的資訊，例如語者資訊、場景的類型、新聞的內容、轉換的時間點以及如何錄製新聞語料的步驟等等，讓我們能快速的瞭解與進一步分析新聞語料的內容。我們也利用了這份語料當成我們效能評估的依據。

在本論文中，我們分析各種場景時域與頻域的特性，使用高斯混合模型來做場景的切割與分類的分類器。在場景切割方面，我們透過新聞主播報導的開始與結束時間點去決定場景的轉換點，使用每秒移動策略，其 Deletion Rate 爲 5.56%，Insertion Ration 爲 5.56%。在場景分類方面，我們使用了 MFCC、LER、HZCRR、SF 與 MFS 去將經過真實轉換點切割出的一段段聲音去作分類，可以達到 92.5%的平均正確率。由於上述的方法耗費計算的時間較久，因此我們也開發了一套快速策略，計算時間節省了大約 1/2，其 Deletion Rate 爲 2.27%，Insertion Ration 爲 5.4%。本論文與其他研究不同在於我們利用知識爲基礎的想法、當人類碰到此問題實是如何解決的，如何去判斷場景的不同與場景轉換時會發生那些現象，將這些解決的想法轉成知識，進一步將這知識設計一套演算法而把場景的切割與分類自動化，而傳統處理場景轉換點偵測是利用相異度的判斷，但是這種方法會容易受到環境的影響（如環境聲、音樂聲、噪音等等）而導致誤判，所以有較高的 Deletion Rate 與 Insertion Rate。最後，我們期望未來可以利用我們分類出來的四種場景去擷取出更多的音訊類型，並研究更多頻域與時域的有效特徵值，另一方面，結合視覺特徵及文字上的資訊，開發更快速正確的方法，將場景的自動切割與分類辨識率提升。

6. 參考文獻

[1] Hsin-min Wang, Shi-sian Cheng, and Yong-cheng Chen, “The SoVideo broadcast news retrieval system for Mandarin Chinese.” International Conference on Spoken Language Processing 2004

[2] Yih-Ru Wang, Chi-Han Huang, “Speaker-and-environment change detection in broadcast news using the common component GMM-based divergence measure.”, International Conference on Spoken Language Processing 2004, pp1069-1072.

[3] Jincheng Huang, Zhu Liu, Yao Wang, ”Joint scene classification and segmentation based on hidden

Markov model", Multimedia, IEEE Transactions on Volume 7, Issue 3, June 2005 Page(s):538 - 550
[4] Zhu Liu, Yao Wang, Tsuhan Chen, "Audio feature extraction and analysis for scene segmentation and classification." Journal of VLSI Signal Processing Systems 1998, Vol.20, pp61 – 79.
[5]Lie Lu, Hong-Jiang Zhang, Hao Jiang, "Content analysis for audio classification and segmentation." IEEE Trans. on Speech and Audio Processing 2002, Vol.10, No.7, pp.504-516.
[6] Lie Lu, Hong-Jiang Zhang, Stan Li, "Content-based audio classification and segmentation by using support vector machines."ACM Multimedia Systems Journal 2003, pp. 482-492.
[7] Lie Lu, Hao Jiang, Hong-Jiang Zhang, "A robust audio classification and segmentation method." ACM International Conference on Multimedia 2001, pp203-211.
[8]Tong Zhang, C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification." IEEE Transactions on Speech and Audio Processing 2001, Vol.9, No.4.
[9]Lekha Chaisorn and Tat-Seng Chua, "The Segmentation and Classification of Story Boundaries in News Video" , Proceeding of 6th IFIP working conference on Visual Database Systems VDB6 2002, Australia 2002
[10]TingYao Wu, Lie Lu, Hong-Jiang Zhan, "UBM-based real-time speaker segmentation for broadcasting news."Speech and Signal Processing (ICASSP) 2003, Vol. II, pp. 193-196.
[11]Ting-Yao Wu, Lie Lu, Ke Chen, Hong-Jiang Zhang, "Universal background models for real-time speaker change detection."Proc. of the 9th International Conference on Multi-Media Modeling 2003, pp.135-149.

Improving Translation of Unknown Proper Names Using a Hybrid Web-based Translation Extraction Method

Min-Shiang Shia Jiun-Hung Lin Scott Yu Wen-Hsiang Lu

Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan, R.O.C.
{foreverdream, jhlin, scottyu}@csie.ncku.edu.tw, whlu@mail.ncku.edu.tw

Abstract

Recently, we have proposed several effective Web-based term translation extraction methods exploring Web resources to deal with translation of Web query terms. However, many unknown proper names in Web queries are still difficult to be translated by using our previous Web-based term translation extraction methods. Therefore, in this paper we propose a new hybrid translation extraction method, which combines our pervious Web-based term translation extraction method and a new Web-based transliteration method in order to improve translation of unknown proper names. In addition, to efficiently construct a good quality transliteration model, we also present a mixed-syllable-mapping transliteration model and a Web-based semi-supervised learning algorithm to explore search-result pages further for collecting large amounts of English-Chinese transliteration pairs from the Web.

1 Introduction

In machine translation (MT) (Brown et al. 1993) or cross-language information retrieval (CLIR) (Jaleel and Larkey 2003; Pirkola et al. 2003), unknown term translation are still problematic and remain to be solved. Conventionally, most of the existing MT or CLIR systems rely mainly on general-purpose bilingual dictionaries, which usually lack translations of proper names or technical terms, and thus are unable to deal with such problems. We have proposed an effective Web-based approach to exploring abundant language-mixed texts on the Web like anchor texts and search-result pages for alleviating the difficulty of unknown query term translation (Lu et al. 2002, 2004; Cheng et al. 2004). However, the approach employing statistical techniques still suffers from the problem of data sparseness and indirect association errors in finding translations of low-frequency unknown terms (Melamed, 2000).

According to the report in previous research (Davis et al. 1998), around 50% of unknown terms are proper names. To improve translation of unknown proper names, in this paper, we propose a hybrid translation extraction method, which is composed of our pervious search-result-based term translation

extraction method (Section 3.2) and a new Web-based transliteration method (Section 3.3). Transliteration is the process that converting a sequence of substrings or characters in the source language (e.g., English) into a pronunciation-approximate substring/character sequence in the target language (e.g., Chinese). Many researchers have proposed phoneme-based mapping techniques for proper name transliteration (Jung et al. 2000; Knight & Graehl 1998; Lin & Chen 2002; Meng et al. 2001; Virga & Khudanpur 2003), but converting an English word from phonemic representation to Chinese Pinyin and from Pinyin to Chinese characters may cause double errors. Taking this problem into consideration, we thus try to adopt direct orthographical mapping for proper name transliteration and propose a simple mixed-syllable-mapping transliteration model which can effectively increase the correct mapping between an English-Chinese transliteration pair with different number of transliteration unit (syllable), such as "Ericsson" (易利信) with four English transliteration units "e", "ri", "c", "sson" and three Chinese transliteration units "易", "利", "信" (Section 3.3). Additionally, to train a good quality transliteration model which is used to filter out impossible transliteration candidates in the process of extracting translation of unknown proper names, we also present a Web-based semi-supervised learning algorithm to collect large amounts of English-Chinese transliteration pairs from the Web. Experimental results show that our new approach can make improvements for translation of unknown proper names.

