相似度比率式鑑別分析應用於大詞彙連續語音辨識

Likelihood Ratio Based Discriminant Analysis for Large Vocabulary Continuous Speech Recognition

李鴻欣　Hung-Shin Lee
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
hungshin@live.com

陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@ntnu.edu.tw

摘要

在近十年來所發展出的自動語音辨識(automatic speech recognition, ASR)技術中，仍有許多研究者嘗試僅藉由前端處理來產生具有鑑別性的語音特徵，而獨立於後端模型訓練與分類器特性。本論文即在此思維下提出嶄新的鑑別式特徵轉換方法，稱爲普遍化相似度比率鑑別分析(generalized likelihood ratio discriminant analysis, GLRDA)，其旨在利用相似度比率檢驗(likelihood ratio test)的概念尋求一個維度較低的特徵空間。在此子空間中，我們不僅考慮了全體資料的異方差性(heteroscedasticity)，即所有類別之共變異矩陣可被彈性地視爲相異，並且在分類上，因著我們也將類別間最混淆之情況（由虛無假設(null hypothesis)所描述）的發生率降至最低，而達到有助於分類正確率提升的效果。同時，我們也證明了傳統的線性鑑別分析(linear discriminant analysis, LDA)與有名的異方差性線性鑑別分析(heteroscedastic linear discriminant analysis, HLDA)可被視爲我們所提出之普遍化相似度比率鑑別分析(GLRDA)的兩種特例。此外，爲了增進語音特徵的強健性，我們所提出的方法更可進一步地與辨識器所提供的實際混淆資訊結合，而獲得在中文大詞彙連續語音辨識的實驗中，相較於以上兩種傳統方法更高的辨識正確率。

關鍵詞：語音辨識、特徵擷取、相似度比率、鑑別分析、混淆資訊

一、緒論

爲了降低計算量與模型的複雜度，語音特徵轉換(feature transformation)在自動語音辨識(automatic speech recognition, ASR)中扮演了很重要的角色。它的目標在於尋求一個

線性轉換$\Theta \in \Re^{n\times d}$，將原有在 n 維空間的聲學特徵向量，投影至 d 維的子空間($d < n$)，使得新的特徵在資料類別間具有較好的鑑別力[1]。而在實務上，語音特徵轉換的技術可被分爲兩種範疇[2]：相依於分類器(classifier-dependent)與獨立於分類器(classifier-independent)。在相依於分類器的範疇中，如某些基於最小音素錯誤(minimum phone error, MPE)[3]與最小分類錯誤(minimum classification error, MCE)[4]的鑑別式方法，轉換矩陣是結合聲學模型(acoustic models)中的參數估測或是分類器所掌握的分類規則一併求得。相對地，獨立於分類器的範疇則是基於各種不同的類別分離度標準，特別是幾何分離度，在聲學模型訓練之前就依據既有的類別統計資訊求出轉換矩陣。例如，線性鑑別分析(linear discriminant analysis, LDA)即試圖最大化類別間的平均馬氏距離平方(squared Mahalanobis distance)[5]；而做爲線性鑑別分析(LDA)的普遍化，異質性鑑別性分析(heteroscedastic linear discriminant analysis, HLDA)則在最大化相似度(maximum likelihood)的框架下，處理每一類別具有相異之共變異矩陣的情形[6]。另外，異質性鑑別分析(heteroscedastic discriminant analysis, HDA)個別地考慮了每一類別的分佈，而產生新的目標函式[7]；爲了保留比線性鑑別分析(LDA)和異質性鑑別性分析(HLDA)更多的鑑別資訊，最大化交互資訊(maximum mutual information, MMI)和最小分類錯誤(MCE)標準也被引入此範疇中使用[8]。此外，最近的研究者開始將成對經驗錯誤率(pairwise empirical error rate)列入考量，期望辨識器在前端處理與後端分類階段的不一致性能夠降低至一定程度[9-11]。

對於獨立於分類器的範疇來說，雖然在類別分離度與辨識結果之間仍存有較大的差距，也就是較高的類別分離度，並不必然保證有較低的辨識錯誤率。但在本論文中，我們仍將研究重點聚焦於此，原因在於：當語音特徵擷取完全與後端聲學模型分離，對於較複雜的自動語音辨識系統，聲學模型訓練模式的改變，就較不會影響到前端的訊號處理，使得此系統較易於被分析解構。而當某些系統的聲學模型機制是固定的，或是以硬體方式呈現，那麼我們就能在無法更動硬體的情況下，對前端訊號處理進行研究或改善[12]。更重要的是，我們相信在此範疇中所設計出的方法，能夠更廣泛地應用在其他圖型辨識(pattern recognition)的領域，如人臉辨識等，而不侷限於系統模型較爲複雜的語音辨識。

在本論文中，我們提出了一個嶄新的鑑別式特徵擷取方法，稱爲普遍化相似度比率鑑別分析(generalized likelihood ratio discriminant analysis, GLRDA)，其旨在利用相似度比率檢驗(likelihood ratio test, LRT)的概念來尋求一個維度較低的特徵空間。在此子空間中，我們不僅考慮了全體資料的異方差性(heteroscedasticity)，即所有類別母體之共變異矩陣可被彈性地視爲相異，並且在分類上，因著我們也將類別間最混淆之情況（由虛無假設(null hypothesis)所描述）的發生率降至最低，而達到有助於分類正確率提升的效果。此外，若我們假設所有類別母體均遵循高斯分佈(Gaussian distribution)，且針對其共變異矩陣給予不同的限制，則普遍化相似度比率鑑別分析(GLRDA)可被化約至傳統的線性鑑別分析與有名的異方差性線性鑑別分析。而爲了增進聲學特徵的強健性，我們的方法更可進一步地與辨識器所提供的經驗混淆資訊結合。

二、普遍化相似度比率鑑別分析

（一）背景

根據統計式假設檢定(statistical hypothesis testing)的定義[13]，相似度比率檢定(LRT)是一種廣爲使用的方法，藉著它我們可獲得虛無假設(null hypothesis)H_0與完全普遍化之對立假設(alternative hypothesis)H_1間相互比較的檢定統計量。在本論文中，虛無假設H_0通常表示不利於我們的目標設定，或我們不願見到的情況，在鑑別式特徵擷取上，即爲使類別母體不具鑑別性的情況。值得一提的是，虛無假設和對立假設之聯集(union)恰爲完整的參數空間。

若Ω表示完整的參數空間(parameter space)，而ω表示被虛無假設H_0所限制的參數子空間，則相似度比率檢定針對虛無假設H_0和對立假設H_1之間的標準爲

$$LR = \frac{\sup L_{\omega}}{\sup L_{\Omega}} \tag{1}$$

其中，L表示訓練樣本(sample)資料的相似度，$\sup L_S$則表示以 S 爲參數子空間時的最大相似度。由式(1)可看出，相似度比率檢定是由兩個部分組成：最大相似度與比率。使用最大化相似度估計法(maximum likelihood estimation, MLE)的用意在於找出最適合兩個統計假設或最具代表性的參數估計量。而相似度比率其背後的邏輯則在於，若我們不考慮任何信心度量測(confidence measure)且虛無假設H_0絕對爲真(true)，則在完整參數空間Ω中的最大相似度參數估測必定發生在參數子空間爲ω的情況；因此，$\sup L_{\omega}$與$\sup L_{\Omega}$必定會非常接近，使 LR 趨近於 1。反之，若 H_0絕對爲假(false)，則最大相似度發生的參數空間必定不是ω；因此，$\sup L_{\omega}$將會遠小於$\sup L_{\Omega}$。

（二）基本概念

相似度比率檢定在語音處理上的應用並不廣泛，近年來它常被用於評估音素間的混淆程度(phonetic confusions)[14]或是語音活動偵測(voice activity detection, VAD)[15]。在鑑別式語音特徵擷取技術中，我們並不打算緊密地遵照相似度比率檢定的過程，而目標也不在依據統計量來檢定虛無假設是真是假。我們的目標在於尋找一個投影子空間，使得虛無假設在此子空間中盡可能不會爲真。爲了使所有類別母體在此子空間中具有鑑別性，我們設計了以下的鑑別式統計假設：

$$\begin{cases} H_0 : \text{所有類別母體均相同} \\ H_1 : \text{所有類別母體均相異} \end{cases}$$

因此，我們所找到的子空間$\boldsymbol{\Theta} \in \Re^{n \times d}$，必須盡可能地推翻不具鑑別性的虛無假設 H_0，也就是使其相似度最小。普遍化相似度比率鑑別分析(generalized likelihood ratio discriminant analysis, GLRDA)目標函式便可寫成

$$J_{\mathrm{GLRDA}}(\boldsymbol{\Theta}) = LR_{\mathrm{GLRDA}}(\boldsymbol{\Theta}) = \frac{\sup L_{\text{所有類別母體均相同的參數子空間}}(\boldsymbol{\Theta})}{\sup L_{\text{完整的參數空間}}(\boldsymbol{\Theta})} \tag{2}$$

轉換矩陣$\boldsymbol{\Theta}$便可藉由最小化$J_{\mathrm{GLRDA}}(\boldsymbol{\Theta})$求得。

（三）同方差性(Homoscedasticity)

一般來說，我們會以類別母體之期望值向量的估計量所形成的空間作爲判斷所有類

別母體是否相同的參數空間。若所有類別母體具同方差性(homoscedasticity)，也就是每一類別母體具有相同的共變異矩陣，則令 $\boldsymbol{\mu}_i$ 爲每一類別母體 C_i 的期望值向量， $\boldsymbol{\Sigma}_i$ 爲每一類別母體 C_i 的共變異矩陣， H_0^{homo} 和 H_1^{homo} 可設定爲：

$$\begin{cases} H_0^{\text{homo}} : \text{對於每一類別} C_i,\ \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}，\ \boldsymbol{\mu}_i = \boldsymbol{\mu}。 \\ H_1^{\text{homo}} : \text{對於每一類別} C_i,\ \text{且} \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}，\ \boldsymbol{\mu}_i \text{不受任何限制。} \end{cases}$$

H_0^{homo} 代表了一種極端的情況，若它爲真，則所有類別母體幾近完全重疊，也就沒有任何鑑別性。因此，普遍化相似度比率鑑別分析(GLRDA)的任務即在 H_0^{homo} 最不可能爲真的情況下，找出最合適的投影子空間。

以下的命題證明了若每一類別母體均遵循高斯分佈(Gaussian distribution)，則線性鑑別分析(LDA)轉換矩陣，等同於將 H_0^{homo} 與 H_1^{homo} 置於普遍化相似度比率鑑別分析(GLRDA)的框架下求解。

命題一：若每一類別母體 C_i 都具有高斯分佈，則最小化普遍化相似度比率鑑別分析(GLRDA)的目標函式

$$J_{\text{GLRDA}}^{\text{homo}}(\boldsymbol{\Theta}) = \frac{\sup L_{H_0^{\text{homo}}}(\boldsymbol{\Theta})}{\sup L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})} \tag{3}$$

等同於最大化線性鑑別分析(LDA)的目標函式

$$J_{\text{LDA}}(\boldsymbol{\Theta}) = \frac{|\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta}|}{|\boldsymbol{\Theta}^T \mathbf{S}_W \boldsymbol{\Theta}|} \tag{4}$$

其中， $\mathbf{S}_B \in \Re^{n \times n}$ 和 $\mathbf{S}_W \in \Re^{n \times n}$ 分別代表類別間散佈矩陣(between-class scatter matrix)與類別內散佈矩陣(within-class scatter matrix)[16]。

證明：爲了方便起見，我們先將式(3)取對數，這並不影響 $\boldsymbol{\Theta}$ 的求解：

$$\log J_{\text{GLRDA}}^{\text{homo}}(\boldsymbol{\Theta}) = \sup \log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta}) - \sup \log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta}) \tag{5}$$

而 $\log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta})$ 和 $\log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})$ 可被進一步分別表示爲樣本所有資料 $\mathbf{x}_1^N$ 在所屬於高斯分佈之類別母體下的相似度：

$$\begin{aligned} &\log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta}) = \log p(\mathbf{x}_1^N, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}) \\ &= g(N,d) - \sum_{i=i}^{C} \frac{n_i}{2} \left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}) + \text{trace}(\tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{S}}_i) + \log |\tilde{\boldsymbol{\Sigma}}| \right) \end{aligned} \tag{6}$$

$$\begin{aligned} &\log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta}) = \log p(\mathbf{x}_1^N, \{\boldsymbol{\mu}_i\}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}) \\ &= g(N,d) - \sum_{i=i}^{C} \frac{n_i}{2} \left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i)^T \tilde{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i) + \text{trace}(\tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{S}}_i) + \log |\tilde{\boldsymbol{\Sigma}}| \right) \end{aligned} \tag{7}$$

其中， $g(N,d) = (-Nd/2)\log(2\pi)$ ，N 爲樣本所有資料總數， n_i 爲類別 C_i 的資料數，C 爲類別總數，d 爲投影後之特徵子空間的維度（或特徵數）； $\tilde{\mathbf{m}}_i$ 和 $\tilde{\mathbf{S}}_i$ 分別爲經過 $\boldsymbol{\Theta}$ 轉換

後的樣本期望值向量與共變異矩陣，而$\tilde{\boldsymbol{\mu}}$、$\{\tilde{\boldsymbol{\mu}}_i\}$和$\tilde{\boldsymbol{\Sigma}}$則是根據統計假設$H_0^{\text{homo}}$與$H_1^{\text{homo}}$而設定之經過$\boldsymbol{\Theta}$轉換後的母體期望值向量與共變異矩陣，即我們所要估計的參數。

欲求得在假設H_0^{homo}下的最大相似度估計量$\tilde{\boldsymbol{\mu}}_0^{\text{homo}}$和$\tilde{\boldsymbol{\Sigma}}_0^{\text{homo}}$，可將式(6)分別對$\tilde{\boldsymbol{\mu}}$和$\tilde{\boldsymbol{\Sigma}}$偏微分，並令其爲 0，可得：

$$\begin{aligned}
&\frac{\partial \log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta})}{\partial \tilde{\boldsymbol{\mu}}} = \frac{\partial\left(-\sum_{i=i}^{C}\frac{n_i}{2}\left((\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}})+\text{trace}(\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{S}}_i)+\log|\tilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial\tilde{\boldsymbol{\mu}}} \\
&= -\sum_{i=i}^{C} n_i\left(\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}})\right)=0 \\
&\Rightarrow \tilde{\boldsymbol{\mu}}_0^{\text{homo}} = \sum_{i=i}^{C}\frac{n_i}{N}\tilde{\mathbf{m}}_i = \tilde{\mathbf{m}}
\end{aligned} \tag{8}$$

$$\begin{aligned}
&\frac{\partial \log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta})}{\partial \tilde{\boldsymbol{\Sigma}}} = \frac{\partial\left(-\sum_{i=i}^{C}\frac{n_i}{2}\left((\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{homo}})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_0^{\text{homo}})+\text{trace}(\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{S}}_i)+\log|\tilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial\tilde{\boldsymbol{\Sigma}}} \\
&= -\sum_{i=i}^{C} n_i\tilde{\boldsymbol{\Sigma}}^{-1} + \sum_{i=i}^{C} n_i\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{S}}_i\tilde{\boldsymbol{\Sigma}}^{-1} + \sum_{i=i}^{C} n_i\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{m}}_i-\tilde{\mathbf{m}})\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{m}}_i-\tilde{\mathbf{m}})^T\tilde{\boldsymbol{\Sigma}}^{-1}=0 \\
&\Rightarrow \tilde{\boldsymbol{\Sigma}}_0^{\text{homo}} = \left(\sum_{i=i}^{C}\frac{n_i}{N}\tilde{\mathbf{S}}_i\right)+\left(\sum_{i=i}^{C}\frac{n_i}{N}(\tilde{\mathbf{m}}_i-\tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i-\tilde{\mathbf{m}})^T\right)=\tilde{\mathbf{S}}_W+\tilde{\mathbf{S}}_B=\tilde{\mathbf{S}}_T
\end{aligned} \tag{9}$$

其中$\mathbf{S}_T \in \Re^{n\times n}$爲全體散佈矩陣(total scatter matrix)[16]。將$\tilde{\boldsymbol{\mu}}_0^{\text{homo}}=\tilde{\mathbf{m}}$和$\hat{\boldsymbol{\Sigma}}_0^{\text{homo}}=\tilde{\mathbf{S}}_T$代入式(6)，可得在假設$H_0^{\text{homo}}$下的最大對數相似度：

$$\begin{aligned}
&\sup\log L_{H_0^{\text{homo}}}(\boldsymbol{\Theta}) = \max\log p(\mathbf{x}_1^N, \tilde{\boldsymbol{\mu}}_0^{\text{homo}}, \tilde{\boldsymbol{\Sigma}}_0^{\text{homo}}, \boldsymbol{\Theta}) \\
&= g(N,d) - \frac{N}{2}\log|\tilde{\mathbf{S}}_T| - \frac{Nd}{2}
\end{aligned} \tag{10}$$

同理，欲求得在假設H_1^{homo}下的最大相似度估計量$\{\tilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}}\}$和$\tilde{\boldsymbol{\Sigma}}_1^{\text{homo}}$，可將式(7)分別對$\tilde{\boldsymbol{\mu}}_i$和$\tilde{\boldsymbol{\Sigma}}$偏微分，並令其爲 0，可得：

$$\begin{aligned}
&\frac{\partial \log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})}{\partial \tilde{\boldsymbol{\mu}}_i} = \frac{\partial\left(-\sum_{i=i}^{C}\frac{n_i}{2}\left((\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_i)^T\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_i)+\text{trace}(\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{S}}_i)+\log|\tilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial\tilde{\boldsymbol{\mu}}_i} \\
&= n_i\left(\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{m}}_i-\tilde{\boldsymbol{\mu}}_i)\right)=0 \\
&\Rightarrow \tilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}} = \tilde{\mathbf{m}}_i
\end{aligned} \tag{11}$$

$$\frac{\partial \log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta})}{\partial \widetilde{\boldsymbol{\Sigma}}} = \frac{\partial\left(-\sum_{i=i}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}})+\text{trace}(\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}|\right)\right)}{\partial \widetilde{\boldsymbol{\Sigma}}}$$
$$= -\sum_{i=i}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1} + \sum_{i=i}^{C} n_i\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\mathbf{S}}_i\widetilde{\boldsymbol{\Sigma}}^{-1} = 0 \tag{12}$$
$$\Rightarrow \widetilde{\boldsymbol{\Sigma}}_1^{\text{homo}} = \left(\sum_{i=i}^{C}\frac{n_i}{N}\widetilde{\mathbf{S}}_i\right) = \widetilde{\mathbf{S}}_W$$

將 $\widetilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}} = \widetilde{\mathbf{m}}_i$ 和 $\widetilde{\boldsymbol{\Sigma}}_1^{\text{homo}} = \widetilde{\mathbf{S}}_W$ 代入式(7)，可得在假設 H_1^{homo} 下的最大對數相似度：

$$\sup \log L_{H_1^{\text{homo}}}(\boldsymbol{\Theta}) = \max \log p(\mathbf{x}_1^N, \{\widetilde{\boldsymbol{\mu}}_{1,i}^{\text{homo}}\}, \widetilde{\boldsymbol{\Sigma}}_1^{\text{homo}}, \boldsymbol{\Theta})$$
$$= g(N,d) - \frac{N}{2}\log|\widetilde{\mathbf{S}}_W| - \frac{Nd}{2} \tag{13}$$