2 Related Work

2.1 Parallel-Corpus-based Term Translation Extraction

Term translation extraction is a significant research topic in the field of machine translation. A number of related researches (Gale and Church 1991; Kupiec 1993; Melamed 2000; Smadja et al. 1996) have used sentence-aligned parallel corpora to extract translations since the advent of statistical translation model (Brown et al. 1990, 1993). For example, Melamed (2000) proposed statistical translation models to improve the techniques of word alignment by taking advantage of pre-existing knowledge and overcome the problems of indirect association errors, i.e., erroneous translational correspondence arose from highly co-occurred relevant terms. Although high accuracy of translation extraction can be easily achieved by these techniques, sufficiently large parallel corpora for various subject domains and language pairs currently are not always available.

2.2 Comparable-Corpus-based Term Translation Extraction

However, less attention has been devoted to automatic extraction of term translations from comparable or even unrelated texts, since such methods encountered more difficulties due to lacking parallel correlation aligned between documents or sentence pairs. Rapp (1999) proposed an approach to utilizing non-parallel corpora based on the assumption that the contexts of a term should be similar to the contexts of its translation in any language pairs. Fung et al. (1998) also proposed a similar approach that uses vector-space model and takes a bilingual lexicon (called seed words) as feature set to estimate the similarity between a word and its translation candidates. These works are important for automatic

extraction of new terminology and unknown proper names in diverse domains. It is a pity that comparable corpora are easier to obtain, however, how to achieve better performance for higher translation coverage is still a challenging task.

2.3 Web-based Term Translation Extraction

The Web is becoming the largest data repository in the world, which consists of huge amounts of multilingual and wide-scoped hypertext resources. A number of studies have been concentrated in the use of the Web to complement insufficient corpora (Cao & Li 2002; Kilgarriff et al. 2003). How to utilize the Web resources to benefit translations of unknown terms is worthy to investigate.

As mentioned above, the conventional term translation methods suffer from the problems of the lack of large-size parallel corpora and the shortage of translation coverage of comparable corpora in medical domain. Thus, we have proposed several Web-based methods to effectively deal with translation of frequent Web query terms by exploring Web anchor text and search-result pages. Although the anchor-text-based approach has been proven effective in extracting multilingual translations (Lu et al. 2002, 2004), it requires crawling the Web to gather sufficient training data as well as more network bandwidth and storage. For the reason to reduce such costs, this paper only adopts the search-result-based approach to extract translation candidates for term translation (describes in Section 3.2). However, many proper names are still difficult to be translated correctly using the search-result-based approach. Therefore, in this paper we intend to further explore search results to collect English–Chinese transliteration pairs, and build a good quality transliteration model which can be used to filtered out impossible translation candidates to improve translation of unknown proper names.

2.4 Proper Name Transliteration

For name transliteration between Latin-alphabet languages and some Asian languages with different writing forms, such as English and Chinese, researchers have proposed phoneme-based mapping techniques (Jung et al. 2000; Knight & Graehl 1998; Lin & Chen 2002; Meng et al. 2001; Virga & Khudanpur 2003). Knight and Graehl used an English-katakana dictionary, katakana-English phoneme mapping, and the CMU Speech Pronunciation Dictionary to deal with transliteration between English words and Katakana sequences. Lin et al. (2003) proposed a statistical transliteration model and apply the model to extract proper names and their transliterations in a parallel corpus with high average precision and recall rates. However, Li et al. (2004) have pointed out that the transliteration precision of the phoneme-based approaches could be limited by two main constraints. First, Latin-alphabet foreign names from different origins have different phonic rules (Pirkola et al. 2003), such as French and English. Second, converting English words to Chinese characters will need two steps: converting from phonemic representation to Chinese Pinyin and from Pinyin to Chinese characters. Two cascaded converting steps may cause double errors. Taking this problem into consideration, we try to adopt direct orthographical mapping for name transliteration (described in Section 3.3).

3 Extracting Translation of Unknown Proper Names

3.1 Problem and Challenge

Actually, search-result page is a good resource for extracting translation of frequent unknown query terms. However, a number of unknown proper names are still not extracted correctly due to the problems of data sparseness. Thus, our idea is to integrate name transliteration techniques into the process of extracting translation of proper names in order to filter impossible transliterated candidates for improving the performance of translation extraction. To deal with the problem, first we need to extract terms from the search-result pages as translation candidates, and then filter out impossible candidates based on the name transliteration model. In fact, it is challenging to build a good quality transliteration model while lacking sufficient transliteration pairs for training. We therefore propose a Web-based semi-supervised learning algorithm to collect large amounts of English-Chinese transliteration pairs from the Web (see Section 3.3).

3.2 Extracting Translation Candidates Using a Search-Result-based Translation Extraction Method

We have proposed an effective search-result-based method to explore language-mixed search-result pages and utilize co-occurrence relation and context information for extracting unknown query term translation. In this section, we will simply describe candidate selection methods using the search-result-based method. For more details, please refer to our previous work (Cheng et al. 2004).

(1) Chi-Square Test Method: On the basis of co-occurrence analysis, chi-square test (χ^2) is adopted to estimate semantic similarity between the source term *E* and the target candidate *C*. The similarity measure is defined as

$$S_{\chi^2}(E, \mathrm{C}) = \frac{N \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)}, \qquad (1)$$

where *a*, *b*, *c* and *d* are the numbers of pages retrieving from search engines by submitting Boolean queries: "*E* and *C*", "*E* and not *C*", "not *E* and *C*", and "not *E* and not *C*", respectively; *N* is the total number of pages, i.e., $N = a + b + c + d$.

(2) Context-Vector Analysis Method: Due to the nature of Chinese-English mixed texts often appearing in Chinese pages, the source term *E* and the target candidate *C* may share common contextual terms in the search-result pages (Fung & Yee 1998; Rapp 1999). The similarity between *E* and *C* will be computed based on their context feature vectors in the vector-space model. The conventional *tf-idf* weighting scheme is used and defined as

$$w_{t_i} = \frac{f(t_i, p)}{\max_j f(t_j, p)} \times \log(\frac{N}{n}), \qquad (2)$$

where $f(t_i, p)$ is the frequency of term t_i in search-result page *p*, *N* is the total number of Web pages, and *n* is the number of the pages containing t_i. Finally, we use the cosine measure to estimate the similarity as follows:

$$S_{CV}(E,\mathrm{C}) = \frac{\sum_{i=1}^{m} w_{e_i} \times w_{c_i}}{\sqrt{\sum_{i=1}^{m}(w_{e_i})^2 \times \sum_{i=1}^{m}(w_{c_i})^2}}. \qquad (3)$$

3.3 Filtering Translation Candidates Using a Web-Based Name Transliteration Method

(1) English Letter Substring Segmentation: Wan and Verspoor (1998) have developed a fully rule-based algorithm to transliterate English proper names into Chinese names. We simplify their syllabification techniques to generate a few simple heuristic rules of segmenting an English name into letter substrings. Each English substring is regarded as a transliteration unit (TU) in this paper and had at most one corresponding character of the Chinese transliterated name. Initially, we used only five rules listed below:

- a, e, i, o, u are vowels, and y is also regarded as a vowel if it appears behind a consonant. All other letters are consonants.
- Separate two consecutive vowels except the following cases: ai, au, ee, ea, ie, oa, oo, ou, etc.
- Separate two consecutive consonants except the following cases: bh, ch, gh, ph, th, wh, ck, cz, zh, zk, ng, sc,ll, tt, etc.
- l, m, n, r are combined with the left vowel only if they are not followed by a vowel.
- A consonant and a following vowel are regarded as a TU.

For example, “amaya” (阿馬雅) is segmented into three substrings “a”, ”ma”, “ya”, and “mobley” (莫布里) is segmented into three substrings “mo”, “b”, “ley”. Currently, some English names may be segmented incorrectly, but it is easy to manually add new rules for improving English letter substring segmentation.

(2) Mixed-Syllable-Mapping Transliteration Model: To avoid double errors of converting English phonemic representation to Chinese Pinyin and from Pinyin to Chinese characters, we thus adopted direct orthographical mapping to deal with the alignment between any English name, $E = e_1e_2\ldots e_m$, and its Chinese transliterated name, $C = c_1c_2\ldots c_n$. Each English TU e_i is mapped to a Chinese character c_i with the probability $P(c_i \mid e_i)$. Initially, to efficiently train a Web-based transliteration model based on the collected transliteration pairs from the Web for filtering out impossible transliteration candidates, we adopt a simple name transliteration model called **forward-syllable-mapping transliteration model**, which computes the forward syllable mapping score between E and C using the following formula:

$$S_{FSM}(E,C) = P(C \mid E) \approx \prod_{i=1}^{\min(m,n)} [(1-\alpha)P(c_i \mid e_i) + \alpha], \qquad (4)$$

where α is the smoothing weight.

For an English-Chinese transliteration pair with different number of transliteration unit, such as “Rusedski” (魯塞斯基) with the five English segmented substrings “ru”, ”se”, “d”, “s”, “ki” and four Chinese characters “魯”, “塞”, “斯”, “基”, to increase the correct mapping between English TUs and Chinese characters, we propose an alternative transliteration mapping model called

reverse-syllable-mapping transliteration model, which is used to compute the reverse syllable mapping score as follows:

$$S_{RSM}(E,C) \approx \begin{cases} \prod_{i=m}^{m-n+1} [(1-\alpha)P(c_{i-(m-n)} \mid e_i) + \alpha], & m \geq n; \\ \prod_{i=n}^{n-m+1} [(1-\alpha)P(c_i \mid e_{i-(n-m)}) + \alpha], & m < n. \end{cases} \quad (5)$$

To cover all possibly correct mapping between English TUs and Chinese transliterated characters for the distinct types of English-Chinese transliteration pairs with the same or different transliteration units, we propose a simple **mixed-syllable-mapping transliteration model,** which combine the forward-syllable-mapping and reverse-syllable-mapping transliteration models, to estimate the mapping score as follows:

$$S_{MSM}(E,C) = \sqrt{S_{FSM}(E,C) \times S_{RSM}(E,C)}. \quad (6)$$

(3) Web-based Semi-Supervised Learning Algorithm: We intend to take advantages of abundant language-mixed texts on the Web to collect English-Chinese transliteration pairs and then train a good quality transliteration model. Thus, we design a semi-supervised learning process of transliteration mapping. The process is composed of three main stages: extraction of Chinese transliterated names, extraction of English original names, and learning of transliteration mapping, and described below as well as the algorithm in Figure 1.

- **Extraction of Chinese Transliterated Names:** Xiao et al. (2002) have proposed a bootstrapping algorithm that uses only five frequent Chinese transliterated characters as initial seed character set: {阿, 爾, 巴, 斯, 基} to automatically collect over 100,000 of Chinese transliterated names by utilizing search-result pages. Inspired by Xiao et al., we design a different bootstrapping algorithm which uses the same seed character set to automatically find large amounts of Chinese transliterated names from search-result pages. Initially, we select two frequent Chinese transliterated characters from the seed character set, and then send them to search engines for getting search-results pages. To efficiently extract more Chinese transliterated names from the search-result pages, we use the CKIP tagger (Ma & Chen 2003), which is a representative Chinese POS tagger with the ability of segmenting Chinese texts into meaningful words and extracting unknown words.
- **Extraction of English Original Names:** We first use the search-result-based translation extraction method (Section 3.2) to find possible candidates of English original names, and then filter out the impossible candidates which are included in general-purpose bilingual dictionaries. Finally, to collect English-Chinese transliteration name pairs with high quality, we may need to take some manual efforts to examine the correct transliteration pairs.

Web-based Semi-Supervised Learning Algorithm for Collecting English-Chinese Transliteration Pairs and Training a Transliteration Model

Input: Chinese seed character set C_s and a general-purpose bilingual dictionary D
Output: English-Chinese transliteration pair set V_{ec}, and a transliteration model T

1. **Extraction of Chinese transliterated names:**
   - **1.1. Seed character selection:** select two frequent characters from the Chinese seed character set C_s.
   - **1.2. Search-result crawling:** send the two selected characters to a search engine and get search-result pages.
   - **1.3. Chinese transliterated name identification:** use CKIP tagger to find unknown terms in the search-result pages, and then take the unknown terms containing the two Chinese seed characters as potential Chinese transliterated names and add them into V_c .
   - **1.4. Seed character set updating:** update C_s by adding the new characters from the new Chinese transliterated names.
   - **1.5. Repeat step1** until the desired number of the Chinese transliterated name in the V_c is reached.
2. **Extraction of English original names:** for each potential Chinese transliterated name in V_c, perform the following sub-steps:
   - **2.1. Potential English name extraction:** use search-result-based translation extraction method (Section 3.2) to find potential candidates of English name.
   - **2.2. Candidate filtering:** filter out impossible English name candidates included in D.
   - **2.3. English name identification:** take some manual efforts to examine the correct original English names.
   - **2.4. English-Chinese transliteration pair updating:** update V_{ec} by adding the new transliteration pair.
3. **Learning of English-Chinese transliteration mapping:** use the proposed mixed-syllable-mapping transliteration model (equation (6)) to train a Web-based transliteration model T based on the extracted English-Chinese transliteration pairs.

Figure 1. Algorithm for collecting transliteration pairs and training a transliteration model.

- **Learning of Transliteration Mapping:** On the basis of the English letter substring segmentation rules and the proposed mixed-syllable-mapping transliteration model described above, we will train a Web-based transliteration model based on the collected transliteration pairs from the Web.

3.4 The Proposed Approach to Translation Extraction

Currently, for some unknown proper names, it is still difficult to effectively extract translation by using our previous search-result-based translation extraction method. Therefore, we try to combine a new Web-based transliteration method to enhance our previous search-result-based translation extraction method.