最後，將式(10)與式(13)代入式(5)，可得在同方差性假設下的對數相似度比率：

$$\log J_{\text{GLRDA}}^{\text{homo}}(\boldsymbol{\Theta})$$
$$= \left(g(N,d) - \frac{N}{2}|\widetilde{\mathbf{S}}_T| - \frac{Nd}{2}\right) - \left(g(N,d) - \frac{N}{2}\log|\widetilde{\mathbf{S}}_W| - \frac{Nd}{2}\right)$$
$$= \frac{N}{2}(\log|\widetilde{\mathbf{S}}_W| - \log|\widetilde{\mathbf{S}}_T|) = \frac{N}{2}(\log|\widetilde{\mathbf{S}}_W| - \log(|\widetilde{\mathbf{S}}_B| + |\widetilde{\mathbf{S}}_W|) \tag{14}$$
$$= \frac{N}{2}\log\frac{|\widetilde{\mathbf{S}}_W|}{|\widetilde{\mathbf{S}}_B| + |\widetilde{\mathbf{S}}_W|} = \frac{N}{2}\log\frac{1}{|\widetilde{\mathbf{S}}_B| \big/ |\widetilde{\mathbf{S}}_W| + 1}$$

$\boldsymbol{\Theta}$ 可經由最小化式(14)來求出。因爲對數函數爲單調遞增(monotonically increasing)函數，所以$\boldsymbol{\Theta}$ 也可藉由最大化式(14)中的$|\boldsymbol{\Theta}^T\mathbf{S}_B\boldsymbol{\Theta}|/|\boldsymbol{\Theta}^T\mathbf{S}_W\boldsymbol{\Theta}|$求得，而此項正好是線性鑑別分析(LDA)的目標函式（式(4)）。 ■

（四）異方差性(Heteroscedasticity)

現在，我們考慮異方差性的統計假設[17]：

$$\begin{cases} H_0^{\text{heter}}: \text{每一類別}C_i\text{均呈高斯分布，且}\boldsymbol{\mu}_i = \boldsymbol{\mu}\text{，}\boldsymbol{\Sigma}_i\text{不受任何限制。} \\ H_1^{\text{heter}}: \text{每一類別}C_i\text{均呈高斯分布，且}\boldsymbol{\mu}_i\text{與}\boldsymbol{\Sigma}_i\text{均不受任何限制。} \end{cases}$$

仿照命題一的方式，普遍化相似度比率鑑別分析(GLRDA)的目標函式可寫成：

$$\log J_{\text{GLRDA}}^{\text{heter}}(\boldsymbol{\Theta}) = \sup \log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta}) - \sup \log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta}) \tag{15}$$

而 $\log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta})$ 和 $\log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})$ 可被分別進一步表示爲

$$\log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta}) = \log p(\mathbf{x}_1^N, \boldsymbol{\mu}, \{\boldsymbol{\Sigma}_i\}, \boldsymbol{\Theta})$$
$$= g(N,d) - \sum_{i=i}^{C} \frac{n_i}{2}\left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}) + \text{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i) + \log|\tilde{\boldsymbol{\Sigma}}_i|\right) \tag{16}$$

$$\log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta}) = \log p(\mathbf{x}_1^N, \{\boldsymbol{\mu}_i\}, \{\boldsymbol{\Sigma}_i\}, \boldsymbol{\Theta})$$
$$= g(N,d) - \sum_{i=i}^{C} \frac{n_i}{2}\left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i)^T \tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i) + \text{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i) + \log|\tilde{\boldsymbol{\Sigma}}_i|\right) \tag{17}$$

欲求得在假設 H_0^{heter} 下的最大相似度估計量 $\tilde{\boldsymbol{\mu}}_0^{\text{heter}}$ 和 $\tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}}$ ，可將式(16)分別對 $\tilde{\boldsymbol{\mu}}$ 和 $\tilde{\boldsymbol{\Sigma}}_i$ 偏微分，並令其爲 0，可得：

$$\frac{\partial \log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta})}{\partial \tilde{\boldsymbol{\mu}}} = \frac{\partial\left(-\sum_{i=i}^{C} \frac{n_i}{2}\left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}) + \text{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i) + \log|\tilde{\boldsymbol{\Sigma}}_i|\right)\right)}{\partial \tilde{\boldsymbol{\mu}}}$$
$$= -\sum_{i=i}^{C} n_i\left(\tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}})\right) = 0 \tag{18}$$
$$\Rightarrow \tilde{\boldsymbol{\mu}}_0^{\text{heter}} = \left(\sum_{i=i}^{C} n_i \tilde{\boldsymbol{\Sigma}}_i^{-1}\right)^{-1} \sum_{i=i}^{C} n_i \tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{m}}_i \approx \left(\sum_{i=i}^{C} n_i \tilde{\mathbf{S}}_i^{-1}\right)^{-1} \sum_{i=i}^{C} n_i \tilde{\mathbf{S}}_i^{-1}\tilde{\mathbf{m}}_i$$

$$\frac{\partial \log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta})}{\partial \tilde{\boldsymbol{\Sigma}}_i} = \frac{\partial\left(-\sum_{i=i}^{C} \frac{n_i}{2}\left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T \tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_0^{\text{heter}}) + \text{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i) + \log|\tilde{\boldsymbol{\Sigma}}_i|\right)\right)}{\partial \tilde{\boldsymbol{\Sigma}}_i}$$
$$= -\frac{1}{2} n_k\left(-\tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_0^{\text{heter}})(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T \tilde{\boldsymbol{\Sigma}}_i^{-1} - \tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i\tilde{\boldsymbol{\Sigma}}_i^{-1} + \tilde{\boldsymbol{\Sigma}}_i^{-1}\right) = 0 \tag{19}$$
$$\Rightarrow \tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}} = (\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_0^{\text{heter}})(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T + \tilde{\mathbf{S}}_i = \tilde{\mathbf{B}}_i + \tilde{\mathbf{S}}_i$$

其中， $\tilde{\mathbf{B}}_i = (\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_0^{\text{heter}})(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_0^{\text{heter}})^T$ 。值得一提的是，在式(18)中， $\tilde{\boldsymbol{\mu}}$ 的形式中含有尚未估計出的 $\tilde{\boldsymbol{\Sigma}}_i$ ，因此我們只能先令 $\tilde{\boldsymbol{\Sigma}}_i = \tilde{\mathbf{S}}_i$ ，得到 $\tilde{\boldsymbol{\mu}}$ 的近似估計量。將式(18)和式(19)中估計出的 $\tilde{\boldsymbol{\mu}}_0^{\text{heter}}$ 和 $\tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}}$ 代入式(16)，可得在假設 H_0^{heter} 下的最大對數相似度：

$$\sup \log L_{H_0^{\text{heter}}}(\boldsymbol{\Theta}) = \max \log p(\mathbf{x}_1^N, \tilde{\boldsymbol{\mu}}_0^{\text{heter}}, \{\tilde{\boldsymbol{\Sigma}}_{0,i}^{\text{heter}}\}, \boldsymbol{\Theta})$$
$$= g(N,d) - \frac{Nd}{2} - \sum_{i=1}^{C} \frac{n_i}{2} \log|\tilde{\mathbf{B}}_i + \tilde{\mathbf{S}}_i| \tag{20}$$

欲求得在假設 H_1^{heter} 下的最大相似度估計量 $\tilde{\boldsymbol{\mu}}_1^{\text{heter}}$ 和 $\tilde{\boldsymbol{\Sigma}}_{1,i}^{\text{heter}}$ ，可將式(17)分別對 $\tilde{\boldsymbol{\mu}}_i$ 和 $\tilde{\boldsymbol{\Sigma}}_i$ 偏微分，並令其爲 0，可得：

$$\frac{\partial \log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})}{\partial \tilde{\boldsymbol{\mu}}_i} = \frac{\partial\left(-\sum_{i=i}^{C} \frac{n_i}{2}\left((\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i)^T \tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i) + \text{trace}(\tilde{\boldsymbol{\Sigma}}_i^{-1}\tilde{\mathbf{S}}_i) + \log|\tilde{\boldsymbol{\Sigma}}_i|\right)\right)}{\partial \tilde{\boldsymbol{\mu}}_i}$$
$$= n_i \tilde{\boldsymbol{\Sigma}}_i^{-1}(\tilde{\mathbf{m}}_i - \tilde{\boldsymbol{\mu}}_i) = 0 \tag{21}$$
$$\Rightarrow \tilde{\boldsymbol{\mu}}_{1,i}^{\text{heter}} = \tilde{\mathbf{m}}_i$$

$$\frac{\partial \log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta})}{\partial \widetilde{\boldsymbol{\Sigma}}_i} = \frac{\partial\left(-\sum_{i=i}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_{1,i}^{\text{heter}})^T\widetilde{\boldsymbol{\Sigma}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_{1,i}^{\text{heter}})+\text{trace}(\widetilde{\boldsymbol{\Sigma}}_i^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\boldsymbol{\Sigma}}_i|\right)\right)}{\partial \widetilde{\boldsymbol{\Sigma}}_i}$$
$$= -\frac{n_i}{2}\left(-\widetilde{\boldsymbol{\Sigma}}_i^{-1}\widetilde{\mathbf{S}}_i\widetilde{\boldsymbol{\Sigma}}_i^{-1}+\widetilde{\boldsymbol{\Sigma}}_i^{-1}\right)=0 \tag{22}$$
$$\Rightarrow \widetilde{\boldsymbol{\Sigma}}_{1,i}^{\text{heter}} = \widetilde{\mathbf{S}}_i$$

將式(21)和式(22)中估計出的$\widetilde{\boldsymbol{\mu}}_1^{\text{heter}}$和$\widetilde{\boldsymbol{\Sigma}}_{1,i}^{\text{heter}}$代入式(17)，可得在假設$H_1^{\text{heter}}$下的最大對數相似度：

$$\begin{aligned}
&\sup\log L_{H_1^{\text{heter}}}(\boldsymbol{\Theta}) = \max\log p(\mathbf{x}_1^N,\{\widetilde{\boldsymbol{\mu}}_1^{\text{heter}}\},\{\widetilde{\boldsymbol{\Sigma}}_{1,i}^{\text{heter}}\},\boldsymbol{\Theta})\\
&= g(N,d)-\sum_{i=1}^{C}\frac{n_i}{2}\left((\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}}_i)^T\widetilde{\mathbf{S}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\mathbf{m}}_i)+\text{trace}(\widetilde{\mathbf{S}}_i^{-1}\widetilde{\mathbf{S}}_i)+\log|\widetilde{\mathbf{S}}_i|\right)\\
&= g(N,d)-\sum_{i=1}^{C}\frac{n_i}{2}\left(d+\log|\widetilde{\mathbf{S}}_i|\right)\\
&= g(N,d)-\frac{Nd}{2}-\sum_{i=1}^{C}\frac{n_i}{2}\log|\widetilde{\mathbf{S}}_i|
\end{aligned} \tag{23}$$

最後，將式(20)與式(23)代入式(15)，經過整理，可得：

$$\begin{aligned}
&\log J_{\text{GLRDA}}^{\text{heter}}(\boldsymbol{\Theta})\\
&=\left(g(N,d)-\frac{Nd}{2}-\sum_{i=1}^{C}\frac{n_i}{2}\log|\widetilde{\mathbf{B}}_i+\widetilde{\mathbf{S}}_i|\right)-\left(g(N,d)-\frac{Nd}{2}-\sum_{i=1}^{C}\frac{n_i}{2}\log|\widetilde{\mathbf{S}}_i|\right)\\
&=-\sum_{i=1}^{C}\frac{n_i}{2}(\log|\widetilde{\mathbf{B}}_i+\widetilde{\mathbf{S}}_i|-\log|\widetilde{\mathbf{S}}_i|)=-\sum_{i=1}^{C}\frac{n_i}{2}\log|\mathbf{I}_{(p\times p)}+\widetilde{\mathbf{S}}_i^{-1}\widetilde{\mathbf{B}}_i|\\
&=-\sum_{i=1}^{C}\frac{n_i}{2}\log|\mathbf{I}_{(p\times p)}+\widetilde{\mathbf{S}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_0^{\text{heter}})(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_0^{\text{heter}})^T|\\
&=-\sum_{i=1}^{C}\frac{n_i}{2}\log\left(1+(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_0^{\text{heter}})^T\widetilde{\mathbf{S}}_i^{-1}(\widetilde{\mathbf{m}}_i-\widetilde{\boldsymbol{\mu}}_0^{\text{heter}})\right)
\end{aligned} \tag{24}$$

因此，我們可得到異方差性之普遍化相似度比率鑑別分析(GLRDA)的目標函式：

$$G_{\text{H}}(\boldsymbol{\Theta}) = -\sum_{i=1}^{C}\frac{n_i}{2}\log\left(1+(\boldsymbol{\Theta}^T\mathbf{m}_i-\boldsymbol{\Theta}^T\boldsymbol{\mu}_0^{\text{heter}})^T(\boldsymbol{\Theta}^T\mathbf{S}_i\boldsymbol{\Theta})^{-1}(\boldsymbol{\Theta}^T\mathbf{m}_i-\boldsymbol{\Theta}^T\boldsymbol{\mu}_0^{\text{heter}})\right) \tag{25}$$

爲了使用梯度下降等遞迴式的最佳化技術求解$\boldsymbol{\Theta}$，式(25)對$\boldsymbol{\Theta}$的一階偏導數可寫成：

$$\frac{\partial G_{\text{H}}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = -\sum_{i=1}^{C} n_i\frac{(-\mathbf{S}_i\boldsymbol{\Theta}\widetilde{\mathbf{S}}_i^{-1}\widetilde{\mathbf{B}}_i+\mathbf{B}_i\boldsymbol{\Theta})\widetilde{\mathbf{S}}_i^{-1}}{1+\text{trace}(\widetilde{\mathbf{S}}_i^{-1}\widetilde{\mathbf{B}}_i)} \tag{26}$$

其中，$\mathbf{B}_i=(\mathbf{m}_i-\boldsymbol{\mu}_0^{\text{heter}})(\mathbf{m}_i-\boldsymbol{\mu}_0^{\text{heter}})^T$，$\widetilde{\mathbf{B}}_i=\boldsymbol{\Theta}^T\mathbf{B}_i\boldsymbol{\Theta}$，$\widetilde{\mathbf{S}}_i=\boldsymbol{\Theta}^T\mathbf{S}_i\boldsymbol{\Theta}$。

（五）討論與比較

表一、普遍化相似度比率鑑別分析(GLRDA)在不同假設下的統計量歸納表

統計假設	期望值向量 估計量	共變異矩陣 估計量	含重要項之 最大對數相似度
$H_0^{\text{homo}}\begin{cases}\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}\\ \boldsymbol{\mu}_i=\boldsymbol{\mu}\end{cases}$	$\tilde{\mathbf{m}}$	$\tilde{\mathbf{S}}_T$	$-\frac{N}{2}\log\lvert\tilde{\mathbf{S}}_T\rvert$
$H_1^{\text{homo}}\begin{cases}\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}\\ \boldsymbol{\mu}_i:\text{無限制}\end{cases}$	$\tilde{\mathbf{m}}_i$	$\tilde{\mathbf{S}}_W$	$-\frac{N}{2}\log\lvert\tilde{\mathbf{S}}_W\rvert$
$H_0^{\text{heter}}\begin{cases}\boldsymbol{\Sigma}_i:\text{無限制}\\ \boldsymbol{\mu}_i=\boldsymbol{\mu}\end{cases}$	$\left(\sum_{i=i}^{C}n_i\tilde{\mathbf{S}}_i^{-1}\right)^{-1}\sum_{i=i}^{C}n_i\tilde{\mathbf{S}}_i^{-1}\tilde{\mathbf{m}}_i$	$\tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i$	$-\sum_{i=1}^{C}\frac{n_i}{2}\log\lvert\tilde{\mathbf{B}}_i+\tilde{\mathbf{S}}_i\rvert$
$H_1^{\text{heter}}\begin{cases}\boldsymbol{\Sigma}_i:\text{無限制}\\ \boldsymbol{\mu}_i:\text{無限制}\end{cases}$	$\tilde{\mathbf{m}}_i$	$\tilde{\mathbf{S}}_i$	$-\sum_{i=1}^{C}\frac{n_i}{2}\log\lvert\tilde{\mathbf{S}}_i\rvert$

表一歸納了普遍化相似度比率鑑別分析(GLRDA)在不同假設下的統計量。它是一個較大之相似度比率的框架，不僅是線性鑑別分析(LDA)的普遍化形式（其轉換矩陣可由 H_0^{homo} 與 H_1^{homo} 的相似度比率得到），以下命題亦證明了它也是異方差性線性鑑別分析(HLDA)的普遍化形式。

命題二：異方差線性鑑別分析(HLDA)轉換矩陣可由普遍化相似度比率鑑別分析(GLRDA)中的 H_0^{homo} 與 H_1^{heter} 的相似度比率得到。也就是說，普遍化相似度比率鑑別分析(GLRDA)是異方差線性鑑別分析(HLDA)的普遍化形式。

證明：根據[6]，HLDA 目標函式爲

$$J_{\text{HLDA}}(\boldsymbol{\Theta})=-\frac{N}{2}\log\lvert\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}\rvert-\sum_{i=1}^{C}\frac{n_i}{2}\log\lvert\boldsymbol{\Theta}_d^T\mathbf{S}_i\boldsymbol{\Theta}_d\rvert+N\log\lvert\boldsymbol{\Theta}\rvert \tag{27}$$

因爲在此，$\boldsymbol{\Theta}$ 爲全秩矩陣(full-rank matrix)，且 $\boldsymbol{\Theta}_{(n\times n)}=[\boldsymbol{\Theta}_d,\boldsymbol{\Theta}_{(n-d)}]$，我們可證明[18]

$$\begin{aligned}&\lvert\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}\rvert=\lvert\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d\rvert\times\lvert\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}\rvert\\ &\Rightarrow\log\lvert\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}\rvert=\log\lvert\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d\rvert+\log\lvert\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}\rvert\\ &\Rightarrow\log\lvert\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}\rvert=\log\lvert\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}\rvert-\log\lvert\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d\rvert\end{aligned} \tag{28}$$

$$\begin{aligned}&N\log\lvert\boldsymbol{\Theta}\rvert-\frac{N}{2}\log\lvert\boldsymbol{\Theta}^T\mathbf{S}_T\boldsymbol{\Theta}\rvert\\ &=N\log\lvert\boldsymbol{\Theta}\rvert-\frac{N}{2}\log\lvert\boldsymbol{\Theta}\rvert-\frac{N}{2}\log\lvert\mathbf{S}_T\rvert-\frac{N}{2}\log\lvert\boldsymbol{\Theta}\rvert\\ &=-\frac{N}{2}\log\lvert\mathbf{S}_T\rvert\end{aligned} \tag{29}$$

將式(28)代入式(27)，且考慮式(29)所推導出的結果，我們可得出異方差線性鑑別分析(HLDA)的另一種目標函式表示法：

表二、MATBN 訓練語料之音素辨識中前 10 組最易混淆之音素模型配對

K	類別（音素）配對　（RCD 模型）		錯誤音框數
1	in　(ㄧㄣ)	ing　(ㄧㄥ)	66,353
2	an　(ㄢ)	eng　(ㄥ)	42,550
3	i　(ㄧ)	sil　(靜音)	31,796
4	u　(ㄨ)	sil　(靜音)	29,082
5	sic_e　(ㄜ的空聲母)	sil　(靜音)	26,134
6	sic_i　(ㄧ的空聲母)	sil　(靜音)	25,709
7	ing　(ㄧㄥ)	sil　(靜音)	21,629
8	g_u　(ㄍㄨ)	sil　(靜音)	19,197
9	ian　(ㄧㄢ)	ie　(ㄧㄝ)	17,212
10	sic_i　(ㄧ的空聲母)	i　(ㄧ)	17,022

$$J_{\mathrm{HLDA}}(\boldsymbol{\Theta}) = \underbrace{-\sum_{i=1}^{C}\frac{n_i}{2}\log|\boldsymbol{\Theta}_d^T\mathbf{S}_i\boldsymbol{\Theta}_d|}_{\sup\log L_{H_1^{\mathrm{heter}}}(\boldsymbol{\Theta})} \underbrace{-\left(-\frac{N}{2}\log|\boldsymbol{\Theta}_d^T\mathbf{S}_T\boldsymbol{\Theta}_d|\right)}_{\sup\log L_{H_0^{\mathrm{homo}}}(\boldsymbol{\Theta})} - \frac{N}{2}\log|\mathbf{S}_T| \qquad (30)$$

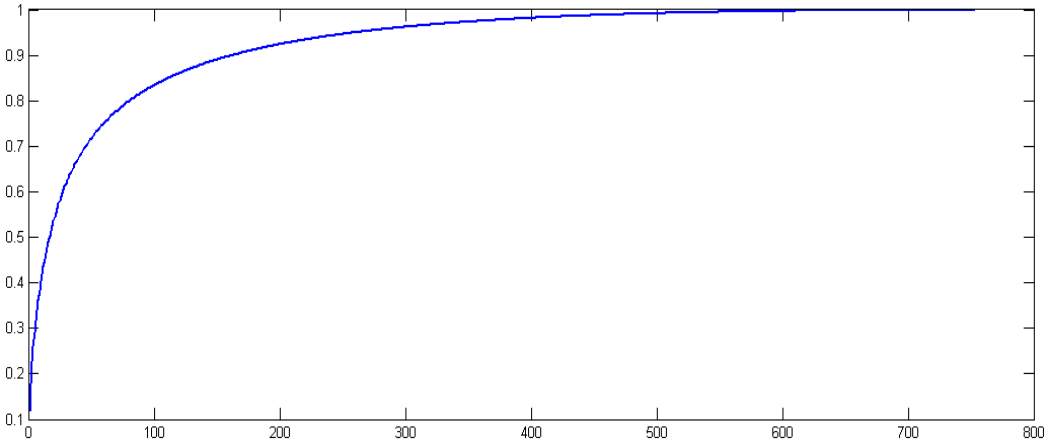
很明顯地，若不考慮常數$(-N/2)\log|\mathbf{S}_T|$，則式(30)等同於普遍化相似度比率鑑別分析(GLRDA)中，H_0^{homo}與H_1^{heter}含重要項之最大對數相似度比率（見表一）。 ■

由命題二可看出，異方差線性鑑別分析(HLDA)與異方差性之普遍化相似度比率鑑別分析(GLRDA)的主要差別在於虛無假設H_0的設定，異方差線性鑑別分析(HLDA)的虛無假設H_0^{homo}較爲嚴格，相對地要使它盡可能不發生的難度也較低。反之，異方差性之普遍化相似度比率鑑別分析(GLRDA)的虛無假設H_0^{heter}所產生的參數空間就比較大，應會使得找出的投影子空間在類別鑑別性上就較爲強健(robust)。

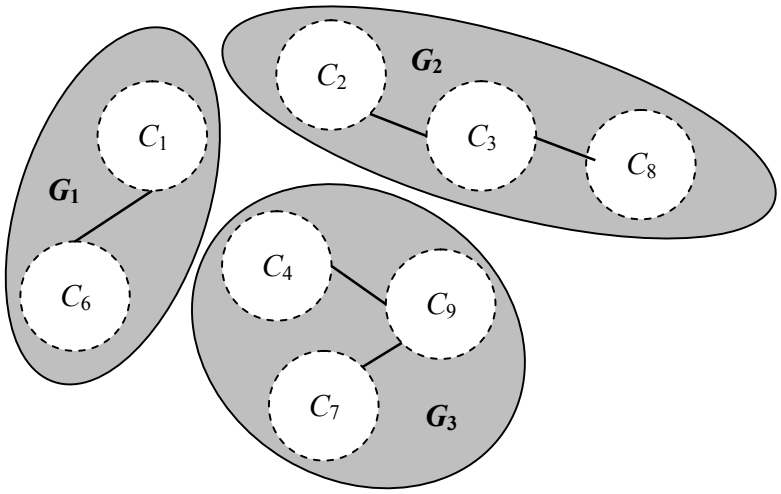
三、混淆資訊的延伸

由異方差性之普遍化相似度比率鑑別分析(GLRDA)的虛無假設H_0^{heter}來看，它設想每一類別母體的期望值向量在投影子空間中幾乎重疊在一起，但事實上這個假設的設定未必精確。在使用線性鑑別分析(LDA)作爲聲學特徵擷取方法之 MATBN 訓練語料（參見第四節）的音素辨識結果中，我們將所有類別配對，依其混淆程度由大至小排序，可得到前 K 個類別配對，如表二。舉例來說，在表二中，最易於混淆的音素類別配對是(in, ing)，其錯誤音框數爲 66,353，表示原本屬於音素 in 和 ing，卻分別被辨識器錯分至音素 ing 和 in 的音框總數。也就是說，以經驗資訊產生的事實來看，我們最不願意見到的假設，就是使這些易於混淆之類別配對的期望值向量幾乎重疊在一起，而非更廣泛的假設全體類別母體之期望值向量重疊，因爲這些易於混淆之類別配對才是全體錯誤的主要來源。再者，從圖一中我們可以發現約前 10%易於混淆之類別配對主導了約 80%的錯誤音框總數。

但是，若我們只考慮這些類別配對之期望值向量重疊的假設，則會發生以下情況：



圖一、前 K 組易於混淆之類別配對與累積錯誤音框比率圖
（橫軸爲 $K/10$，縱軸爲累積錯誤音框比率）



圖二、類別配對與群聚形成示意圖

若類別配對 C_1 與 C_2 爲最混淆之配對，則虛無假設可設定爲 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$；而類別配對 C_2 與 C_3 爲次混淆之配對，則虛無假設可增加 $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$。因此，整個虛無假設可合併爲 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$。也就是說，我們必須在類別配對集合中找到所有相關的類別配對以組成混淆群聚(confusable cluster)，如圖二。若我們把所有類別視爲圖形(graph)中的點(vertex)，而由易於混淆之類別配對所建立的關係視爲兩點之間的邊(edge)，則混淆群聚的產生可被視爲尋找圖形(graph)中所有的連通子圖(connected subgraph)。所以，我們可以使用一些圖論中的演算法，如滿水填充演算法(flood fill algorithm)[19]，來解決這個問題。

因此，我們可以將異方差性之普遍化相似度比率鑑別分析(GLRDA)改良成基於混淆資訊之普遍化相似度比率鑑別分析(confusion information based GLRDA, CI-GLRDA)：令 $G:\{G_k\}$ 爲所有根據前 K 組易於混淆之類別配對，並利用滿水填充演算法求出之群聚的集合，則其虛無假設與對立假設可設定如下：

表三、異方差性之普遍化相似度比率鑑別分析在不同期望值估計下之正確率(%)

GLRDA	Without MLLT	With MLLT
權重平均(weighted mean)	62.34	74.88
算術平均(arithmetic mean)	58.68	74.45

$$\begin{cases} H_0^{\mathrm{CI}}: \text{每一類別} C_i \text{均呈高斯分布，且} \boldsymbol{\Sigma}_i \text{不受任何限制，而若} C_i \in G_{l_i}\text{，則} \boldsymbol{\mu}_i = \boldsymbol{\mu}_{l_i}\text{。} \\ H_1^{\mathrm{CI}}: \text{每一類別} C_i \text{均呈高斯分布，且} \boldsymbol{\mu}_i \text{與} \boldsymbol{\Sigma}_i \text{均不受任何限制。} \end{cases}$$

其中，l_i 爲群聚編號，用來標示類別 C_i 所屬的群聚。

類似於上一節的最大化相似度估計法，我們可得到基於混淆資訊之普遍化相似度比率鑑別分析(confusion information based GLRDA, CI-GLRDA)目標函式：

$$G_{\mathrm{CI}}(\boldsymbol{\Theta}) = -\sum_{i=1, G_{l_i} \in G}^{C} \frac{n_i}{2} \log\left(1 + (\boldsymbol{\Theta}^T \mathbf{m}_i - \boldsymbol{\Theta}^T \boldsymbol{\mu}_{l_i}^{\mathrm{CI}})^T (\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta})^{-1} (\boldsymbol{\Theta}^T \mathbf{m}_i - \boldsymbol{\Theta}^T \boldsymbol{\mu}_{l_i}^{\mathrm{CI}})\right) \tag{31}$$

而式(31)對 $\boldsymbol{\Theta}$ 的一階偏導數亦可表示成：

$$\frac{\partial G_{\mathrm{CI}}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = -\sum_{i=1, G_{l_i} \in G}^{C} n_i \frac{(-\mathbf{S}_i \boldsymbol{\Theta} \widetilde{\mathbf{S}}_i^{-1} \widetilde{\mathbf{B}}_{l_i} + \mathbf{B}_{l_i} \boldsymbol{\Theta}) \widetilde{\mathbf{S}}_i^{-1}}{1 + \mathrm{trace}(\widetilde{\mathbf{S}}_i^{-1} \widetilde{\mathbf{B}}_{l_i})} \tag{32}$$

其中，$\mathbf{B}_{l_i} = (\mathbf{m}_i - \boldsymbol{\mu}_{l_i}^{\mathrm{CI}})(\mathbf{m}_i - \boldsymbol{\mu}_{l_i}^{\mathrm{CI}})^T$ ，$\widetilde{\mathbf{B}}_{l_i} = \boldsymbol{\Theta}^T \mathbf{B}_{l_i} \boldsymbol{\Theta}$ ，$\widetilde{\mathbf{S}}_i = \boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta}$ 。

四、實驗結果與分析

（一）實驗設定

本論文主要使用的語料庫爲 MATBN 中文電視新聞語料[20]，內含外場記者的語料總共約 27 小時，其中 24.5 小時（5,774 句，再切成 34,672 個短句供聲學模型訓練之用）做爲聲學模型訓練的語料，1 小時（230 句）爲辨識評估的資料，另有 1.5 小時（292 句）則爲發展集(developing set)，用來決定特殊參數的調整，如梯度下降等最佳化方法的遞迴次數或在上節中提到的 K。

在本論文中使用梅爾倒頻譜係數(Mel-frequency cepstral coefficients, MFCCs)作爲最基本的語音特徵參數。在聲學模型部分，我們將每個中文字視爲由一個聲母與一個韻母組成，採用傳統式由左至右的隱藏式馬可夫模型(hidden Markov models, HMMs)，分別爲聲母及韻母建立 INITIAL 與 FINAL 模型，並且考慮聲母會受右相連韻母影響其發音特性，所以採用右相關聯模型，(right-context-dependent model, RCD model)，加上一個靜音(silence)模型，總共有 151 個聲學模型[21]。聲學模型會透過 EM 演算法(expectation-maximization algorithm)，經過 10 次遞迴的最大化相似度訓練而得。而在語言模型方面，我們使用了詞二連以及詞三連語言模型(word bigram and trigram language models)，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個中文字語料作爲背景語言模型訓練時的訓練資料。本論文中的語言模型訓

表四、各種特徵擷取方法在中文大詞彙辨識系統下之詞正確率(%)

方法	Without MLLT	With MLLT
線性鑑別分析(LDA)	71.46	74.33
異方差性線性鑑別分析(HLDA)	70.28	74.88
異方差性鑑別分析(HDA)	71.36	74.53
異方差性之普遍化相似度比率鑑別分析(Heteroscedastic GLRDA)	62.34	**74.88**
基於混淆資訊之普遍化相似度比率鑑別分析(CI-GLRDA)	63.62	**75.26**

練工具採用 SRI Language Modeling Toolkit (SRILM)[22]。就以梅爾倒頻譜係數(MFCCs)爲基礎實驗(baseline)而言，其詞辨識正確率(character accuracy)爲 72.23%。

（二）實驗結果

本論文中所提到的重要特徵抽取方法，如線性鑑別分析(LDA)、異方差性線性鑑別分析(HLDA)、異方差性鑑別分析(HDA)，以及我們所提出的普遍化相似度比率鑑別分析(GLRDA)，都是作用在 162 維(n = 162)的超級向量(super-vector)上進行降維處理。此超級向量是由連續 9 個音框之梅氏濾波器組(Mel-frequency filterbank)所輸出的 18 維特徵向量串接而成，目的在於捕捉音框間的動態資訊。而目標維度則設定爲 39 維(d = 39)，其目的則在於使我們能在子空間維度固定的情況下，定性地比較各種方法的優劣。分類的最小單位則是以隱藏式馬可夫模型(HMMs)中的狀態(state)爲主，並經由一個辨識效果較高的系統針對每一訓練語句進行強制校準(forced alignment)，從而產生語句中的類別（音素和狀態）分界。而混淆資訊的獲得即以此爲正確答案，針對每一音框進行類別比對而得。由於這些特徵轉換方法所擷取出的特徵向量並不會使得各個類別的共變異矩陣爲對角化，會造成後端隱藏式馬可夫模型(HMMs)參數的估計失真，所以我們嘗試在這些方法後各自加上最大化相似度線性轉換(maximum likelihood linear transformation, MLLT)[23]。

在異方差性之普遍化相似度比率鑑別分析(GLRDA)的部分，由於在目標函式(25)中，$\boldsymbol{\mu}_0^{\text{heter}}$的部分是根據近似估計來的，我們在此採取了兩種近似方式，一種就是式(18)所提到的權重平均(weighted mean)，表示如下：

$$\boldsymbol{\mu}_0^{\text{heter}} = \left(\sum_{i=i}^{C} n_i \mathbf{S}_i^{-1} \right)^{-1} \sum_{i=i}^{C} n_i \mathbf{S}_i^{-1} \mathbf{m}_i \tag{33}$$

而另一種則是全體資料的算術平均(arithmetric mean)，表示如下：

$$\boldsymbol{\mu}_0^{\text{heter}} = \frac{1}{N} \sum_{i=i}^{N} \mathbf{x}_i \tag{34}$$

表三顯示了異方差性之普遍化相似度比率鑑別分析(GLRDA)在不同期望值估計量下之詞正確率。我們可以看出，以權重平均作爲$\boldsymbol{\mu}_0^{\text{heter}}$估計量的效果較好，因此，在之後基於混淆資訊的普遍化相似度比率鑑別分析(CI-GLRDA)，我們也會採用同樣的設定。此外，

我們也可以發現未經過 MLLT 處理的異方差性之普遍化相似度比率鑑別分析(GLRDA)，辨識率仍偏低。這是因爲根據其標準函式求出的轉換矩陣，並沒有使得每一類別的共變異矩陣趨於對角化的效果。

表四顯示出各種特徵擷取方法在大詞彙連續語音辨識之詞正確率。其中，異方差性鑑別分析(HDA)的目標函式可表示爲：

$$J_{\mathrm{HDA}}(\boldsymbol{\Theta}) = \sum_{i=1}^{C} n_i \log |\boldsymbol{\Theta}^T \mathbf{S}_i \boldsymbol{\Theta}| + N \log |\boldsymbol{\Theta}^T \mathbf{S}_B \boldsymbol{\Theta}| \qquad (35)$$

我們可以發現，各種針對線性鑑別分析(LDA)的改進方法均在辨識率上有些微進步。值得注意的是，在未加 MLLT 技術的情況，以線性鑑別分析(LDA)爲基礎的方法並不會優於梅爾倒頻譜係數（MFCCs），這也驗證了在大詞彙語音辨識工作上，以線性鑑別分析(LDA)爲基礎的方法會因著語料的不同而有或好或差的結果[24-25]。而在加入了 MLLT 技術之後，虛無假設內具有異方差性的方法，如異方差性線性鑑別分析(HLDA)與異方差性之普遍化相似度比率鑑別分析(Heteroscedastic GLRDA)，會有一致的實驗結果（74.88%），也都能看出打破傳統線性鑑別分析(LDA)中同方差性的效果。當我們進一步地加入混淆資訊後，並由發展集語料決定出最佳的 K 值（K = 100），辨識效果更好（75.26%），也證明了以經驗資訊作爲輔助，根據類別的特性設計出適當的虛無假設是有助於分類的。

五、結論與未來展望

本文的主要貢獻在於，基於相似度比率檢驗(LRT)的另一種涵義下，提出了嶄新的、更普遍化的框架進行鑑別式特徵擷取。我們的方法不僅可以使用在語音處理上，也能夠應用在其他需要特徵擷取的領域中，而獲得更具鑑別性的特徵。此外，我們的方法也能與類別混淆資訊結合，使其成爲理論與經驗兼具的方法。

普遍化相似度比率鑑別分析(GLRDA)的確具有進一步研究的空間：未來我們會嘗試各種將混淆類別分群的方法，找出屬於語音資料中，最混淆或使得辨識率最差的虛無假設表示法，使其更具分類上的代表性。

六、致謝

本研究承蒙國科會研究計畫 NSC 98-2221-E-003-011-MY3 、 NSC 96-2628-E-003-015-MY3、NSC 97-2631-S-003-003 的部分補助，僅此致謝。

參考文獻

[1] B. D. Ripley, *Pattern Recognition and Neural Networks*. New York: Cambridge University Press, 1996.
[2] X. Wang and K. K. Paliwal, "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern Recognition,* vol. 36, pp.

2429-2439, 2003.
[3] D. Povey, *et al.*, "fMPE: discriminatively trained features for speech recogntion," in *Proc. ICASSP*, 2005, pp. 961-964.
[4] X.-B. Li, *et al.*, "Dimensionality reduction using MCE-optimized LDA transformation," in *Proc. ICASSP*, 2004, pp. 137-140.
[5] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics,* vol. 8, pp. 376-386, 1938.
[6] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication,* vol. 26, pp. 283-297, 1998.
[7] G. Saon, *et al.*, "Maximum likelihood discriminant feature spaces," in *Proc. ICASSP*, 2000, pp. 1129-1132.
[8] K. Demuynck, *et al.*, "Optimal feature sub-space selection based on discriminant analysis " in *Proc. Eurospeech*, 1999, pp. 1311-1314.
[9] H.-S. Lee and B. Chen, "Linear discriminant feature extraction using weighted classification confusion information," in *Proc.Interspeech*, 2008, pp. 2254-2257.
[10] H.-S. Lee and B. Chen, "Improved linear discriminant analysis considering empirical pairwise classification error rates," in *Proc. ISCSLP*, 2008, pp. 149-152.
[11] H.-S. Lee and B. Chen, "Empirical error rate minimization based linear discriminant analysis," in *Proc. ICASSP*, 2009.
[12] X. Cui, *et al.*, "Stereo-based stochastic mapping with discriminative training for noise robust speech recognition," in *Proc. ICASSP*, 2009, pp. 2933-2936.
[13] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.
[14] Y. Liu and P. Fung, "Acoustic and phonetic confusions in accented speech recognition," in *Proc. Interspeech*, 2005, pp. 3033-3036.
[15] J. M. Górriz, *et al.*, "Generalized LRT-based voice activity detector," *IEEE Signal Processing Letters,* vol. 13, pp. 636-639, 2006.
[16] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic Press, 1990.
[17] N. A. Campbell, "Canonical variate analysis with unequal covariance matrices - generalizations of the usual solution," *Mathematical Geology,* vol. 16, pp. 109-124, 1984.
[18] M. Sakai, *et al.*, "Linear discriminant analysis using a generalized mean of class covariances and its application to speech recognition," *IEICE Trans. Information and Systems,* vol. E91-D, pp. 478-487, 2008.
[19] J. D. Foley, *et al.*, *Computer Graphics: Principles and Practice in C*, 2nd ed.: Addison-Wesley, 1995.
[20] H.-M. Wang, *et al.*, "MATBN: A mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics and Chinese Language Processing,* vol. 10, pp. 219-235, 2005.
[21] B. Chen, *et al.*, "Lightly supervised and data-driven approaches to mandarin broadcast news transcription," in *Proc. ICASSP*, 2004, pp. 777-780.
[22] A. Stolcke, *SRI Language Modeling Toolkit (Version 1.5.2)*.
[23] R. A. Gopinth, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP*, 1998, pp. 661-664.
[24] L. Wood, *et al.*, "Improved vocabulary-independent sub-word HMM modelling," in *Proc. ICASSP*, 1991, pp. 181-184.
[25] G. Yu, *et al.*, "Discriminant analysis and supervised vector quantization for continuous speech recognition," in *Proc. ICASSP*, 1990, pp. 685-688.