(1) Linear Combination Method: Intuitively, a simple method is to directly combine the above three different methods: the chi-square test method, the context-vector analysis method, and the Web-based transliteration method. Under consideration of the large difference of ranges of similarity values among the above methods, we would use a linear combination of inverse ranks to compute the similarity measure as follows:

$$S_{Combined}(E,C) = \sum_{m} \frac{\alpha_m}{R_m(E,C)}, \qquad (7)$$

where α_m is an assigned weight for each similarity measure S_m, and $R_m(E, C)$ represents the similarity rank of each target candidate C with respect to its source term E and is assigned to be from 1 to k (candidate number) according to similarity measure $S_m(E, C)$ in decreasing order.

Note that this liner combination method is only used as baseline in comparison with our proposed hybrid translation extraction method described below in the following experiments (Section 4.2).

(2) Hybrid Method: For some unknown proper names, the simple linear combination method might not make good improvements while these respective methods can't obtain high ranks for possibly correct transliteration candidates. Therefore, we propose a new hybrid translation extraction method in order to obtain better performance. First, we use the search-result-based translation extraction method described above to extract k (k = 20) terms with high similarity score as transliteration candidates. Second, some impossible candidates included in general-purpose bilingual dictionaries are filtered out, and then each of the rest transliterated candidates is ranked according to transliteration mapping score with the test proper name which is computed based on the Web-based transliteration model (Equation (4) and (6)).

4 Experimental Results and Analysis

We conducted the following experiments to examine the performance of the proposed hybrid translation extraction method and the comparison with the simple linear combination method. Particularly, the focus of the experiments is mainly emphasized on the effectiveness of translations of unknown proper names using the proposed mixed-syllable-mapping transliteration model and hybrid translation extraction method.

Collected data: Initially, our proposed Web-based semi-supervised learning algorithm is employed to efficiently collect about 11,000 English-Chinese transliteration pairs for training a transliteration model.

Test set: We constructed one test set of unknown English query terms, **NTCIR proper name set**, which contains 22 unknown transliteration names from a total of 100 NTCIR2 and NTCIR3 title queries that contain 175 and 183 unique query terms respectively (Chen & Chen 2001).

Evaluation Metric: The average top-*n* inclusion rate was adopted as a metric on the extraction of translation equivalents. For a set of terms to be translated, its top-*n* inclusion rate was defined as the percentage of the terms whose translations could be found in the first *n* extracted translations (Cheng et al. 2004).

Table 1. Comparison of translation results between the forward-syllable-mapping model and the mixed-syllable-mapping model.

Translation Method	Forward-Syllable-Mapping Transliteration Model			Mixed-Syllable-Mapping Transliteration Model		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Name Transliteration	14%	27%	27%	27%	32%	32%
Hybrid	36%	41%	45%	45%	50%	55%

Table 2. Comparison of translation results between the different translation methods.

Translation Method	Mixed-Syllable-Mapping Transliteration Model		
	Top-1	Top-3	Top-5
Search-Result-based	36%	36%	45%
Name Transliteration	27%	32%	32%
Linear Combination	32%	41%	55%
Hybrid	45%	50%	55%

Table 3. Effective results of translation extraction using the hybrid translation extraction method (underlined terms indicate correct translation).

Test Query	Translation Method	Top 5 Translation Candidates	
		Forward-Syllable-Mapping Transliteration Model	Mixed-Syllable-Mapping Transliteration Model
Michael	Search-Result-based	麥可布雷,麥克傑克森,施文彬,華納,個人	麥可布雷,麥克傑克森,施文彬,華納,個人
	Name Transliteration	蜜可艾爾,麥可艾爾,蜜可艾德,蜜可埃爾,蜜高艾爾	蜜可艾爾,密可艾爾,麥可艾爾,蜜可埃爾,蜜麥艾爾
	Linear Combination	麥可布雷,麥克傑克森,蜜可艾爾,施文彬,華納	麥可布雷,麥克傑克森,蜜可艾爾,施文彬,華納
	Hybrid	麥可傑克森,麥可布雷,麥克傑克森,麥克,喬丹	麥可布雷,麥可傑克森,麥克,麥克傑克森,舒馬克
Kosovar	Search-Result-based	譯音無限次，發行公司，散財，科索沃，譯音無限	譯音無限次，發行公司,散財,科索沃，譯音無限
	Name Transliteration	可索瓦,克索瓦,茉索瓦,可蘇瓦,可喬瓦	可索雷,克索雷,可索瓦,克索瓦,科索雷
	Linear Combination	譯音無限次，發行公司，可索瓦，散財，科索沃	譯音無限次，發行公司,可索雷，散財，科索沃
	Hybrid	科索沃，譯音無限次，發行公司，散財，譯音無限	科索沃,譯音無限次，發行公司，散財，譯音無限

4.1 Mixed-Syllable-Mapping Transliteration Model vs. Forward-Syllable-Mapping Transliteration Model

To test the effectiveness of the mixed-syllable-mapping transliteration model, we carried out a comparative experiment with different ranking. The results are shown in Table 1. Actually, the mixed-syllable-mapping transliteration model is effective to improve the top-*n* inclusion rate. For translation extraction of the NTCIR proper names, the mixed-syllable-mapping transliteration model can achieve 27% and 45% top-1 inclusion rates for the name transliteration method and the hybrid

translation method, respectively. Obviously, the reason is that for many English-Chinese transliteration pairs with different number of TU, reverse-syllable-mapping transliteration model can aid in learning correct mapping between English substrings and Chinese characters. Additionally, the model has the same assist effect to many partially matching transliteration pairs collected by using our proposed Web-based transliteration method. For the given proper name "Michael" (麥克) shown in Table 3, the better rank of its correct translation can be obtained by using the mixed-syllable-mapping transliteration model.

4.2 Hybrid Translation Extraction Method vs. Linear Combination Method

To determine the effectiveness of the proposed hybrid translation extraction method compared with other methods, we also did several comparative experiments with different ranking. The results are also shown in Table 2. For the NTCIR test set, surprisingly, the hybrid translation extraction method made a great improvement compared with the search-result-based translation extraction method, name transliteration method, or linear combination method. The hybrid translation extraction method with mixed-syllable-mapping transliteration model can achieve 45% top-1 inclusion rate. The main reason is that most of the incorrect translation candidates extracted by using the search-result-based translation extraction method can be filtered out by using the Web-based transliteration method. For example, given the proper name "Kosovar" (see Table 3), the correct Chinese transliterated name "科索沃" can be ranked to the top one from the fourth rank using only the search-result-based translation extraction method. However, the simple linear combination method seems not effective to improve translation performance since the name transliteration method is still limited in generating correct transliterated candidates even though it can generate many pronunciation-proximate candidates.

4.3 Discussions

Our proposed mixed-syllable-mapping model and hybrid translation extraction method is effective to improve performance in extracting translation of unknown proper names. However, the hybrid translation extraction method sometimes performs not good as linear combination method. An example such as "Viagra" (威而剛) is shown in Table 4. Currently, our Web-based semi-supervised learning algorithm is limited by insufficient transliteration training from our collected transliteration pairs which are still in the need of examining by large amounts of manual labor. In the future, we will develop an unsupervised learning algorithm to automatically collect much more amounts of English-Chinese

Table 4. Ineffective results of translation extraction using the hybrid translation extraction method (underlined terms indicate correct translation).