Wavelet Energy-Based Support Vector Machine for Noisy Word Boundary Detection With Speech Recognition Application

Chia-Feng Juang, Chun-Nan Cheng and Chiu-Chuan Tu

Department of Electrical Engineering
National Chung-Hsing University,
Taichung, 402 Taiwan, R.O.C.
e-mail: cfjuang@dragon.nchu.edu.tw

Abstract

Word boundary detection in variable noise-level environments by support vector machine (SVM) using Low-band Wavelet Energy (LWE) and Zero Crossing Rate (ZCR) features is proposed in this paper. The Wavelet Energy is derived based on Wavelet transformation; it can reduce the affection of noise in a speech signal. With the inclusion of ZCR, we can robustly and effectively detect word boundary from noise with only two features. For detector design, a Gaussian-kernel SVM is used. The proposed detection method is applied to detection word boundaries for an isolated word recognition system in variable noisy environments. Experiments with different types of noises and various signal-to-noise ratios are performed. The results show that using the LWE and ZCR parameters-based SVM, good performance is achieved. Comparison with another robust detection method has also verified the performance of the proposed method.

Keywords: Speech detection, word boundary detection, support vector machine, wavelet transform, noisy speech recognition.

1. INTRODUCTION

For speech recognition, the detection of speech affects recognition performance. A robust word boundary detection method in the presence of variable-label noises is necessary and is studied in this paper. Depending on the characteristics of speech, a variety of parameters have been proposed for boundary detection. They include the time energy (the magnitude in time domain), zero crossing rate (ZCR) [1] and pitch information [2]. These parameters usually fail to detect word boundary when signal-to-noise ratio (SNR) is low. Another parameter concerning frequency domain has also been recently proposed. According to the frequency energy, the time-frequency (TF) parameter [3] which sums the energy in time domain and the frequency energy was presented. The TF-based algorithm may work well for fixed-level background noise. However, its detection performance degrades for background noise of various levels. For this problem, some modified TF parameters are proposed [4]. In [5], the idea of using Wavelet transform features as speech detection features was proposed. In this paper, we present a new Low-band Wavelet Energy (LWE) parameter which separates the speech from noise in the domain of Wavelet transform. Computation of the WE parameter is easier than the modified TF parameters, and it is shown in the experiment section that a better detection performance is achieved.

After the features for detection have been extracted, the next step is to determine thresholds and decision rules. Many decision methods based on computational intelligence techniques have been proposed, such as fuzzy neural networks (FNNs) [4] and neural networks (NNs) [6]. Generalization performance may be poor when FNNs and NNs are over-trained. To cope with the low generalization ability problem, a new learning method, the Support Vector Machine (SVM), has been proposed [7, 8]. SVM is a new and useful learning method whose formulation is based on the principle of structural risk minimization. Instead of minimizing an objective function based on training, SVM attempts to minimize a bound on the generalization error. SVM has gained wide acceptance due to its high generalization abilities for a wide range of applications. For this reason, this paper used a SVM as a detector.

The rest of the paper is organized as follows. Section II introduces the derivation and analysis of the WE and ZCR parameters. Section III describes the SVM detector. Experiments on word boundary detection for noisy speech recognition are studied in Section IV. Finally, Section V draws conclusions.

2. ROBUST DETECTION PARAMETERS

Wavelet Transform (WT) is a technique for analyzing the time-frequency domain that is most suited for a non-stationary signal [9]. For short-time analysis and discrete speech signal, discrete-time WT (DTWT) is used. Let the amplitude of the k th point in the i th frame of a noisy speech signal be denoted by $s(i,k)$ and the frame length in sample number be represented by N. The DTWT of the i-th speech frame is as follows,

$$\mathrm{DTWT}(m,n)=\frac{1}{\sqrt{a_0^m}}\sum_{k=1}^{N}s(i,k)\psi(a_0^{-m}k-n\tau_0)\,, \tag{1}$$

where $\psi(\cdot)$ represents a wavelet basis function, a_0^m is the scale and τ_0 is a translation parameter which is set to a_0^{-m} in this paper. The commonly used value a_0=2 is used in this paper, resulting in a binary dilation. Thus, Eq. (1) can be written as

$$\mathrm{DTWT}(m,n)=\frac{1}{\sqrt{2^m}}\sum_{k=1}^{N}s(i,k)\psi[2^{-m}(k-n)] \tag{2}$$

.

In this paper, the Harr wavelet is used in Eq. (2), where

$$\psi[2^{-m}(k-n)]=\begin{cases}1\,, & 0\le 2^{-m}(k-n)\le\frac{1}{2}\\ -1\,, & \frac{1}{2}\le 2^{-m}(k-n)\le 1\\ 0\,, & \text{otherwise}\end{cases} \tag{3}$$

Generally, the DTWT is computed at scales a_0^m for, theoretically, all m. The output of DTWT can be regarded as finding the output of a bank of band-pass filters, where different values of scales corresponds to different band-pass filters. The outputs of DTWT at different

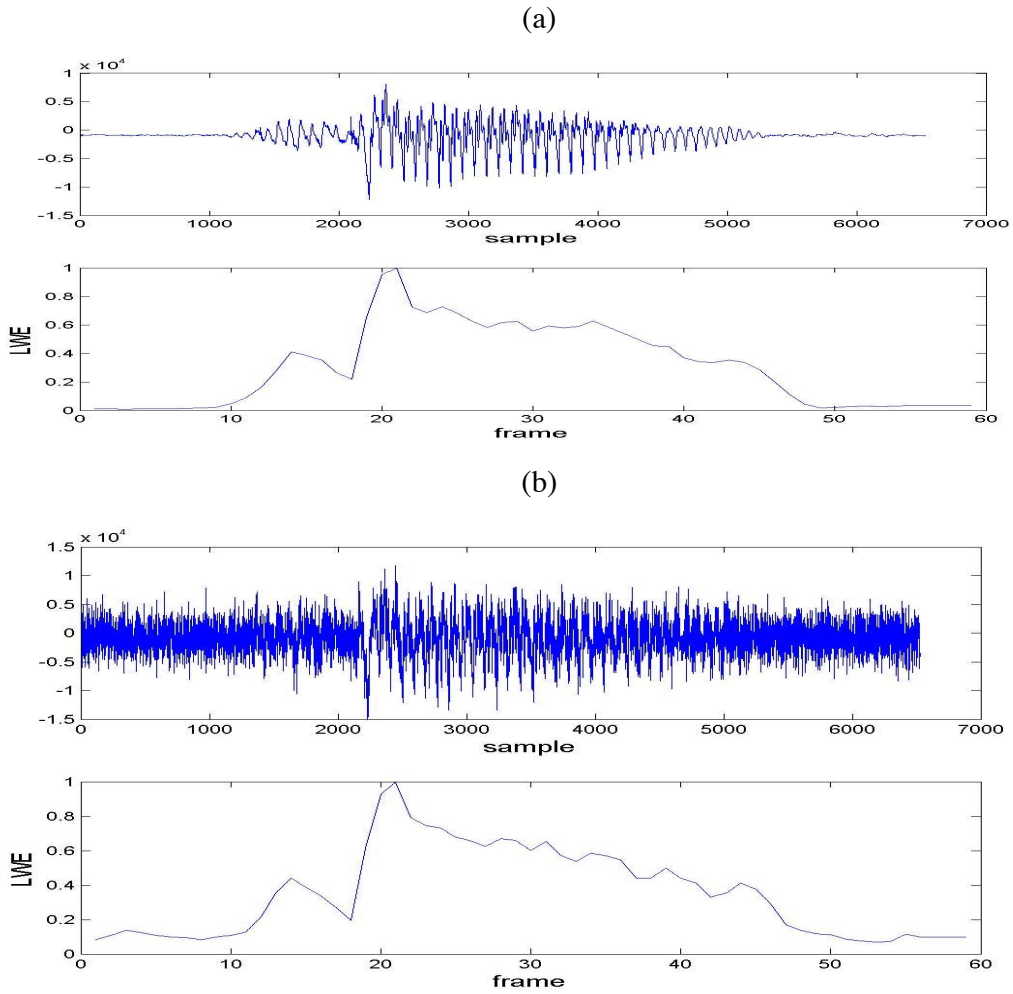


Fig. 1. (a) The LWEs of clean speech (b) The LWEs of speech with white noise added at SNR5.

scales contain different amounts of speech and noise information, and only the crucial scale(s) that contains maximum word signal information and is robust to noise should be used. Therefore, energy of the crucial scale is adopted as detection parameter for distinction between speech and noise in this paper.

To find the crucial scale, some observations on the effect of additive noise are made on different scales of DTWT. It is found that at the scale of $a_0^m = 2^6$, distribution of the STWT amplitudes matches well with the speech interval.

After computing DTWT for each time frame of a speech signal at the scale $a_0^m = 2^6$, the next step is to find an energy parameter to stand for the amount of word signal information at this scale. It is found the speech section corresponds to large DTWT amplitude values. Thus, summation of the amplitudes over n can be used as a parameter to stand for the amount of word signal information. It is also found that the amplitudes of noise tend to become larger when translation index n is larger than $0.8N$. Thus, summation is performed only from n=0 to $n = 0.8N$. This novel detection parameter, called low-band

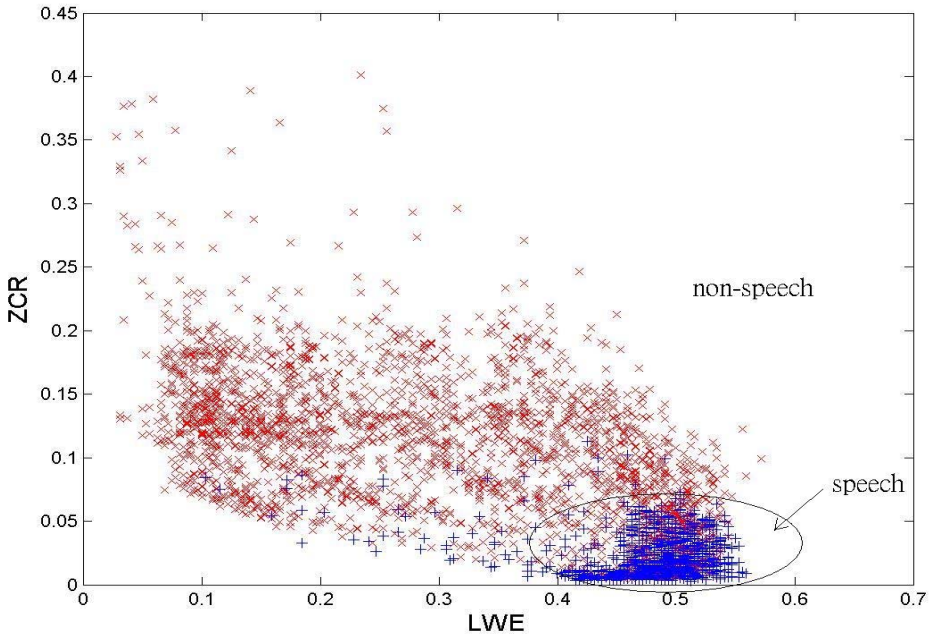


Fig. 2. Distributions of speech/non-speech frames in the LWE-ZCR plane with noise ranging from SNR20 to SNR0, where "×" and " +" denote non-speech and speech, respectively.

wavelet energy (LWE), is computed as follows,

$$\mathrm{LWE} = \sum_{n=0}^{0.8N} \left| \frac{1}{2^{-3}} \sum_{k=1}^{N} s(i,k)\psi(2^{-6}(k-n)) \right| \tag{4}$$

For illustration, a clean speech and its corresponding WE parameters of each frame are shown in Fig. 1(a). The speech with white noise and its corresponding WE parameters at SNR5 is shown in Fig. 1(b). This example shows that the WE parameter can robustly represent the energy of speech signal at different SNRs.

In addition to the WE parameter which is used to measure speech energy, the other parameter used for speech detection is the Zero Crossing Rate (ZCR). The reason for using the ZCR is that it is particularly suitable for un-voiced detection due to the high-frequency nature of the majority of fricatives.

Figure 2 shows distributions of speech/non-speech frames in the LWE-ZCR plane with noise levels SNR=20, 15, 10, and 5. The results show that the speech frames locate in a certain region of the two dimensional feature space.

3. SUPPORT VECTOR MACHINE DETECTOR

SVM is based on the statistical learning theory developed by Vapnik [7]. SVM first maps the input points into a high dimensional feature space and finds a separating hyperplane that maximizes the margin between two classes in this space. Suppose we are given a set S of labeled training set, $S = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \cdots, (\vec{x}_N, y_N)\}$, where $\vec{x}_i \in \mathbb{R}^n$, and $y_i \in \{+1, -1\}$. Considering that the training data is linearly non-separable, the goal of SVM is to find an optimal hyperplane such that

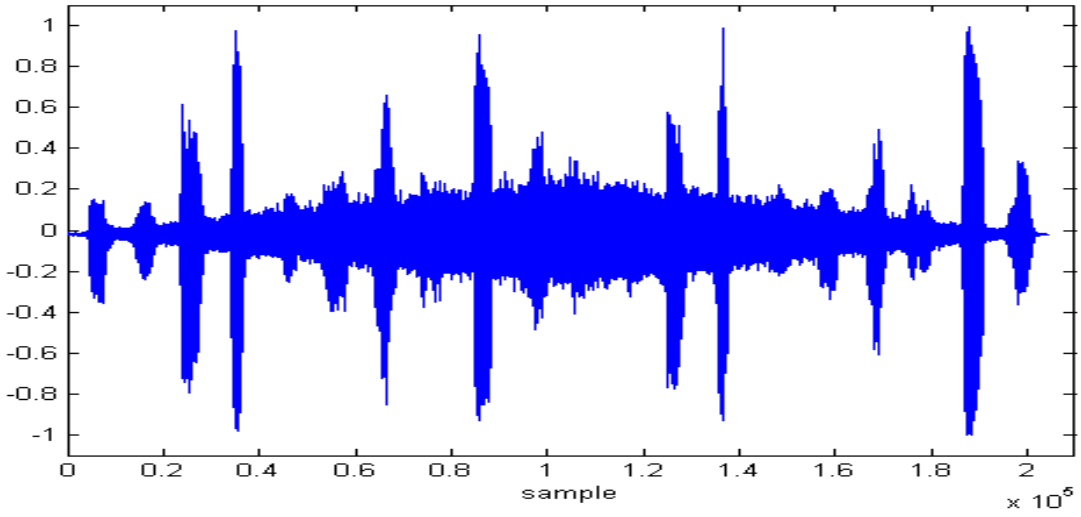


Fig. 3. The sequence of speech used for SVM training.

$$y_i\left(\vec{w}^T\vec{x}_i+b\right)\geq 1-\xi_i,\ i=1,\ldots,N \tag{5}$$

where $\vec{w}\in\mathbb{R}^n$, $b\in\mathbb{R}$, and $\xi_i\geq 0$ is a slack variable. For $\xi_i>1$, the data are misclassified. To find an optimal hyperplane is to solve the following constrained optimization problem:

$$\begin{aligned}&\text{Min}_{w,\xi}\quad \frac{1}{2}\vec{w}^T\vec{w}+C\sum_{i=1}^{N}\xi_i\\&\text{Subject to}\quad y_i(\vec{w}^T\vec{x}_i+b)\geq 1-\xi_i\end{aligned} \tag{6}$$

where C is a user defined positive cost parameter and $\sum\xi_i$ is an upper bound on the number of training errors. After solving Eq. (2), the final hyperplane decision function is achieved, and

$$f(\vec{x})=\text{sign}(\vec{w}^T\vec{x}+b)=\text{sign}(\sum_{i=1}^{N}y_i\alpha_i<\vec{x},\vec{x}_i>+b)=\text{sign}(\sum_{i\in SV}y_i\alpha_i<\vec{x},\vec{x}_i>+b) \tag{7}$$

where α_i is a Lagrange multiplier and the training samples for which $\alpha_i\neq 0$ are support vectors (SVs). A detailed derivation process can be found in [8].

The above linear SVM can be readily extended to a nonlinear classifier by first using a nonlinear operator Φ to map the input data into a higher dimensional feature space. In this way, it can solve nonlinear problems. By replacing $\vec{x}$ in Eqs. (1) and (2) with the feature space $\Phi(\vec{x})$ and solving the constrained optimization problem, the decision function

$$\begin{aligned}f(\vec{x})&=\text{sign}(\sum_{i=1}^{N}y_i\alpha_i<\Phi(\vec{x}),\Phi(\vec{x}_i)>+b)\\&=\text{sign}(\sum_{i=1}^{N}y_i\alpha_iK(\vec{x},\vec{x}_i)+b)\\&=\text{sign}(\sum_{i\in SV}y_i\alpha_iK(\vec{x},\vec{x}_i)+b)\end{aligned} \tag{8}$$

is achieved, where $K(\vec{x},\vec{x}_j)=\Phi(\vec{x})\cdot\Phi(\vec{x}_j)$ is called a kernel function. This paper uses a Gaussian-kernel SVM with $K(\vec{x},\vec{x}_j)=\exp(-\left\|\vec{x}-\vec{x}_j\right\|^2/\gamma)$, where γ is the width of a Gaussian-kernel. The two-dimensional inputs of the Gaussian-kernel SVM detector are ZER and LWE. For SVM, there is only one output and the desired output is "1" and "−1" if the input frame is speech and non-speech, respectively. During test, the SVM output indicates where or not the input frame is speech.

4. EXPERIMENTS

The wave files of speech are recorded by 11.025 kHz sample rate, mono channel and 16-bit resolution. For SVM training, the training sequence length is 13 seconds and is shown in Fig. 3. It consists of 20 words and is corrupted by white noise whose energy level increases from the start to SNR=0 and then decreases till the end of the sequence. For testing, the speech database is built of sequences of transcriptions from the same male speaker, where each sequence consists of ten isolated Mandarin words "0", "1",⋯, "9". There are a total of 50 test sequences used for playing the judicial role in performance comparison. The noise added to the speech sequence is of variable noise level during the sequence. Figure 4(a) shows the flowchart of training by

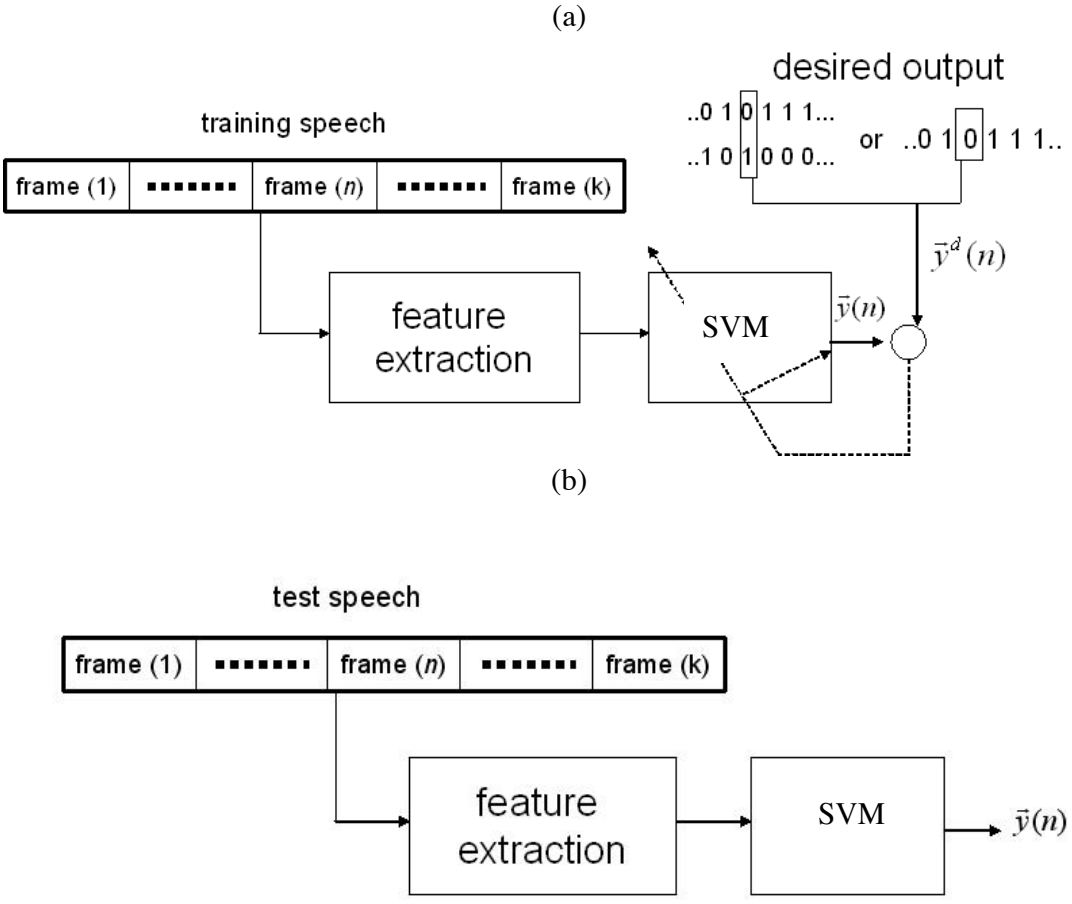


Fig. 4. (a) Flowchart of SVM training. (b) Flowchart of LWE-based SVM for test data.

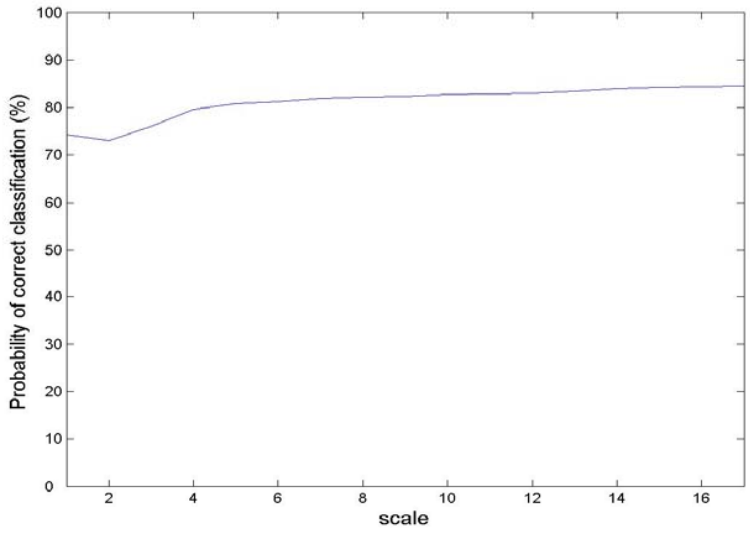


Fig. 5. Training performance of C in the range in the range [1, 85], where the range is spaced to 17 equal scales.

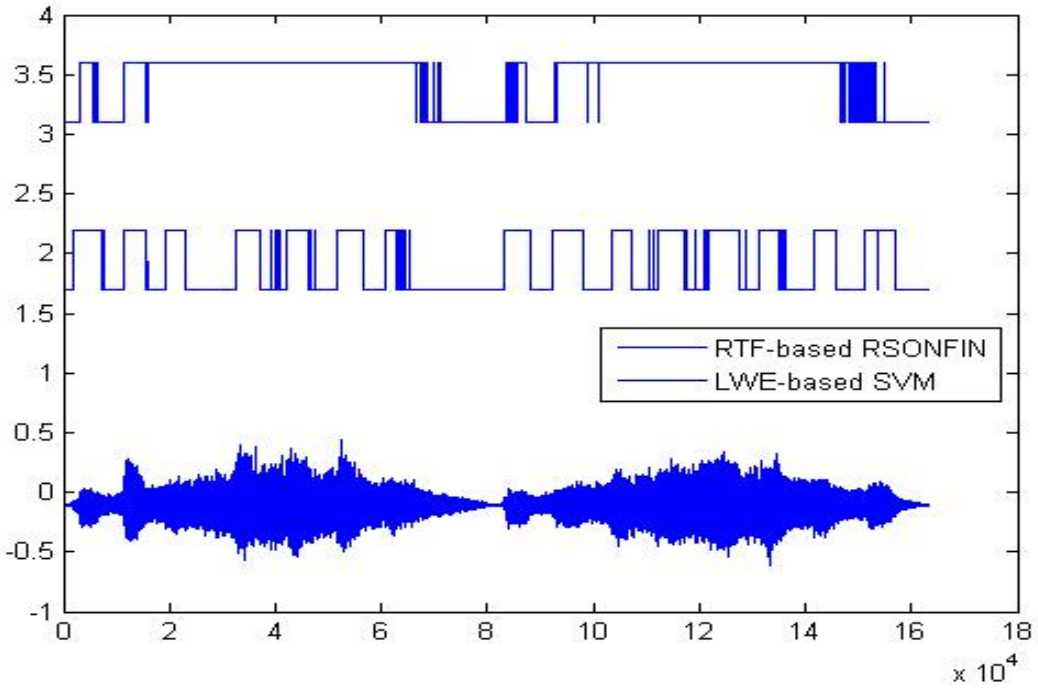


Fig. 6. Word boundary results by LWE-based SVM and RTF-based RSONFIN in variable noise level environment.

LWE-based SVM is shown, and Fig.4 (b) shows test of LWE-based SVM.

The classification rate defined in Eq. (9) is used as training performance index.

$$\text{Classification rate} = \frac{\text{Correctly detected frame number}}{\text{total frame number in training sequance}} \tag{9}$$

For SVM, the value of C influences the training performance. Fig. 5 shows the training performance of C in the range [1, 85], where the range is spaced to 17 equal scales. The cost value C is set to 40 in the following experiments, where there are a total of 1050

(a) (b)

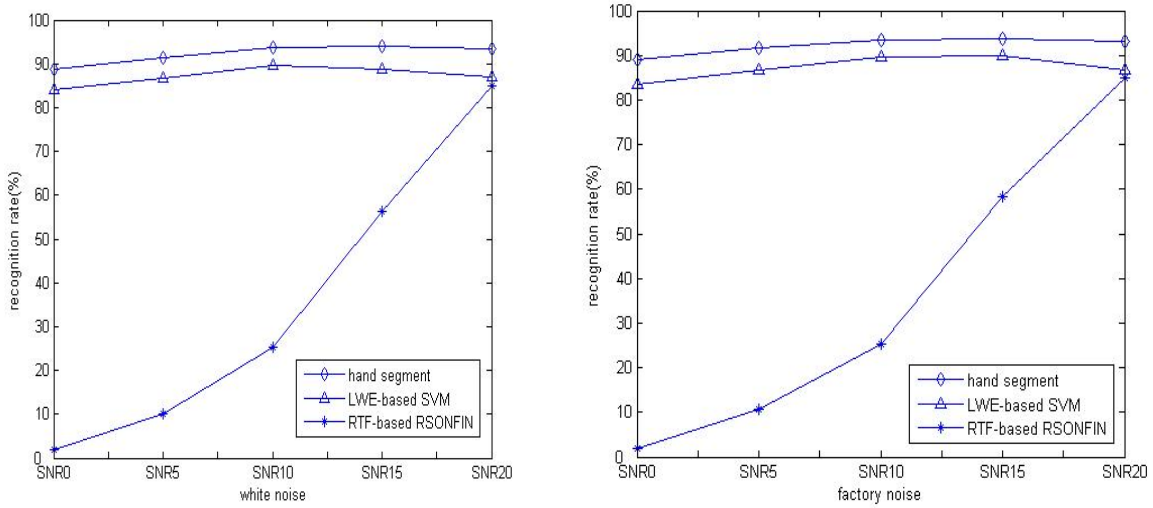


Fig. 7. Noisy speech recognition results by different word boundary detection methods. (a) white noise (b) factory noise.

SVs in the trained SVM.

To get a quick view on the test performance, some illustrative examples are experimented and shown in Fig. 6, where white noise with sharp variation in amplitude is added to the clean speech. Most word boundaries are correctly detected. For comparison, Fig. 6 also shows the performance of refined time-frequency (RTF) feature –based recurrent self-organizing neural fuzzy inference network (RTF-based RSONFIN) [4] detection method. The result shows that RTF-based RSONFIN almost fails to detect most of the words, and the performance of LWE-based SVM shows much better performance than RTF-based RSONFIN.

Next, the ten Mandarin digital words in each sequence of transcriptions in the test database are to be recognized. The words in each sequence are detected by the two methods respectively. When the number of successive frames being detected as speech is larger than 0.1 second, we regard it as word for recognition, otherwise these frames are discarded. So the number of words

detected in each sequence of transcription may be larger or smaller than exact ten words. Considering this phenomenon, we define the following recognition rate

$$\text{recognition rate} = \frac{\mathrm{T}-\mathrm{E}-\mathrm{U}-\mathrm{S}}{\mathrm{T}} \times 100\% , \tag{10}$$

where T is the total number of words in the reference transcriptions, E is the number of words recognized incorrectly, U is un-detect words of reference transcriptions, and S is surplus words of reference transcriptions.

For the recognizer, the hierarchical singleton-type recurrent neural fuzzy network (HSRNFN) [9] that put SNR20 white noise as training data is used. The reason we use HRNFN is that it achieves high recognition rate and is robust to different types of noise under different SNR. With HSRNFN recognizer, the recognition results by hand-segment,

LWE-based SVM, and RTF-based RSONFIN methods under white and factory noise are shown in Fig. 7. The results show that recognition rate of the LWE-based SVM method is slightly lower than that of hand segmentation, but is much larger than that of the RTF-RSONFIN method.

5. CONCLUSIONS

Two research results on robust speech detection in variable noise-level environment have been presented this paper, one is the robust LWE-based parameters, and the other is detector design by SVM. Variable noise-level instead of fixed noise-level is added to each sequence of transcript. Distributions of the LWE-based parameters in the 2-dimensional feature space for different SNRs have shown that the LWE-based parameters are feasible for speech detection over variable level noise. The LWE-based SVM can be applied to a speech recognition system as demonstrated in the experiments.

REFERENCES

[1] M. H. Savoji, A robust algorithm for accurate end-pointing of speech signals, *Speech Communication*, vol. 8. no. 1, 1989, pp. 45-60.

[2] J. Rouat, Y. C. Liu, and D. Morissette, Pitch determination and voiced/unvoiced decision algorithm for noisy speech, *Speech Communication*, vol. 21, no. 3, 1997, pp. 191-207.

[3] J. C. Junqua, B. Mak, and B. Reaves, A robust algorithm for word boundary detection in the presence of noise, *IEEE Trans. Speech and Audio Processing*, vol. 2, 1994, pp. 406-412.

[4] G. D. Wu and C. T. Lin A recurrent neural fuzzy network for word boundary detection in variable noise-level environments, *IEEE Transactions on systems, Man, and cybernetics*, vol. 31, no. 1, 2001, pp. 84-97.

[5] J. F. Wang and S. H. Chen, "A C/V segmentation algorithm for mandarin speech signal based on Wavelet transforms," *Proc. of ICASSP*, vol.1, pp.417-420, March 1999.

[6] Y. Qi and B. R. Hunt, Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier, *IEEE Trans. Speech and Audio Processing*, vol. 1, 1993, pp. 250-255.

[7] C. Cortes, and V. Vapnik, "Support vector networks," *International Journal on Machine Learning*, vol. 20, pp. 1-25, 1995.

[8] N. Cristianini and J. S.-Taylor, *An Introduction to Support Vector Machines And Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[9] Y. T. Chan, Wavelet Basics, 1995, Kluwer Academic Publishers.

[10] C. F. Juang, C. T. Chiou, and C. L. Lai, "Hierarchical singleton-type recurrent neural fuzzy networks for noisy speech recognition," *IEEE Trans. Neural Networks*, vol. 18, no. 3, pp. 833-843, May 2007.

Noise-Robust Speech Features Based on Cepstral Time Coefficients

Ja-Zang Yeh
Department of Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan
ycc97m@cse.nsysu.edu.tw
Chia-Ping Chen
Department of Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan
cpchen@cse.nsysu.edu.tw

Abstract

In this paper, we investigate the noise-robustness of features based on the **cepstral time coefficients** (CTC). By cepstral time coefficients, we mean the coefficients obtained from applying the discrete cosine transform to the commonly used mel-frequency cepstral coefficients (MFCC). Furthermore, we apply temporal filters used for computing delta and acceleration dynamic features to the CTC, resulting in delta and acceleration features in the frequency domain. We experiment with five different variations of such CTC-based features. The evaluation is done on the Aurora 3 noisy digit recognition tasks with four different languages. The results show all but one such feature set performance gain, the other feature sets actually lead to performance gains. The best feature set achieves an improvement of 25% over the baseline feature set of MFCC.

Keywords: **MFCC, CTC, delta, robust feature**

1. Introduction

A front-end of a speech recognition system may consist of several stages for noise-robustness to achieve good performance. In the early stage of spectral domain, well-known methods such as spectral subtraction [1] and Wiener filter [2] may be applied. In the middle stage of cepstral domain, the mel-frequency cepstral coefficients (MFCC) are commonly used as the static feature set. In the post-processing stage, there may be normalization, temporal information integration, and transformation modules.

It has been observed that simple normalization approaches, such as the cepstral mean subtraction (CMS) [3], cepstral variance normalization (CVN) [4], and histogram normalization (HEQ) [5] can lead to significant performance improvement in recognition accuracy in noisy environment. Apparently such methods are capable of alleviating the *mismatch* between the clean and noisy data.

In this paper we investigate novel features based on simple transformation methods. Specifically, we insert a window of static cepstral vectors in a matrix and then apply the *discrete cosine transform* (DCT) along the temporal axis. The coefficents after the DCT is called the cepstral time coefficients,

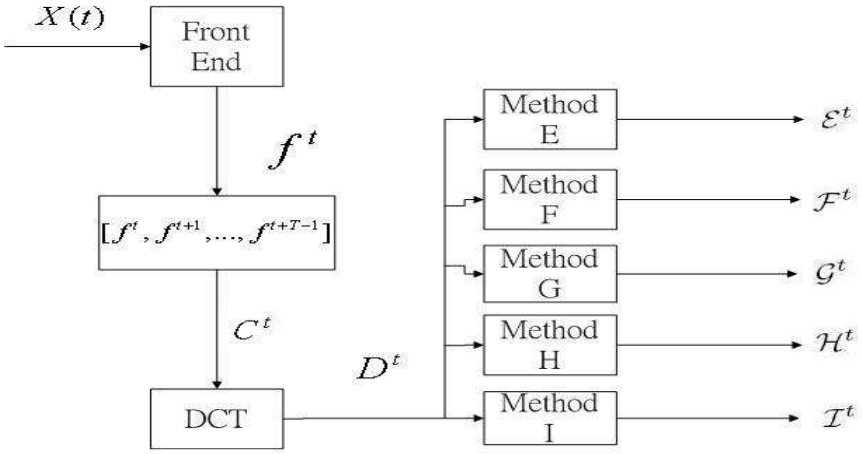


Figure 1: The block diagram of the proposed feature transformation methods.

and the resultant matrix is called the cepstral time matrix (CTM) [6,7]. After CTM for each frame is extracted, we further apply normalization and routines for delta and acceleration feature extraction to the cepstral time coefficients. The transformed features are combined with the static MFCC features to form the final feature vector.

This paper is organized as follows. Section 2 defines the cepstral time matrix and introduces the investigated feature transformations. The experimental setup and recognition results are described in Section 3. In Section 4, we draw conclusions.

2. Feature Transformations

Our feature extraction and transformation process is illustrated in Figure 1. We begin with a review of the cepstral time matrix, which is followed by the mathematical definition of the proposed additive transformation methods.

2.1. Cepstral Time Coefficients

We first insert a fixed number of adjacent feature vectors in a matrix

$$C^t \triangleq \begin{bmatrix} C^t_{11} & C^t_{12} & \dots & C^t_{1T} \\ \vdots & \ddots & & \vdots \\ C^t_{K1} & C^t_{K2} & \dots & C^t_{KT} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{f}^t & \mathbf{f}^{t+1} & \dots & \mathbf{f}^{t+T-1} \end{bmatrix}. \quad (1)$$

Here K is the feature vector dimension, and $\mathbf{f}^t$ is the feature vector of frame t, C^t is the matrix whose column vectors are the T consecutive feature vectors starting from frame t.

The cepstral time matrix at frame t, D^t, is related to C^t by the discrete-cosine transform. Each **row** of D^t is the discrete-cosine transform of the corresponding row of C^t. That is,

$$D^t_{i:} = DCT(C^t_{i:}). \quad (2)$$

Here $D^t_{i:}$ is the i-th row of matrix D.[1] We call D^t_{in} the nth cepstral time coefficient (CTC) of channel i at frame t. D is also called cepstral time matrix (CTM). It represents the spectral information of

cepstral coefficient in an analysis window of frames. [1]Since our matrix index starts from 1 instead of 0, here the DCT needs to be

$$D_{in}^{t} = \sum_{\tau=1}^{T} C_{i\tau}^{t} \cos\left(\frac{(2\tau-1)(n-1)\pi}{2T}\right). \tag{3}$$

2.2. CTC-Based Features

In this paper, we have 5 different transforms applied to CTC, each leading to a different feature vector.

2.2.1. Method E

The first transform is dividing the first column of D^t by the number of frames (T), while leaving other columns unchanged. Let E^t be the new feature matrix, we have

$$\begin{cases} E_{:1}^{t} &= D_{:1}^{t}/T \\ E_{:n}^{t} &= D_{:n}^{t}, \;\; n \neq 1 \end{cases} \tag{4}$$

Note $E_{:1}^{t}$ has a physical meaning. According to (2), it is the mean of the cepstral coefficients within an analysis window (while $D_{:1}^{t}$ is the sum).