Test Query	Translation Method	Top 5 Translation Candidates Mixed-Syllable-Mapping Transliteration Model
Viagra	Search-Result-based	偉哥,食品藥物,<u>威而剛</u>,藥物管理局,藥物
	Name Transliteration	薇阿格拉,薇亞格拉,薇艾格拉,薇阿葛拉,薇亞葛拉
	Linear Combination	偉哥,食品藥物,薇阿格拉,<u>威而剛</u>,藥物管理局
	Hybrid	萬艾可,藥物管理,食品管理,輝瑞,<u>威而剛</u>

transliteration pairs from the Web for training good quality transliteration model. Besides them, there are still a number of cases that are still difficult to deal with by using the simple mixed-syllable-mapping transliteration model and need to be further investigated in the future.

5 Conclusions

We have presented a new hybrid translation extraction method that works well for improving extraction of translation of known proper names by effectively combining a previous search-result-based translation extraction method and our proposed Web-based name transliteration method. Additionally, our proposed simple mixed-syllable-mapping transliteration model and Web-based semi-supervised learning algorithm are also effective to collect English-Chinese transliteration pairs and then train a transliteration model for filtering out incorrect transliteration candidates in the process of extracting proper name translation.

References

N. A. Jaleel and L. S. Larkey. 2003. Statistical transliteration for English-Arabic cross language information retrieval. *CIKM* 2003: 139-146.

P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79-85.

P. F. Brown, , S. A. D. Pietra, V. D. J. Pietra and R. L. Mercer. 1993. The Mathematics of Machine Translation. *Computational Linguistics*, 19(2): 263-312.

Y.-B. Cao and H. Li. 2002. Base noun phrase translation using Web data and the EM algorithm. In *Proc. of COLING* 2002: 127-133.

K.-H. Chen and H.-H. Chen. 2001. The Chinese Text Retrieval Tasks of NTCIR Workshop 2. In *Proc. of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*.

P.-J. Cheng, Y.-C. Pan, W.-H. Lu, L.-F. Chien. 2004. Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora. In *Proc. of ACL* 2004: 535-542.

M. W. Davis and W. C. Ogden. 1998. Free Resources and Advanced Alignment for Cross-Language Text Retrieval. In *Proc. of the Sixth Text Retrieval Conference* (*TREC*6): 385-394.

P. Fung and L.-Y. Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of ACL* 1998: 414-420.

W. A. Gale and K. W. Church. 1991. Identifying Word Correspondances in Parallel Texts, In Proc. of DARPA Speech and Natural Language Workshop.

W. Gao, K.-F. Wong and W. Lam. 2004. Phoneme-based Transliteration of Foreign Names for OOV Problem. *In Proc. of IJCNLP* 2004: 274-381.

J. Halpern. 2000. Lexicon-based orthographic disambiguation in CJK intelligent information retrieval. In *Proc. of Workshop on Asian Language Resources and International Standardization.*

S. Y. Jung, S. L. Hong and E. Paek. 2000. An English to Korean Transliteration Model of Extended Markov Window. In *Proc. of COLING* 2000.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29(3): 333-348.

K. Knight and J. Graehl. 1998. Machine Transliteration, *Computational Linguistics* 24(4): 599-612.

J. M. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proc. of ACL* 1993: 17-22.

H. Li, M. Zhang and J. Su. 2004. A Joint Source-Channel Model for Machine Transliteration. In *Proc. of ACL* 2004: 160-167.

T. Lin, C.-C. Wu, J.-S. Chang. 2003. Word-Transliteration Alignment, In *Proc. of ROCLING XV*, 1-16.

W.-H. Lin and H.-H. Chen. 2002. Backward machine transliteration by learning phonetic similarity. In *Proc. of CONLL* 2002: 139-145.

W.-H. Lu., L.-F. Chien and H.-J. Lee. 2002. Translation of Web Queries using Anchor Text Mining, *ACM Transactions on Asian Language Information Processing* (TALIP), 159-172.

W.-H. Lu., L.-F. Chien and H.-J. Lee. 2004. Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. *ACM Transactions on Information Systems* 22(2): 242-269.

W.-Y. Ma and K.-J. Chen. 2003. A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction, In *Proc. of ACL workshop on Chinese Language Processing* 2003*: 31-38.*

I. D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221-249.

H. Meng, W.-K. Lo, B. Chen and K. Tang. 2001. *Generate Phonetic Cognates to Handle Name Entities in English-Chinese Cross-Language Spoken Document Retrieval*, ASRU 2001.

A. Pirkola, J. Toivonen, H. Keskustalo, K. Visala and K. Jarvelin. 2003. Fuzzy Translation of Cross-Lingual Spelling Variants, In *Proc. of SIGIR* 2003: 345-352.

Y. Qu and G. Grefenstette. 2004. Finding Ideographic Representations of Japanese Names Written in Latin Script via Language Identification and Corpus Validation In *Proc. of ACL* 2004: 184-191.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora, In *Proc. of ACL* 1999: 519-526.

R. Schwartz and Y.-L. Chow. 1990. The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypothesis. In *Proc. of ICCASP* 1990: 81-84.

F. Smadja, K. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1-38.

P. Virga and S. Khudanpur. 2003. Transliteration of Proper Names in Cross-Lingual Information Retrieval. *ACL 2003 workshop MLNER*.

S. Wan and C. M. Verspoor. 1998. Automatic English-Chinese name transliteration for development of multilingual resources. In *Proc. of ACL* 1998: 1352-1357.

J. Xiao, J. Liu and T.-S. Chua. 2002. "Extracting pronunciation-translated names from Chinese texts using bootstrapping approach", the 1st SIGHAN workshop on Chinese Language Processing , Taipei, Taiwan, Aug 2002.

Translation Divergence Analysis and Processing for Mandarin-English Parallel Text Exploitation

Shun-Chieh Lin, Jia-Ching Wang, and Jhing-Fa Wang
Department of Electrical Engineering, National Cheng Kung University.
701 No.1, Ta-Hsueh Road, Tainan City Taiwan R.O.C.
Tel: 886-6-2757575 Ext. 62341, Fax: 886-6-2761693
E-mail: wangjf@csie.ncku.edu.tw

Abstract

Previous work shows that the process of parallel text exploitation to extract transfer mappings between language pairs raises the capability of language translation. However, while this process can be fully automated, one thorny problem called "divergence" causes indisposed mapping extraction. Therefore, this paper discuss the issues of parallel text exploitation, in general, with special emphasis on divergence analysis and processing. In the experiments on a Mandarin-English travel conversation corpus of 11,885 sentence pairs, the perplexity with the alignments in IBM translation model is reduced averagely from 13.65 to 5.18 by sieving out inappropriate sentences from the collected corpus.