We then compute a novel feature set based on E^t. Specifically, we treat the columns in E^t as a temporal sequence and apply the delta and acceleration feature extraction steps. That is,

$$\begin{cases} \breve{E}_{:2}^{t} &= E_{:2}^{t} - E_{:1}^{t} \\ \breve{E}_{:3}^{t} &= E_{:3}^{t} - 2E_{:2}^{t} + E_{:1}^{t}. \end{cases} \tag{5}$$

We add the $\breve{E}_{:2}^{(t)}$ and $\breve{E}_{:3}^{(t)}$ to the static MFCCs, resulting in a feature vector of

$$\mathcal{E}^{t} = \begin{bmatrix} C_{:1}^{t} \\ \breve{E}_{:2}^{t} \\ \breve{E}_{:3}^{t} \end{bmatrix}. \tag{6}$$

2.2.2. Method F

An alternative transform is to normalize the feature values in the first column to the range of $[-1, 1]$. This is achieved by dividing $D_{:1}^{t}$ by the maximum magnitude of the first column. Let F^t be defined by

$$\begin{cases} F_{:1}^{t} &= D_{:1}^{t}/N^{t} \\ F_{:n}^{t} &= D_{:n}^{t}, \;\; n \neq 1 \end{cases} \tag{7}$$

where N^t is the maximum magnitude in the first column, i.e.,

$$N^{t} = \max_{d} |D_{d1}^{t}|.$$

The remaining operations are similar to Method E. That is,

$$\begin{cases} \breve{F}_{:2}^{t} &= F_{:2}^{t} - F_{:1}^{t} \\ \breve{F}_{:3}^{t} &= F_{:3}^{t} - 2F_{:2}^{t} + F_{:1}^{t}. \end{cases} \tag{8}$$

[1]In general, we will use notation $A_{i:}$ to denote the i-th row vector and $A_{:j}$ to denote the j-th column vector, of matrix A.

We add $\breve{F}_{:2}^{(t)}$ and $\breve{F}_{:3}^{(t)}$ to the static MFCCs, resulting in a feature vector of

$$\mathcal{F}^t = \begin{bmatrix} C_{:1}^t \\ \breve{F}_{:2}^t \\ \breve{F}_{:3}^t \end{bmatrix}. \tag{9}$$

2.2.3. Method G

In Method G, we add the first and second columns of CTM, which represents the zeroth and first cepstral time coefficients, to the static MFCC vector,

$$\mathcal{G}^t = \begin{bmatrix} C_{:1}^t \\ D_{:1}^t \\ D_{:2}^t \end{bmatrix}. \tag{10}$$

2.2.4. Method H

In Method H, we add the second and third columns of CTM, which represent the first and second cepstral time coefficients, to the static MFCC vector,

$$\mathcal{H}^t = \begin{bmatrix} C_{:1}^t \\ D_{:2}^t \\ D_{:3}^t \end{bmatrix}. \tag{11}$$

2.2.5. Method I

In Method I, we no longer use the MFCC. Instead, we simply use the zeroth, first, and second cepstral time coefficients,

$$\mathcal{I}^t = \begin{bmatrix} D_{:1}^t \\ D_{:2}^t \\ D_{:3}^t \end{bmatrix}. \tag{12}$$

2.2.6. Method B

For completeness, we describe our baseline features as Method B. Our baseline simply uses the 12 MFCCs ($c_1, \ldots, c_{12}$), the log energy, and the delta and delta-delta features. Therefore, the feature vector has a dimension of 39, which agrees with other methods. Furthermore, our baseline results agree with the Aurora 3 baseline results [8,9].

3. Experiments

3.1. Experimental Database

We evaluate the proposed CTC-based speech features on the Aurora 3 noisy-digit recognition tasks [8,9]. Aurora 3 is a multi-lingual speech database, consisting of digit-string utterances in Danish, German, Finnish and Spanish. It provides a platform for fair comparison between systems of different front-ends. All the results reported in this paper follow the Aurora 3 evaluation guidelines.

3.2. Results

We first evaluate the number of vectors to be included in C^t, and decide to use $T = 15$. For the static features we use 12 MFCC features and the log energy, making $K = 13$. Therefore, the initial matrix C^t is of size 13×15.

Table 1 lists the experimental results on the Aurora 3 database. The entries in the table are the averaged relative improvements of word error rates over the baseline.

Consistent performance across different methods have been observed in the experiments. Specifically, Method H achieves the best performance, while Method G yields the worst performance, in all languages. Given that Method G and Method H differ only in the cepstral time coefficients they include in the final feature vector, it is fair to say that *the zeroth cepstral time coefficient is detrimental to recognition accuracy*.

Methods E, Method F, and Method I yield mixed results. In Finnish, Method E outperforms Method F and Method I. In Spanish and Danish, Method F outperforms Method I and Method E. Method E and Method F are similar in the sense that the first column (zeroth cepstral time coefficients) are normalized, and then used in procedures similar to delta and acceleration feature extraction, in the frequency domain rather than in the time domain. It is not surprising that they have similar performance level.

Table 1: *The overall (averaged over conditions) relative improvements of the word error rates in the Aurora 3 tasks*.

	German	Spanish	Finnish	Danish
E	-12.4	16.2	16.5	16.3
F	-10.5	22.4	10.8	16.3
G	-58.1	-29.0	-42.9	-19.2
H	7.5	26.6	25.4	23.2
I	-10.8	19.8	8.5	13.1

The comparison of Method G and H concludes that the zeroth CTC is detrimental of recognition accuracy. The zeroth CTC corresponds to the first column of CTM. Therefore in Method E and F, we try schemes of normalizing the first column of CTM. In Method E we divide the first column of CTM by T, and in Mthod F we normalize the value of first column to the range -1 to 1. The performance of E and F given in Table 1 are better than the baseline. Lastly, we also try Method I, which uses only CTCs, and excludes MFCCs. Its recognition accuracy is also better than the baseline.

Figure 2 plots the temporal sequences of the fifth dimension of the third column (Dimension 31 out of 39) of the feature vectors of Method B, F, and H of a pair of Danish utterances. The pair consists of an utterance of Channel 0 (the cleaner instance) and an utterance of Channel 1 (the noisier instance). Specifically, using our previously defined notations, Figure 2(B) is the plot of $\triangle^2 f_5^t$, Figure 2(F) is the plot of $\breve{F}_{53}^t$, and Figure 2(H) is the plot of $\breve{H}_{53}^t$. It appears that the difference between Channel 0 and Channel 1 is smaller in the cases of (F) and (H) than in the case of (B). Therefore the mismatchedness is reduced.

Table 2 lists the experimental results of Method H on the Aurora 3 database, given as percent word error rate (WER) results. These results include the four Aurora 3.0 languages (Finnish, Spanish, German, and Danish) and the Well-Matched(WM), Medium-Matched(MM), and Highly-Mismatched(HM) training/testing cases.

Table 2: *Our most recent Aurora 3.0 results using the method H, given as percent word error rate (WER) results. These results include the four Aurora 3.0 languages (Finnish, Spanish, German, and Danish) and the Well-Matched(WM), Medium-Matched(MM), and Highly-Mismatched(HM) training/testing cases.*

Aurora3 Reference Word Error Rate				
	German	Spanish	Finnish	Danish
WM	9.4	13.1	9.5	20.4
MM	21.9	26.3	27.5	50.6
HM	25.7	57.8	69.6	66.8

Aurora3 Word Error Rate, Method H				
	German	Spanish	Finnish	Danish
Well	9.1	9.7	7.0	15.4
Mid	19.8	18.4	21.3	39.0
High	21.7	45.4	50.2	52.4

Aurora3 Relative Percentage Improvement					
	German	Spanish	Finnish	Danish	Avg.
Well	4.4	26.0	26.2	24.5	20.3
Mid	5.3	29.9	22.7	22.9	20.2
High	15.5	23.0	27.9	21.5	22.0
overall	7.5	26.6	25.4	23.2	20.7

4. Conclusion and Future Work

In this paper, we use five difference feature sets based on the cepstral time coefficients. Method E and F, which first normalize the first column and then apply the delta and delta-delta operations on the first 3 columns of CTM, lead to performance gains over the baseline. Method G and H, which combine different sets of columns of CTM with the raw MFCC vector, lead to mixed results. Method I, which uses all cepstral time coefficients, leads to improvement. Overall, the combination of raw MFCC and the second and the third columns of CTM yields the best results among all experimented feature sets.

5. References

[1] S. Boll, "Supression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] A. Berstein and I. Shallom, "An hypothesized Wiener filtering approach to noisy speechrecognition," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 1991, pp. 913–916.

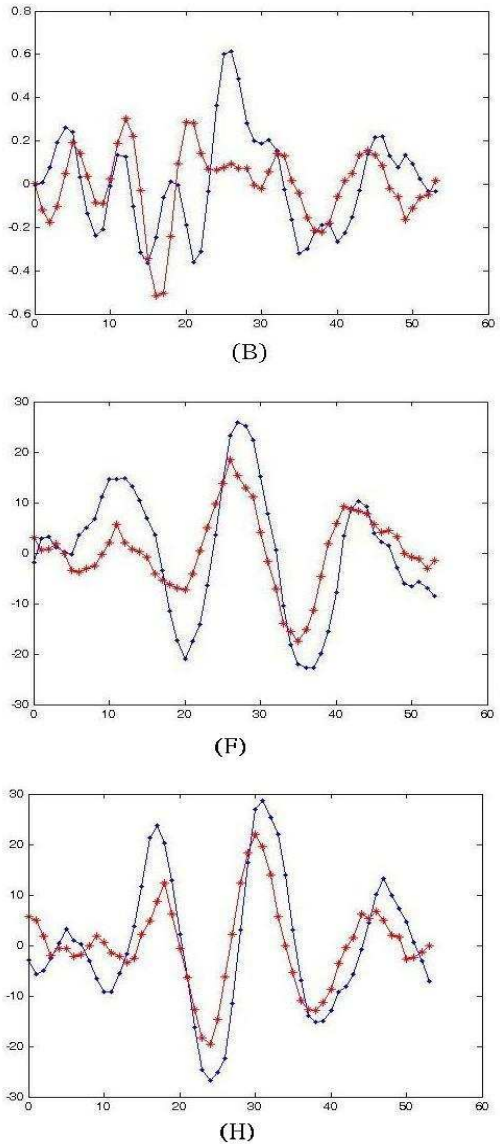


Figure 2: *Plot of Dimension* 31 *(out of* 39*) of a Danish utterance recorded in two mismatched channels. (B) is the* $\triangle^2 f_5^t$*, (F) is* $\breve{F}_{53}^t$*, and (H) is* $\breve{H}_{53}^t$*. The horizontal axis is the frmae index and the vertical axis is the feature value. The dotted line ('.') represents Channel 0 and the starred line ('*') represents Channel 1.*

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[4] O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speechrecognition in noise," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, 1998.

[5] A. de La Torre, A. Peinado, J. Segura, J. Perez-Cordoba, M. Benitez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[6] B. Milner, "Inclusion of temporal information into features for speechrecognition," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 1, 1996.

[7] ——, "A comparison of front-end configurations for robust speechrecognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02)*, vol. 1, 2002.

[8] Motorola Au/374/01, "Small vocabulary evaluation: Baseline mel-cepstrum performances with speech endpoints," October 2001.

[9] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speechdat-car: A large speech database for automotive environments," in *Proceedings of the II LREC Conference*, vol. 1, no. 2, 2000.

強健性語音辨識中分頻段調變頻譜補償之研究
A Study of Sub-band Modulation Spectrum Compensation for Robust Speech Recognition

黃勝源 Sheng-yuan Huang
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s96323530@ncnu.edu.tw

杜文祥 Wen-hsiang Tu
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
aero3016@ms45.hinet.net

洪志偉 Jeih-weih Hung
國立暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
jwhung@ncnu.edu.tw

摘要

雖然語音科技進步迅速，但自動語音辨識仍是一門值得繼續研究開發的課題。因爲目前多數的語音辨識系統應用於不受干擾的安靜環境，雖然能得到相當滿意的辨識效果，但若將其應用於實際的環境中，語音訊號往往會因爲環境雜訊的影響，導致辨識效能有明顯地衰減，發展多年的強健性技術即是針對此項缺點作改進。

在諸多強健性技術中，有一類方法爲對語音特徵作統計上的正規化，傳統上，這些方法都是對全頻段的語音特徵時間序列做正規化處理，然而，在分析此類方法的效能上，通常是以其調變頻譜的正規化程度作爲效能的依據，因此，如果直接在語音特徵之調變頻譜上作正規化，應亦可達到不錯的效果。另外，由於不同頻率的調變頻率成份具有不相等的重要性，但是傳統之特徵時間序列正規化法相對忽略了此性質，基於這些觀察，在本論文中，我們提出了一系列的分頻段調變頻譜統計正規化法，此類方法可以分別正規化不同頻段的統計特性，進而提升語音特徵在雜訊環境下的強健性能；在國際通用的 Aurora-2 連續數字資料庫之語音辨識上，我們所提出的新方法相對於基礎實驗的辨識率而言，可以達到高達 65%的相對錯誤降低率，而這些新的調變頻譜正規化法相對於時間序列正規化法而言，於相對錯誤降低率上也有 7%至 32%的進步空間，此足以驗證這些新方法能夠更有效地提昇語音辨識系統在雜訊環境下的辨識效能。

關鍵詞：語音辨識、調變頻譜、統計正規化、強健性語音特徵參數

Abstract

In this paper, we propose a novel scheme in performing feature statistics normalization techniques for robust speech recognition. In the proposed approach, the processed temporal-domain feature sequence is first converted into the modulation spectral domain. The magnitude part of the modulation spectrum is decomposed into non-uniform sub-band

segments, and then each sub-band segment is individually processed by the well-known normalization methods, like mean normalization (MN), mean and variance normalization (MVN) and histogram equalization (HEQ). Finally, we reconstruct the feature stream with all the modified sub-band magnitude spectral segments and the original phase spectrum using the inverse DFT. With this process, the components that correspond to more important modulation spectral bands in the feature sequence can be processed separately. For the Aurora-2 clean-condition training task, the new proposed sub-band spectral MN, MVN and HEQ provide relative error rate reductions of 18.66% and 23.58% over the conventional temporal MVN and HEQ, respectively.

一、簡介

雖然語音科技進步迅速，但自動語音辨識(automatic speech recognition, ASR)[1]仍是一門值得繼續研究開發的課題。目前多數的語音辨識系統若在不受干擾的安靜環境下，一般而言皆能得到相當滿意的辨識效果，然而若將其應用於實際的生活環境中，辨識效能便會有所衰減，主要是實際生活環境中有許多的變異性(variation)影響辨識效能，其中影響語音辨識的變異性有訓練環境與測試環境之間的環境不匹配(environmental mismatch)、語者變異性(speaker variation)及發音的變異性(pronunciation variation)等因素，這些因素都會明顯影響語音辨識系統的效能。因此在近幾十年來，持續不斷有許多學者研究努力改善上述幾類的語音變異性，進而使語音辨識系統能更有效地運用於真實的生活環境中。

針對環境不匹配所發展的許多強健性方法，大致上包含了特徵補償與模型補償兩大類型，而特徵補償方法中其中有一類別的方向是針對語音辨識所用的特徵參數之統計量作正規化處理，這些處理通常是作在特徵之時間序列域(temporal domain)上，例如倒頻譜平均值正規化法(cepstral mean normalization, CMN)[2]、倒頻譜平均值與變異數正規化法(cepstral mean and variance normalization, CMVN)[3]與統計圖等化法(histogram equalization, HEQ)[4]等。

以上各種方法主要是執行在語音特徵的時間序列域上，但在其效能的分析上，我們通常會去探討雜訊及通道效應對於原始特徵之調變頻譜的失真，及這些方法對於此失真的改善程度，因此近年來，開始有學者提出直接於特徵之調變頻譜域上使用特徵統計正規化法，如調變頻譜統計圖等化法(spectrum histogram equalization, SHE)[5]，此方法是針對調變頻譜的強度頻譜之機率分佈(probability distribution)作正規化處理，驗證了直接針對調變頻譜的強度成份作機率分佈的正規化的確帶來了明顯的特徵強健性效果。但是以上各種技術，皆是直接或間接將語音特徵序列的全調變頻帶資訊作整體的處理，並未對各不同的頻帶有不同的考慮。然而，根據許多的研究[6][7]證實，對語音辨識而言，不同頻率的調變頻譜成份具有不相等的重要性；在文獻[8]中更明確地提到，調變頻譜的偏低頻率成份資訊對於語音辨識有較大的助益，其中又以 1~16 Hz 之調變頻帶範圍的成份最爲重要。藉由以上之各觀點，在本論文中，我們提出了基於強度頻譜之**分頻段**調變頻譜統計正規化法，一方面希望如 SHE 法，直接對於語音特徵序列之調變頻譜作正規化處理，另一方面，則是希望在新方法中能異於過去之全調變頻帶之資訊一併處理的方式，將調變頻帶作一系列的頻段切割，在每個子頻段中加以正規化其調變頻譜，進而更有效地凸顯正規化的效能；在後面章節之一系列的實驗中，我們將呈現所提出之新方法確實可以更有效地提昇語音特徵在雜訊環境的強健性，達到我們以上所提的目的。

本論文其他章節概要如下：在第二章中，我們介紹本論文所提出之分頻段的統計正規化法其背景、原理及其相關的步驟說明。第三章將呈現並討論一系列分頻段調變頻譜統計正規化法的實驗結果，並與其他時間序列域上的強健性技術結合，對此類結合方式

的辨識實驗加以探討與分析，以驗證此類結合方式是否具有良好的加成性。而在第四章裡，則爲一簡要的結論與未來展望。

二、基於強度頻譜之分頻段調變頻譜統計正規化法

在這一章中，我們將對所新提出的分頻段調變頻譜統計正規化法之背景與步驟作詳細的說明，並且將以一段受雜訊干擾的語句爲例，驗證這些新方法在降低雜訊干擾的效能，及與其他相類似方法的初步比較。

(一) 分頻段調變頻譜統計正規化法

在本論文所提的新方法中，我們嘗試將調變頻譜中的強度頻譜(magnitude spectrum)切割成許多子頻段，再分別對各自子頻段的統計值作正規化處理；我們所用的正規化演算法，包括了除了文獻[5]之 SHE 技術所用的統計圖等化法(HEQ)外，也額外使用了較簡易執行的平均值正規化法(MN)與平均值與變異數正規化法(MVN)，以期它們相較於傳統全頻帶式的正規化法而言，能帶來更明顯的效能，或是能有效減低執行的複雜度。我們所提的分頻段調變頻譜正規化法的詳細步驟分列於下：

1. 假設一段語音之梅爾倒頻譜特徵參數序列以下式(2-1)表示：

$$\left\{x^{(m)}[n];1\le n\le N\right\},\quad 1\le m\le M, \qquad \text{式(2-1)}$$

其中 M 爲一語音特徵向量中特徵個數， N 表示爲此單一語句的音框總數。每個特徵序列 $\left\{x^{(m)}[n]\right\}$ 經正規化處理後，以 $\left\{\tilde{x}^{(m)}[n]\right\}$ 表示，我們希望新的特徵序列 $\left\{\tilde{x}^{(m)}[n]\right\}$ 相對於原始特徵序列而言，更具有強健性，使辨識效果有明顯地提升。在之後的敘述，爲了精簡符號的標示，我們省略了上標"(m)"符號。

2. 將特徵序列 $\{x[n];1\le n\le N\}$ 經 N 點離散傅立葉轉換(discrete Fourier transform, DFT)後得到其調變頻譜 $\{X[k]\}$ ，如下式。