1. Introduction

Over the past decade, research has focused on the automatic acquisition of translation knowledge from parallel text corpora. Statistical-based systems build alignment models from the corpora without linguistic analysis [1,2]. Another class of systems analyzes sentences in parallel texts to obtain transfer structures or rules [6]. Previous work shows that the process of parallel text exploitation to extract transfer mappings (models or rules) between language pairs can raise the capability of language translation.

However, previous work is still hampered by the difficulties in transfer mapping extraction of achieving accurate lexical alignment and acquiring reusable structural correspondences. Although automatic extraction methods of lexical alignment and structural correspondences are introduced, they are not capable of handling exceptional cases like "divergence" presented in [4]. In general, divergence arises with variant lexical usage of role, position, and morphology between two languages. Therefore, while mapping extraction can be fully automated from parallel texts, divergence causes indisposed mapping extraction. Furthermore, the existence of translation divergences also makes adaptation from source structures into target structures difficult [5,7,8]. For parallel text exploitation, these divergences make the training process of transfer mapping extraction between languages impractical including parsing and word-level alignment, lexical-semantic lexicography, and syntactic structures. Therefore, study of parallel text exploitation needs a careful study of translation divergence.

The framework of this paper is as follows. A brief overview of parallel text exploitation is discussed in Section 2. In Section 3, translation divergence analysis and processing for Mandarin-English parallel texts is presented. Section 4 shows experimental results with the alignments in IBM translation model. Finally, generalized conclusions are presented in Section 5.

2. Overview of Statistical-based Parallel Text Exploitation

The goal of parallel text exploitation is to acquire the knowledge for translation of a text given in some source ("Mandarin") string of words, *m* into a target ("English") string of words, *e*. For the presented statistical approach [1] to string translation of $\Pr(e|m)$, among all possible target strings, the string will be chosen with the highest probability which is given by Bayes' decision rule as follows:

$$\hat{e} = \arg\max_{e} \Pr(e)\Pr(m \mid e) \tag{1}$$

$\Pr(e)$ is the language model of target language and $\Pr(m| e)$ is the translation model. In order to estimate the correspondence between the words of the target sentence and the words of the source sentence, a sort of pair-wise dependence by considering all word pairs for a given sentence pair [*m*, *e*] is assumed, referred to as alignment models. Figure 1 shows an example for the translation parameters of a sentence pair. In general, these parameters are lexicon probability, ex. $p(m_j \mid e_i)$, sentence length probability, ex. $p(l_m \mid l_e)$, and alignment probability, ex. $p(j \mid i, l_m, l_e)$. Therefore, given more parallel texts, more probability parameters could be estimated for translation.

Mandarin: $l_m = 4$

$m = m_1 m_2 \cdots m_j \cdots m_{l_m}$:(一)$_1$ (晚)$_2$ (多少)$_3$ (錢)$_4$?

English: $l_e = 4$

$e = e_1 e_2 \cdots e_i \cdots e_{l_e}$:(*How much*)$_1$ (*for*)$_2$ (*a*)$_3$ (*night*)$_4$?

$p(m_2 \mid e_4)$: lexicon probability

$p(l_m = 4 \mid l_e = 4)$: sentence length probability

$p(j = 2 \mid i = 4, l_m, l_e)$: alignment probability

Fig. 1. An example for the translation parameters of a sentence pair

However, it is difficult to achieve straightforward and correct estimation for these probability parameters. In the above example, the English word "for" is one major factor called "*divergence*" makes the estimation process between sentence pairs impractical. Therefore, in the next section, we present the analysis and processing of the translation divergence for improving the performance on parallel text exploitation.

3. Translation Divergence Analysis and Processing

3.1 Analysis of Divergence Problems

Dorr's work [3] of divergence analysis is based on English-Spanish and English-German translations. Based on these two language pairs, 5 different categories have been identified. In this section, we discuss more multiform examples among the 5 types of divergences in Mandarin-English parallel texts. For each example, three sentences are given: *e* means an original English sentence in parallel texts, *m* means a Mandarin sentence, and *ẽ* means an amended English sentence which is better for translation parameter training with *m*.

3.1.1 Identification of Thematic Divergence

Thematic divergence often involves a "swap" of the subject and object position and obtains unpredictable word-level alignment. For example,

e: (*Is*)$_1$ (*credit card*)$_2$ (*acceptable*)$_3$ (*to*)$_4$ (*them*)$_5$?
m: (*他們*)$_1$ (*接受*)$_2$ (*信用卡*)$_3$ (*嗎*)$_4$?
ẽ: (*Do*)$_1$ (*they*)$_2$ (*accept*)$_3$ (*credit card*)$_4$?

Here, credit card appears in subject position in *e* and in object position ("*信用卡*") in *m*; analogously, the object them appears as the subject they ("*他們*"). Therefore, for the thematic divergence, the position alignments of 2↔3 and 5↔1 are obtained in a sentence pair [*m*, *e*]. However, if a sentence pair [*m*, *ẽ*] can be provided, the position alignments of 1↔2, 2↔3, and 3↔4 are better for straightforward parameter estimation of $p(j \mid i, l_m, l_e)$.

3.1.2 Identification of Morphological Divergence

Morphological divergence involves the selection of a target-language word that is a morphological variant of the source-language equivalent and it raises the ambiguity of lexical-semantic lexicography.

e: (*May*)$_1$ (*I*)$_2$ (*have*)$_3$ (*your*)$_4$ (*signature*)$_5$ (*here*)$_6$?
m: (*請*)$_1$ (*你*)$_2$ (*在*)$_3$ (*這*)$_4$ (*簽名*)$_5$ (*好嗎*)$_6$?
ẽ: (*Could*)$_1$ (*you*)$_2$ (*sign*)$_3$ (*here*)$_4$?

In this example, the predicate is nominal (*signature*) in *e* but verbal ("*簽名*") in *m*. While inputting two sentence pairs [*m*, *e*] and [*m*, *ẽ*], the parameter estimation of $p(m_j \mid e_i)$ should be reformulated with two morphological translation conditions: $p(m_j, m_j \in V \mid e_i, e_i \in N)$ and $p(m_j, m_j \in V \mid e_i, e_i \in V)$. Therefore, with growing of various morphological translations, more

conditions would raise more complexity of lexicon transfer parameter estimation and cause more ambiguity of lexical-semantic lexicography.

3.1.3 Identification of Structural Divergence

In structural divergence, a verbal argument has a different syntactic realization in the target language and the appearance of the divergence causes additional syntactic structural mapping constructions.

e: (*About*)$_1$ (*the*)$_2$ (*center*)$_3$.
m: (*大概*)$_1$ (*在*)$_2$ (*中間*)$_3$.
ẽ: (*About*)$_1$ (*in*)$_2$ (*the*)$_3$ (*center*)$_4$.

Observe that the place object is realized as a noun phrase (*the center*) in *e* and as a prepositional phrase ("*在 中間*") in *m*. For this example, the divergence causes the alignment of 0←2, which is a null mapping for Mandarin lexicon "*在*". In addition, the divergence also causes alignments of 2↔3 and 3↔3, which result in non-equal mapping number *q*-to-*n* (*q*>1, *n*>1, and *q*≠*n*). For a raised null mapping, the parameter estimation of $p(j \mid i, l_m, l_e)$ and $p(l_m \mid l_e)$ become more complicated by further considering translation of lexicon insertion (*i*=0) and deletion (*j*=0). More raised non-equal mapping number in parallel texts, more parameter estimation of $p(l_m \mid l_e)$ and more length generation condition for translation.

3.1.4 Identification of Conflational Divergence

Conflation is the incorporation of necessary participants (or arguments) of a given action. A conflational divergence arises when there is a difference in incorporation properties between two languages. In addition, there are word compounds in Chinese language by embedding some semantic contiguity. For this divergence, the complexity of training process for transfer mapping extraction is extremely increased.

e: (*Please*)$_1$ (*have*)$_2$ (*him*)$_3$ (*call*)$_4$ (*me*)$_5$.
m: (*請*)$_1$ (*轉告*)$_2$ (*他*)$_3$ (*回*)$_4$ (*個*)$_5$ (*電話*)$_6$ (*給*)$_7$ (*我*)$_8$.
ẽ: (*Please*)$_1$ (*tell*)$_2$ (*him*)$_3$ (*to*)$_4$ (*give*)$_5$ (*me*)$_6$ (*a*)$_7$ (*call*)$_8$.

This example illustrates the conflation of a constitution in *e* that must be overly realized in *m*: the effect of the action (*give me a call*) is indicated by the word "*回 個 電話 給 我*" whereas this information is incorporated into the main verb (*call me*) in *e*. Therefore, this divergence causes most complexity on parameter estimation of translation including $p(m_j \mid e_i)$, $p(l_m \mid l_e)$, and $p(j \mid i, l_m, l_e)$.

3.1.5 Identification of Lexical Divergence

For lexical divergence, the event is lexically realized as the main verb in one language but as a different verb in other language. It typically raises the ambiguity of lexical-semantic lexicography and also can be viewed as a side effect of other divergences. Thus, the formulation thereof is considered to be some combination of those given above, such as a conflational divergence forces the occurrence of a lexical divergence.

e: (*Nothing*)$_1$ (*can*)$_2$ (*beat*)$_3$ (*Phantom of the Opera*)$_4$.
m: (*沒有*)$_1$ (*什麼*)$_2$ (*比得上*)$_3$ (*歌劇魅影*)$_4$.
ẽ: (*Nothing*)$_1$ (*can*)$_2$ (*compare*)$_3$ (*with*)$_4$ (*Phantom of the Opera*)$_5$.

Here the main verb "*beat*" in *e* but as a different verb "*比得上*" (*to compare with*) in *m*. Other examples are like "*cash*", "*have*", "*take*", and etc. in English but "*兌換 成 現金*", "*轉告*", "*坐*", and etc. in Mandarin, respectively.

3.2 Processing of Divergence Evaluation

According to the above divergence analysis, the divergent mappings between sentence pairs are composed of non-equal mapping number (*q*-to-*n*, *q*>1, *n*>1, $q \neq n$), different position mapping ($i \leftrightarrow j$, $i \neq j$), and null mapping ($i \rightarrow 0$ or $0 \rightarrow j$). Unlike non-equal mapping number and different position mapping, the null mapping cannot provide target language translation information for lexical item selection and position generation. Therefore, we want to use a simple and straightforward measurement method to evaluate the possible null mappings.

For example to the Mandarin-English parallel text corpus, given a Mandarin sentence $m = m_1 m_2 \cdots m_j \cdots m_{l_m}$ and an English sentence $e = e_1 e_2 \cdots e_i \cdots e_{l_e}$:, direct lexical mappings in the mapping space can be extracted using the relevant bilingual dictionary [13]. The mapping function is defined as follows:

$$\tau(m_j, e_i) = \delta(m_j - \sigma_k) = \begin{cases} 1 & \text{if } \exists \sigma_k \in \Theta_{p_i}, \ni m_j = \sigma_k \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where m_j is *j*-th Mandarin segmented term; e_i is the *i*-th English phrase, and Θ_{p_i} is represented as a Mandarin lexicon set of the English phrase e_i in the chosen bilingual dictionary. The mapping function $\tau(m_j, e_i)$ has the factor σ_k , which represents *k*-th Mandarin lexicon in Θ_{p_i} . Therefore, if

the translation of e_i found in the bilingual dictionary is the same to m_j, $\tau(m_j, e_i)$ is assigned to 1; otherwise, $\tau(m_j, e_i)$ is assigned to 0. And we can obtain the direct lexical mapping sequence

$$\Delta_M = \left\{ a_j^i \mid 0 \le i \le I \text{ and } 0 \le j \le J \right\} \tag{3}$$

where a_j^i is a mapping referred to as the alignment $i \to j$ if $\tau(m_j, e_i) = 1$ or $i \to 0$ and $0 \to j$ if $\tau(m_j, e_i) = 0$.

If the lexical mapping sequence Δ_M contains more than a particular number, named ε_n, of null mappings ($i \to 0$ and $0 \to j$), then the degree of divergence between the sentence pairs [*m*, *e*] becomes significant. Hence, the content of *m* or *e* should be updated to improve the accuracy and effectiveness of exploration of mapping order between word sequences and derivation of transfer mappings. In this paper, we choose to sieve out the divergent sentence pairs from the parallel texts.

4. Experimental Results

Table 1 shows the basic characteristics of the collected parallel texts extended by travel conversation [11]. The Mandarin words in the corpora were obtained automatically using a Mandarin morphological analyzer at CKIP [10] and an English morphological analyzer referred to LinkGrammar [12].

Table 1. Basic characteristics of the collected parallel texts

	Mandarin	English
Number of sentences	11,885	11,885
Total number of words	80,699	66915
Number of word entries	6,278	5,118
Average number of words per sentence	6.79	5.63

The percentage of the various types of divergences for the collected parallel texts is shown on Fig. 2. For the collected corpus of travel conversation, almost two out of three parallel sentences (65 percent) occur the conflational divergence and less than one out of five parallel sentences (19 percent) occur the lexical divergence. In order to assess the effect of translation divergence in the parallel texts, the system also utilizes an alignment training tool called GIZA, which is a program in an EGYPT toolkit

designed by the Statistical Machine Translation team [9][1]. Based on segmented Chinese, we use the original GIZA for testing in this paper. In relation to the IBM models in GIZA, this study uses models 1-4 and ten iterations of each training models for the collected corpus. The parallel sentences with various types of divergences are sieved out from the collected corpus and perplexity in IBM original GIZA training model with comparison of sieving various types of divergences is shown on Table 2. The perplexity with sieving thematic divergence is similar to that with sieving structural divergence and the perplexity with sieving morphological divergence is similar to that with sieving lexical divergence. For sieving conflational divergence, a noticeable perplexity reduction is obtained among other types of divergence but the cost is that almost two out of three parallel sentences (65 percent) are sieved out from the collected corpus. Table 3 lists the perplexity of the original parallel sentences and that of the evaluated parallel sentences from GIZA. The results demonstrate that more null mappings can result in higher perplexity, i.e. more translation choices for a lexical item, thus increasing the translation ambiguity and lowering the accuracy of lexical mapping extraction. Two amended translation probabilities with evaluation of ε_n <1 are shown in Table 4. The number of translation choices of "*have*" and "*back*" are reduced from 7 to 4 and 7 to 3, respectively. After evaluating the divergence of each sentence pair in parallel texts and retaining those with ε_n <1, i.e. no null mappings in a sentence pair, the perplexity in the alignment training model can be reduced from 13.65 to 5.18 on average.