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi nk}{K}},\quad 0\le k\le\left\lceil\frac{N}{2}\right\rceil \qquad \text{式(2-2)}$$

假設 $\{x[n]\}$ 的音框取樣頻率(frame rate)爲 F_s Hz，則在其調變頻譜域上 $\{X[k]\}$ 的頻率範圍爲 $\left[0,\frac{F_s}{2}\right]$ ；而由於 $X[k]$ 爲一複數，我們以極座標(polar form)表示 $X[k]$ 如下式：

$$X[k]=A[k]e^{j\theta_k} \qquad \text{式(2-3)}$$

其中 $A[k]$ 是 $X[k]$ 的強度成份， $\theta[k]$ 是 $X[k]$ 的相位成份，接下來我們只針對強度成份 $\{A[k]\}$ 作調整，而保留相位成份 $\{\theta[k]\}$ 不變。

3. 將上一步驟調變頻譜的強度成分 $\left\{A[k];0\le k\le\left\lceil\frac{N}{2}\right\rceil\right\}$ 以不等切(non-uniform)且倍頻(octave)的方式，切割成 L 個頻段，每個頻段的範圍如下式(2-4)所示：

$$\begin{cases}\left[0,\dfrac{1}{2^{L-1}}\left(\dfrac{F_s}{2}\right)\right], & \text{if } \ell=1.\\ \left[\dfrac{2^{\ell-2}}{2^{L-1}}\left(\dfrac{F_s}{2}\right),\dfrac{2^{\ell-1}}{2^{L-1}}\left(\dfrac{F_s}{2}\right)\right], & \text{if } \ell=2,3,...,L.\end{cases} \qquad \text{式(2-4)}$$

由上式可以得知，調變頻譜低頻帶的部分被切割成較多個頻段，且每個頻段的長度較

短，相對地，高頻的部分被切割成較少的頻段，且每個頻段的長度較長。在將$\{A[k]\}$作上述的頻段切割後，我們以$\{A_\ell[k']\}$表示其中的第 L 個頻段。此對於頻段不等切的原因，在於我們之前所提，低調變頻帶對於語音辨識較爲重要，理應分較多的頻段來個別處理，而高調變頻帶相對而言較不重要，所以可將較大的頻段範圍一併處理。

4. 我們將上一步驟所得之不同頻段的強度頻譜$\{A_\ell[k']\}$作統計正規化處理。我們使用的正規化法分別爲：平均值正規化法(MN)、平均值與變異數正規化法(MVN)與統計圖等化法(HEQ)，處理後的特徵即以$\{\tilde{A}_\ell[k']\}$表示。詳細地說，平均值正規化法(MN)在此的計算方式以下式(2-5)表示：

$$\tilde{A}_\ell[k'] = A_\ell[k'] - \mu_{\ell,s} + \mu_{\ell,a}, \qquad \text{式(2-5)}$$

其中，$\mu_{\ell,s}$爲單一(single)語句之分頻段強度頻譜的平均值，$\mu_{\ell,a}$爲全部(all)訓練語句之分頻段強度頻譜的平均值。

平均值與變異數正規化法(MVN)在此的計算方式以式(2-6)表示：

$$\tilde{A}_\ell[k'] = \left(\frac{A_\ell[k'] - \mu_{\ell,s}}{\sigma_{\ell,s}}\right)\cdot\sigma_{\ell,a} + \mu_{\ell,a} \qquad \text{式(2-6)}$$

其中，$\mu_{\ell,s}$爲單一語句之分頻段強度頻譜的平均值，$\sigma_{\ell,s}$爲單一語句之分頻段強度頻譜的標準差，$\mu_{\ell,a}$爲全部訓練語句之分頻段強度頻譜的平均值，$\sigma_{\ell,a}$爲全部訓練語句之分頻段強度頻譜的標準差。

統計圖等化法(HEQ)在此的計算方式以式(2-7)表示：

$$\tilde{A}_\ell[k'] = F_{\ell,a}^{-1}\left(F_{\ell,s}\left(A_\ell[k']\right)\right) \qquad \text{式(2-7)}$$

其中$F_{\ell,s}(\bullet)$爲單一語句之分頻段強度頻譜的機率分佈，$F_{\ell,a}(\bullet)$爲全部訓練語句之分頻段強度頻譜的機率分佈。

5. 在處理完每一頻段之後，我們將各頻段的強度頻譜$\{\tilde{A}_\ell[k']\}$照其頻率大小順序重新串接起來，得到新的全頻段強度頻譜$\left\{\tilde{A}[k]; 0 \le k \le \left\lfloor\frac{N}{2}\right\rfloor\right\}$，此即爲統計正規化法處理後的調變頻譜之強度成份，接著將$\{\tilde{A}[k]\}$補回式(2-3)中的原本相位成分$\{\theta[k]\}$，再經逆轉換離散傅立葉轉換(inverse discrete Fourier transform, IDFT)所得新的特徵$\tilde{x}[n]$，如下式(2-8)表示：

$$\tilde{x}[n] = \frac{1}{N}\sum_{k=0}^{N-1}\left(\tilde{A}[k]e^{j\theta[k]}\right)e^{j\frac{2\pi nk}{N}}, \qquad 0 \le n \le N-1. \qquad \text{式(2-8)}$$

由於特徵序列經傅立葉轉換後，具有左右對稱的特性，即$\tilde{A}[k] = \tilde{A}[N-k]$與$\theta[k] = -\theta[N-k]$，因此我們可藉此推得式(2-8)所需用到的$\{\tilde{A}[k]\}$與$\{\theta[k]\}$在$\left\lfloor\frac{N}{2}\right\rfloor < k \le N-1$的每一項。

在步驟 2 中，若我們未對調變頻譜的語音特徵作分頻段處理，即分段數$L=1$，接著在步驟 3 作統計圖等化法(HEQ)的正規化運算，這樣的運算方式相當於[5]中的調變頻譜統計圖等化法(SHE)；爲了之後討論方便起見，我們將上述式(2-5)、式(2-6)、式(2-7)的正規化方法處理，分別命名爲：分頻段調變頻譜平均值正規化法(sub-band spectral mean normalization, SB-SMN)、分頻段調變頻譜平均值與變異數正規化法(sub-band

spectral mean and variance normalization, SB-SMVN)與分頻段調變頻譜統計圖等化法(sub-band spectral histogram equalization, SB-SHE)，而文獻[5]中所用的全頻帶(full-band)之 SHE 技術，我們則以 FB-SHE 來表示。以下將針對這些分頻段正規化法的特點加以討論：

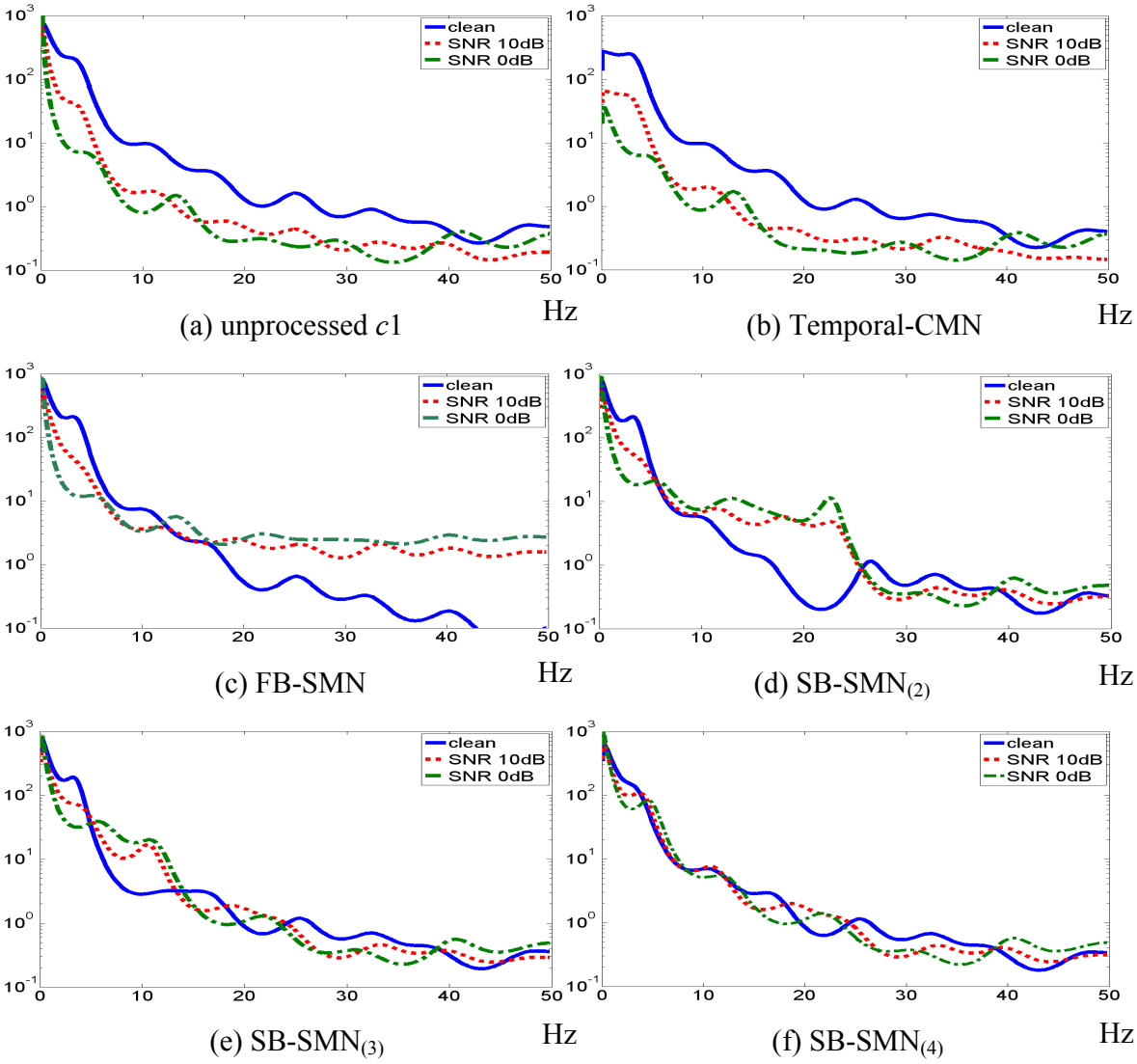
(1) 經由 SB-SMN 與 SB-SMVN 的方法處理之後，所得的調變頻譜強度之部份數值可能爲負值，此明顯違反頻譜強度必然非負的條件，因此當負值的情形出現時，我們將其值重設爲 0。

(2) 在 SB-SMN 與 SB-SMVN 方法中，不同頻段具有各自的目標平均值或目標變異數，同樣地，在 SB-SHE 中，不同頻段使用不同的目標機率分佈作正規化運算。這樣的作法，可以保留不同頻段的頻譜強度之間差異性。

(3) 在這些分頻段調變頻譜正規化法中，全頻段的長度等於各分頻段之長度的和，所以增加分頻段的數目並不會明顯增加運算上的複雜度。然而子頻段的數目不能過多，否則在低頻的子頻段的 $A[k]$ 項數將過少甚至爲零，如此明顯會影響單一頻段所求取之統計值（如平均值、變異數與機率分佈等）的精確性。舉例說明，假設單一語音特徵序列的總點數爲 N 點時，由於此（實數）特徵序列經 N 點傅立葉轉換後，其頻譜的強度成份具有左右對稱的特性，因此實際使用的頻譜點數爲總數的一半，即爲 $\left\lfloor \frac{N}{2} \right\rfloor$ 點，如果我們以不等切的方式切割整個頻帶，所得的頻段不能無限制地增多；例如當我們切 L 個子頻段時，所得的每個子頻段點數由多到少分別爲：$\left\lfloor \frac{N}{4} \right\rfloor, \left\lfloor \frac{N}{8} \right\rfloor, \ldots, \left\lfloor \frac{N}{2^{(L+1)}} \right\rfloor$，所以爲了滿足最少的那一個子頻段的點數不爲零，即每個頻段的資料量至少有一點，我們須滿足 $N \geq 2^{(L+1)}$ 的條件，由此推知，若 $N = 60$，最多只能切 5 個頻段，而若 $N = 30$，則最多只能切 4 個子頻段，若進一步要求若要求每個子頻段點數不能太少，則子頻段數目限制將會更嚴格。

(二) 分頻段調變頻譜正規化法其初步效能的討論：

在這裡，我們將探討本章所提出的三種分頻段調變頻譜正規化法在特徵序列之功率頻譜密度(power spectral density, PSD)降低失真的效果，同時，把這些方法呈現的結果，和第二章所介紹之特徵時間序列域(temporal domain)之正規化技術：倒頻譜平均值正規化法(CMN)、倒頻譜平均值與變異數正規化法(CMVN)、統計圖等化法(HEQ)分別作比較。我們利用了 AURORA-2 資料庫[9]裡 MIP_28826Z4A 語音檔，加入不同訊雜比(SNR)的人聲(babble)雜訊，再經各種正規化法加以處理，最後求取其功率頻譜密度。首先，圖一爲一系列之平均值正規化法（MN）作用於第一維倒頻譜特徵(the first cepstral coefficient, c_1)序列所得之功率頻譜密度圖。在圖一中，藉由圖(a)我們發現，雜訊的存在使乾淨語音與雜訊語音產生明顯的 PSD 失真，而圖(b)中所用 CMN 法，即時域型 MN 法(temporal CMN)，可稍微降低此失真，而頻域型的 MN 法中，由圖(c)至圖(f)發現,將全頻帶逐漸細分至 2 到 4 個子頻段（分別以 SB-SMN$_{(2)}$, SB-SMN$_{(3)}$與 SB-SMN$_{(4)}$表示，下標括號中的數字表示分頻段的個數），此 PSD 失真逐漸降低，其中以圖(f)經 SB-SMN$_{(4)}$ 處理後，PSD 的失真程度最小。由此說明分頻段調變頻譜平均值正規化法對於降低因雜訊所造成的 PSD 失真有明顯的幫助。



(a) unprocessed $c1$ (b) Temporal-CMN

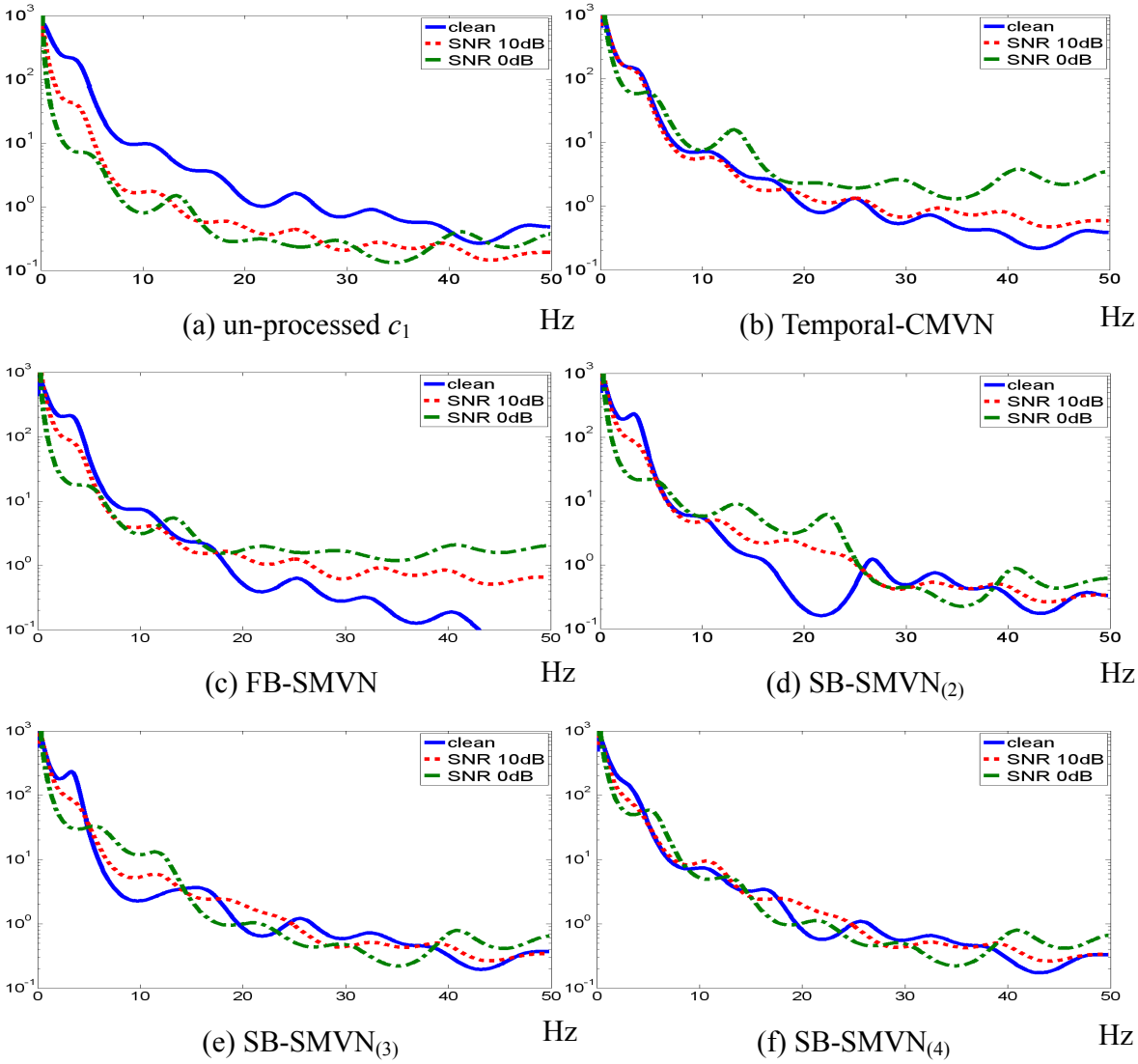
(c) FB-SMN (d) SB-SMN$_{(2)}$

(e) SB-SMN$_{(3)}$ (f) SB-SMN$_{(4)}$

圖一 平均值正規化法作用於不同訊雜比下語音之原始 $c1$ 特徵序列，其調變頻譜曲線圖：(a)原始 $c1$ 特徵序列，(b)時域型 MN 法—CMN，(c)頻域型之全頻帶 MN 法—FB-SMN，(d)頻域型之分頻段 MN 法—SB-SMN$_{(2)}$，(e)頻域型之分頻段 MN 法—SB-SMN$_{(3)}$，(f)頻域型之分頻段 MN 法—SB-SMN$_{(4)}$

接著，圖二爲一系列之平均值與變異數正規化法（MVN）作用於第一維倒頻譜特徵(the first cepstral coefficient, c_1)序列所得之功率頻譜密度圖。在圖二中，藉由圖(b)我們發現，傳統的 CMVN 法，即時域型 MVN 法(temporal CMVN)，相較於圖一的圖(b)之 CMN 而言，降低 PSD 失真的效應更好，意味了額外處理特徵的變異數確實是有幫助的。而在各種頻域型的 MVN 法中，由圖(c)至圖(f)發現，類似 MN 法的效果，當我們將全頻帶逐漸細分至 2 到 4 個子頻段（分別以 SB-SMVN$_{(2)}$, SB-SMVN$_{(3)}$與 SB-SMVN$_{(4)}$表示，下標括號中的數字表示分頻段的個數），PSD 失真也逐漸降低，其中以圖(f)經 SB-SMVN$_{(4)}$ 處理後，對於 PSD 的失真的降低效果最好。由此亦說明了分頻段調變頻譜平均值與變異數正規化法對於降低因雜訊所造成的 PSD 失真也有明顯幫助，同時將圖二與圖一比