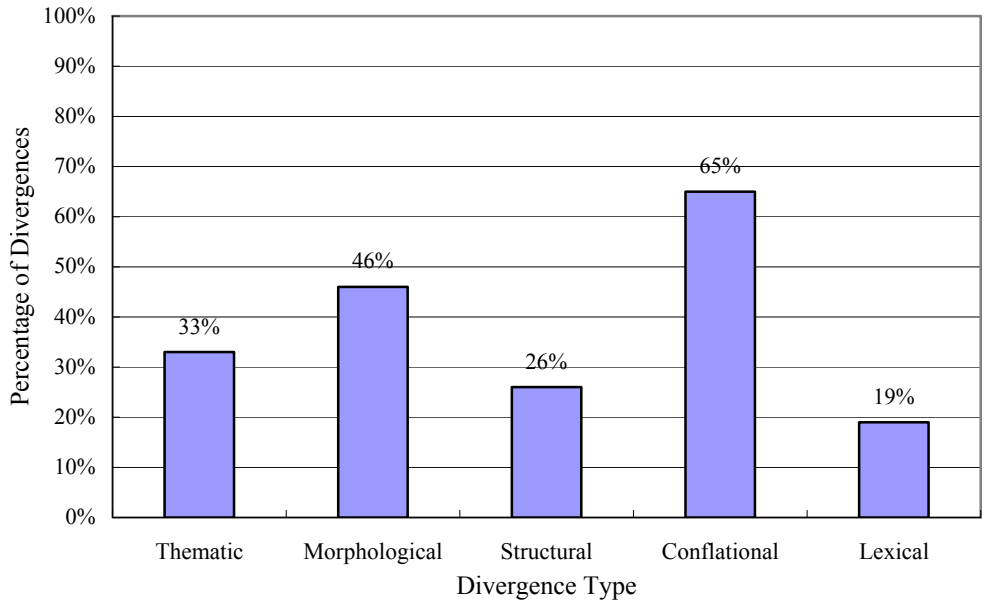


Fig. 2. The percentage of the various types of divergences for the collected parallel texts

[1] This toolkit could be downloaded from http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/

Table 2. Perplexity in IBM original GIZA training model with comparison of sieving various types of divergences.

	Thematic Divergence	Morphological Divergence	Structural Divergence	Conflational Divergence	Lexical Divergence
Model 1	9.91	10.17	9.41	8.46	10.70
Model 2	7.89	10.22	7.84	6.51	9.72
Model 3	8.72	11.83	8.56	7.48	12.35
Model 4	8.65	11.79	8.61	7.44	12.29
Average	8.79	11.00	8.61	7.47	11.26

Table 3. Perplexity in IBM original GIZA training model with comparison of original ($\varepsilon_n < \infty$) /evaluated parallel sentences.

	$\varepsilon_n < \infty$	$\varepsilon_n < 4$	$\varepsilon_n < 3$	$\varepsilon_n < 2$	$\varepsilon_n < 1$
No. of (m, e)	11,885	10,976	9,874	8,618	7,639
Model 1	10.94	10.09	8.26	6.79	5.98
Model 2	12.92	8.52	6.57	5.24	4.43
Model 3	15.39	9.47	7.13	6.21	5.16
Model 4	15.33	9.45	7.11	6.20	5.15
Average	13.65	9.38	7.27	6.11	5.18

Table 4. Examples of two amended English word translation probabilities.

Have			
Translation probability trained with original parallel sentences		Translation probability trained with evaluated parallel sentences	
已經	0.4312746	已經	0.4612446
有	0.346279	有	0.398176
給	0.1231011	給	0.1035049
你	0.0975747	叫	0.0370745
我	0.00146905		
轉告	0.000294352		
在	2.95704e-08		

Back			
Translation probability trained with original parallel sentences		Translation probability trained with evaluated parallel sentences	
回來	0.937283	回來	0.9392834
給	0.0379813	給	0.042981
能	0.01650713	能	0.01871713
錢	0.00786959		
在	2.5874e-06		
轉告	0.000294352		
何時	1.66024e-07		

5. Conclusion

In this work, we discuss one issue of parallel text exploitation, in general, with special emphasis on divergence analysis and processing. Experiments were performed for the languages of Mandarin and

English with the travel conversation corpus of 11,885 sentence pairs. The experimental results show that the analysis and evaluation of divergence for retaining low divergent parallel sentences can reduce the perplexity in IBM translation model averagely from 13.65 to 5.18. For sieving conflational divergence, a noticeable perplexity reduction is obtained among other types of divergence but the cost is that almost two out of three parallel sentences (65 percent) are sieved out from the collected corpus. Future studies will attempt to implement a translation decoder to assess the influence of divergence evaluation on BLEU sore.

References

[1] H. Ney, S. Nießen, F. J. Och, H. Sawaf, C. Tillmann, and S. Vogel, "Algorithms for statistical translation of spoken language," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 1, pp. 24–36, Jan. 2000.

[2] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[3] B. J. Dorr, P. W. Jordan and J. W. Benoit, "A survey of current paradigms in machine translation," in *Advances in Computers*, vol. 49, M. V. Zelkowitz, Ed. Academic Press, 1999.

[4] B. J. Dorr, *Machine translation: A view from the lexicon*. Cambridge, MA: The MIT press, 1993.

[5] B. J. Dorr, "Machine Translation Divergences: A Formal Description and Proposed Solution," *ACL* Vol. 20, No. 4, pp. 597–631, 1994.

[6] A. Menezes and S. D. Richardson, "A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora," *in Proc. Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics*, 2001, pp. 39–46.

[7] J. F. Wang and S. C. Lin, "Bilingual corpus evaluation and discriminative sentence vector expansion for machine translation," in *Proc. ICAIET*, 2002, pp.117–120.

[8] D. Gupta and N. Chatterjee, "Study of divergence for example based English-Hindi machine translation," in *Proc. STRANS*, 2002, pp. 132-140.

[9] *EGYPT toolkit*, developed by the Statistical Machine Translation team, Center for Language and Speech Processing, Johns-Hopkins University, MD, 1999.

[10] L. L. Chang, "The modality words in modern Mandarin," Chinese Knowledge Information Processing Group, Institute of Information Science Academia Sinica, Taiwan, Tech. Rep. 93-06, 1993.

[11] 徐歡, *旅遊英文 Easy Go*, 廣讀書城出版社, 2001.

[12] D. D. Sleator and D. Temperley, "Parsing English with a Link Grammar," Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-91-196, 1993.

[13] *Dr. Eye 譯典通 6.0*, developed by Inventec Corporation, 2004.