較後，可以明顯看出 MVN 法在降低 PSD 失真的效能上優於 MN 法，此吻合我們一般對這兩類方法之效能的認知。



(a) un-processed c_1 (b) Temporal-CMVN

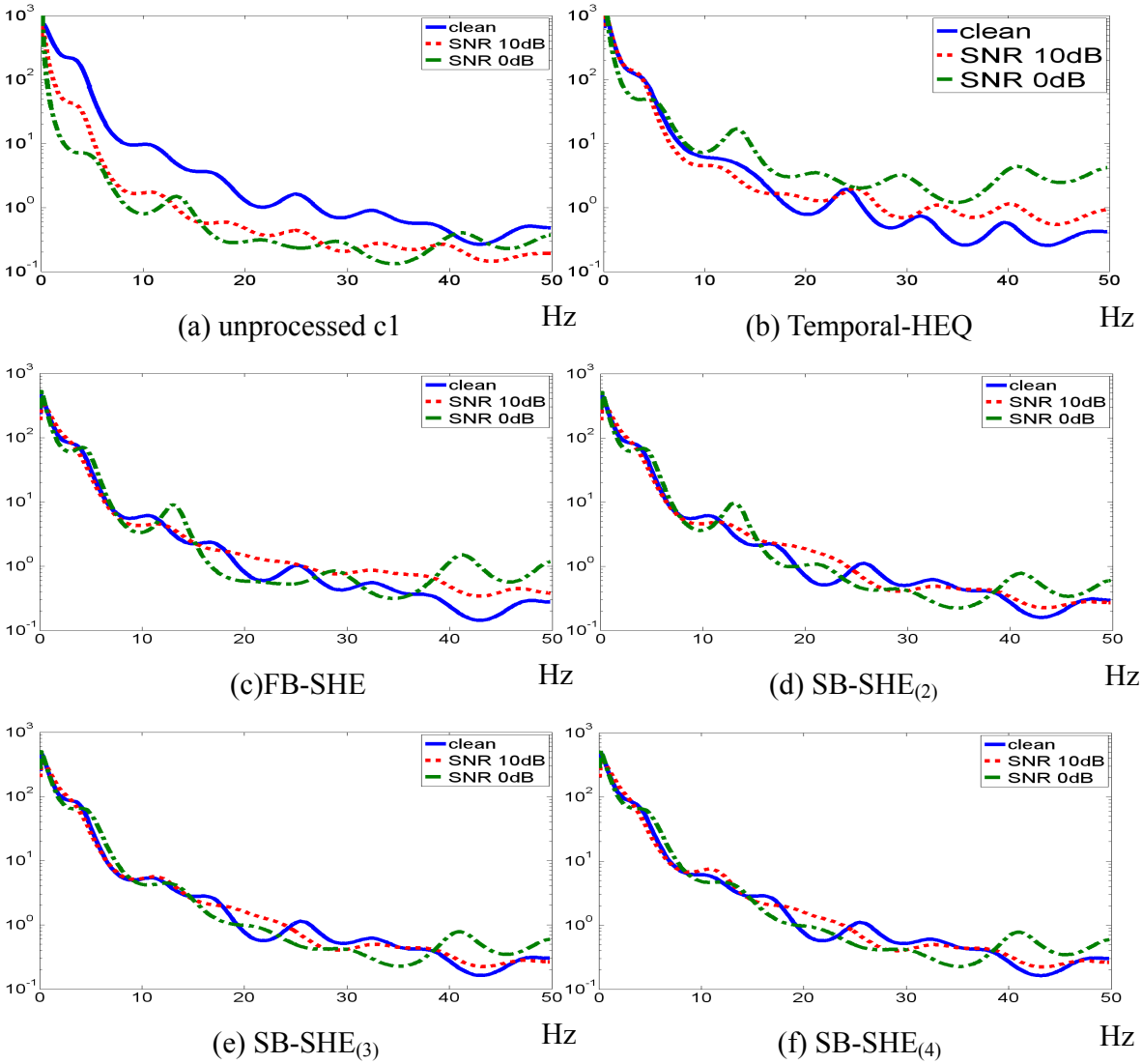
(c) FB-SMVN (d) SB-SMVN$_{(2)}$

(e) SB-SMVN$_{(3)}$ (f) SB-SMVN$_{(4)}$

圖二 平均值與變異數正規化法作用於不同訊雜比下語音之原始$c1$特徵序列，其調變頻譜曲線圖：(a)原始$c1$特徵序列，(b)時域型 MVN 法—CMVN，(c)頻域型之全頻帶 MVN 法—FB-SMVN， (d)頻域型之分頻段 MVN 法—SB-SMVN$_{(2)}$，(e)頻域型之分頻段 MVN 法—SB-SMVN$_{(3)}$，(f)頻域型之分頻段 MVN 法—SB-SMVN$_{(4)}$

最後，圖三爲一系列之統計圖等化法（HEQ）作用於第一維倒頻譜特徵(the first cepstral coefficient, $c1$)序列所得之功率頻譜密度圖。我們將圖三與圖一和圖二比較，可明顯看出正規化整個機率分佈的 HEQ 法，明顯在降低 PSD 的失真上優於只正規化平均值的 MN 法與正規化平均值與變異數的 MVN 法（無論是時域型或頻域型的皆是如此），此外，我們若比較三種全頻式的方法(圖一(c)的 FB-SMN, 圖二(c)的 FB-SMVN 與圖三(c)FB-SHE)，可發現 FB-SHE 相對於 FB-SMN 與 FB-SMVN 而言，從低頻到高頻的 PSD 失真都有明顯降低，而不是像 FB-SMN 與 FB-SMVN 相對只有減少低頻成分的 PSD 失

真，此現象也間接驗證了文獻[5]所提之 FB-SHE 的良好效能。然而在我們所提出的各種分頻段 SHE(SB-SHE)法中，明顯看出它們皆比 FB-SHE 在減低 PSD 失真的效能來得好，由圖(c)至圖(f)發現，類似之前 MN 與 MVN 法的效果，當我們將頻段從全頻段逐漸細分至 2 到 4 個頻段（分別以 SB-HEQ$_{(2)}$, SB- HEQ $_{(3)}$與 SB- HEQ $_{(4)}$表示，下標括號中的數字表示分頻段的個數），PSD 失真逐漸降低，其中以圖(f)經 SB-HEQ$_{(4)}$處理後，對於 PSD 的失真的降低效果最好。由此明顯說明了分頻段調變頻譜統計圖等化法足以有效降低因雜訊所造成的 PSD 失真，在下一章的辨識實驗中，我們將更明顯地看出這些新方法對於提昇語音辨識精確度的效能。



(a) unprocessed c1　(b) Temporal-HEQ

(c)FB-SHE　(d) SB-SHE$_{(2)}$

(e) SB-SHE$_{(3)}$　(f) SB-SHE$_{(4)}$

圖三 統計圖等化法作用於不同訊雜比下語音之原始$c1$特徵序列，其調變頻譜曲線圖：(a)原始$c1$特徵序列，(b)時域型 HEQ 法—Temporal HEQ，(c)頻域型之全頻帶 HEQ 法—FB-HEQ，(d)頻域型之分頻段 HEQ 法—SB-HEQ$_{(2)}$，(e)頻域型之分頻段 HEQ 法—SB-HEQ$_{(3)}$，(f)頻域型之分頻段 HEQ 法—SB-HEQ$_{(4)}$

三、分頻段調變頻譜統計正規化法之實驗結果及分析討論

(一) 語音資料庫簡介

本論文使用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行的 Aurora-2 語音資料庫[9]，它是一套以人工方式錄製的連續英文數字字串，語者由美國成年男女所組成，加上八種來源不同的雜訊，分別爲：地下鐵、人聲、汽車、展覽館、餐廳、街道、機場、火車站等，並以不同程度的訊雜比(signal-to-noise ratio, SNR)加入雜訊，分別爲：clean、20 dB、15 dB、10 dB、5 dB、0 dB 與-5 dB；其通道效應分別爲 G.712 與 MIRS，其爲國際電信聯盟(International Telecommunication Union, ITU)[10]所訂立的兩個通道標準。

(二) 語音特徵參數設定及聲學模型

本論文之相關語音辨識實驗所使用特徵參數爲梅爾倒頻譜係數(MFCC)，附加上其一階差量與二階差量，其詳細語音特徵參數設定分別在表一表示。

取樣頻率	8 kHz
音框長度(frame size)	25 ms, 200 點
音框平移(frame shift)	10 ms, 80 點
預強調濾波器	$1-0.97z^{-1}$
視窗形式(window)	漢明窗(Hamming window)
傅立葉轉換點數(T)	256 點
濾波器組(filter bank)	梅爾刻度三角濾波器組，共 23 個三角濾波器
特徵向量(feature vector)	MFCC 13 維($c1$~$c12$, log-energy) + △MFCC 13 維+△△MFCC 13 維，共 39 維

表一 本論文所使用語音特徵參數設定

我們是以隱藏式馬可夫模型(hidden Markov model, HMM)[11]作爲聲學模型(acoustic models)的型式。包含11個數字模型(zero, one, two, …, nine 及oh)以及靜音(silence)模型，每個數字模型包含16個狀態，各狀態包含20個高斯密度混合。

(三) 語音辨識實驗結果

在這一節中，我們將各種調變頻譜正規化法之實驗結果綜合整理成表二，其中絕對錯誤降低率(absolute error rate reduction, AR)與相對錯誤降低率 1 (relative error rate reduction 1, RR_1)分別爲新辨識率與基礎實驗辨識率(baseline)比較下，所得到的絕對改善率與相對改善率，相對錯誤改善率 2 (relative error rate reduction 2, RR_2)，它是分頻段技術相較於全頻段技術而言所得到的相對錯誤改善率，其計算方式分別由式(3-1)、式(3-2)、式(3-3)所示：

$$AR(\%)=\left(\text{新辨識率}-\text{基礎實驗辨識率}\right)\times 100\% \qquad \text{式(3-1)}$$

$$RR_1(\%)=\left(\frac{\text{新辨識率}-\text{基礎實驗辨識率}}{100\%-\text{基礎實驗辨識率}}\right)\times 100\% \qquad \text{式(3-2)}$$

$$RR_2(\%)=\left(\frac{\text{分頻段法辨識率}-\text{全頻段法辨識率}}{100\%-\text{全頻段法辨識率}}\right)\times 100\% \qquad \text{式(3-3)}$$

由表二觀察中，我們可以得到以下幾點結果：

1. 我們所新提出之各種分頻段調變頻譜正規化法相較於基本實驗而言，皆能使辨識率明顯提升，從 RR_1 的數據看出，它們至少能有 19.00%的相對錯誤降低率；其中 SB-SHE 法的辨識效果比 SB-SMN 法及 SB-SMVN 法更優越，可能之原因如我們預期的，SB-SMN 法及 SB-SMVN 法只對一階動差或一階及二階動差作正規化，而 SB-SHE 法能同時對更高階的動差作正規化處理，使得 SB-SHE 法有較優異的表現。

2. 由於調變頻譜中低頻部分(1~16Hz)佔有較多重要的語音成份，所以我們著重於將低頻部分切割開來分別作正規化處理，從表二可以清楚發現，當低頻部份切割越細，能有效提升語音辨識效能，而三種分頻段調變頻譜補償技術皆以分割四個頻段的效果最爲優越。相對於全頻段式的方法而言，分頻段式的方法其相對錯誤改善率(RR_2)爲：$SB\text{-}SMN_{(4)}$ 的 8.31%，$SB\text{-}SMVN_{(4)}$的 32.64%，$SB\text{-}SHE_{(4)}$的 7.56%。

Method	Set A	Set B	Set C	average	AR	RR_1	RR_2
Baseline	71.98	67.79	78.28	71.56	—	—	—
FB-SMN	77.43	76.26	78.05	77.08	5.52	19.41	—
$SB\text{-}SMN_{(2)}$	77.87	77.26	78.36	77.72	6.16	21.66	2.79
$SB\text{-}SMN_{(3)}$	78.21	76.37	80.82	77.99	6.43	22.61	3.97
$SB\text{-}SMN_{(4)}$	79.12	77.26	82.20	**78.99**	**7.43**	**26.13**	**8.31**
FB-SMVN	79.03	81.19	78.29	79.75	8.19	28.80	—
$SB\text{-}SMVN_{(2)}$	80.06	81.97	79.28	80.67	9.11	32.03	4.54
$SB\text{-}SMVN_{(3)}$	80.84	82.59	80.89	81.55	9.99	35.13	8.89
$SB\text{-}SMVN_{(4)}$	85.94	87.06	85.79	**86.36**	**14.80**	**52.04**	**32.64**
FB-SHE	89.71	90.03	88.27	89.55	17.99	63.26	—
$SB\text{-}SHE_{(2)}$	89.76	90.09	88.40	89.62	18.06	63.50	0.67
$SB\text{-}SHE_{(3)}$	90.13	90.47	88.68	89.98	18.42	64.77	4.11
$SB\text{-}SHE_{(4)}$	90.59	90.69	89.13	**90.34**	**18.78**	**66.03**	**7.56**

表二 調變頻譜統計正規化法之實驗辨識率(%)綜合比較表

(四) 調變頻譜正規化法結合時域型特徵正規化法之實驗結果

在本節中，我們先將原始 MFCC 特徵經各式時域型特徵統計正規化法處理後，再作調變頻譜統計正規化法的處理。在以下各項將呈現並討論各式調變頻譜正規化結合時域型特徵正規化法之實驗結果。

1. 調變頻譜平均值正規化法結合時域型特徵統計正規化法之實驗結果

實驗結果討論：

(1) 表三中，調變頻譜平均值正規化法結合時域型特徵統計正規化法，與其中單一特徵正規化法比較，幾乎皆能有效提升語音辨識效能。舉例而言：$SB\text{-}SMN_{(4)}$結合 CMN 的辨識率爲 88.02%，比起 CMN 的辨識率 81.66%與 $SB\text{-}SMN_{(L=4)}$的辨識率 78.99%，都有相當明顯的改善。惟獨在 $SB\text{-}SMN_{(2)}$結合 CMN 情況下，無法進一步提升辨識率，這可能是該結合方式的分頻段之平均值無法有效逼近訓練語句之分頻段強度頻譜的平均

值，導致辨識率明顯下降。

(2) 從表三也可以清楚發現，當低頻部份切割越細，能有效提升語音辨識效能，而 SMN 法結合其他特徵正規化法皆以分割四個頻段的效果最爲優越。其中 SB-SMN$_{(4)}$結合 CMN 的 RR$_2$ 爲 30.02%，SB-SMN$_{(4)}$結合 CMVN 的 RR$_2$ 爲 15.95%，SB-SMN$_{(4)}$結合 MVA 的 RR$_2$ 爲 12.70%，SB-SMN$_{(4)}$結合 HEQ 的 RR$_2$ 爲 4.98%。

(3) SB-SMN$_{(L=4)}$分別與 CMVN、MVA 及 HEQ 結合，使辨識率幾乎達到 90.00%；而此代表我們用相對簡單的一階統計正規化法(SB-SMN)結合 CMVN、MVA 及 HEQ，即可達到十分突出的效果。

Method		Set A	Set B	Set C	average	AR	RR$_1$	RR$_2$
Baseline		71.98	67.79	78.28	71.56	—	—	—
CMN		80.69	83.41	80.09	81.66	—	—	—
CMN	FB-SMN	82.34	84.06	81.61	82.88	1.22	6.65	—
	SB-SMN$_{(2)}$	80.89	82.22	80.24	81.29	-0.37	-2.02	-9.29
	SB-SMN$_{(3)}$	83.67	84.63	82.69	83.86	2.20	12.00	5.72
	SB-SMN$_{(4)}$	88.09	88.64	86.63	**88.02**	**6.36**	**34.68**	**30.02**
CMVN		83.55	83.75	81.57	83.23	—	—	—
CMVN	FB-SMN	87.81	88.18	86.27	87.65	4.42	26.36	—
	SB-SMN$_{(2)}$	89.08	89.39	87.39	88.87	5.64	33.63	9.88
	SB-SMN$_{(3)}$	89.63	89.97	88.37	89.51	6.28	37.45	15.06
	SB-SMN$_{(4)}$	89.86	90.09	88.20	**89.62**	**6.39**	**38.10**	**15.95**
MVA		86.69	86.89	84.98	86.43	—	—	—
MVA	FB-SMN	88.89	89.19	87.54	88.74	2.31	17.02	—
	SB-SMN$_{(2)}$	89.87	90.17	88.77	89.77	3.34	24.61	9.15
	SB-SMN$_{(3)}$	90.08	90.51	88.94	90.02	3.59	26.46	11.37
	SB-SMN$_{(=4)}$	90.36	90.59	88.94	**90.17**	**3.74**	**27.56**	**12.70**
HEQ		86.90	87.73	87.56	87.36	—	—	—
HEQ	FB-SMN	89.10	89.70	89.20	89.36	2.00	15.82	—
	SB-SMN$_{(L=2)}$	89.20	89.81	89.30	89.46	2.10	16.61	0.94
	SB-SMN$_{(L=3)}$	89.15	89.89	89.35	89.48	2.12	16.77	1.13
	SB-SMN$_{(L=4)}$	89.54	90.28	89.82	**89.89**	**2.53**	**20.02**	**4.98**

表三 SMN 法結合時域型特徵正規化法之實驗綜合比較表

2. 調變頻譜平均值與變異數正規化法結合時域型特徵正規化法之實驗結果

實驗結果討論：

(1) 表四中，調變頻譜平均值與變異數正規化法結合時域上特徵正規化法與單一作特徵正規化法比較，皆能有效提升語音辨識效能。舉而言之：SB-SMVN$_{(4)}$結合 CMVN 的辨識率爲 89.87%，比起 CMVN 的辨識率 83.23%與 SB-SMVN$_{(4)}$的辨識率 86.36%，都有相當明顯的改善。

(2) 從表四也可以清楚發現，當低頻部份切割越細，更能有效提升語音辨識效能，而 SMVN 法結合其他特徵正規化法皆以分割四個頻段的效果最爲優越。其中 SB-SMVN$_{(4)}$ 結合 CMN 的 RR$_2$ 爲 20.73%，SB-SMVN$_{(4)}$結合 CMVN 的 RR$_2$ 爲 21.17%，SB-SMVN$_{(4)}$ 結合 MVA 的 RR$_2$ 爲 16.99%，SB-SMVN$_{(4)}$結合 HEQ 的 RR$_2$ 爲 9.80%。

Method		Set A	Set B	Set C	average	AR	RR$_1$	RR$_2$
Baseline		71.98	67.79	78.28	71.56	—	—	—
CMN		80.69	83.41	80.09	81.66	—	—	—
CMN	FB-SMVN	87.38	87.73	85.61	87.17	5.51	30.04	—
	SB-SMVN$_{(2)}$	88.60	88.89	86.83	88.36	6.70	36.53	9.28
	SB-SMVN$_{(3)}$	89.77	89.90	88.17	89.50	7.84	42.75	18.16
	SB-SMVN$_{(4)}$	90.01	90.35	88.43	**89.83**	**8.17**	**44.55**	**20.73**
CMVN		83.55	83.75	81.57	83.23	—	—	—
CMVN	FB-SMVN	87.33	87.80	85.47	87.15	3.92	23.38	—
	SB-SMVN$_{(2)}$	88.54	88.84	86.80	88.31	5.08	30.29	9.03
	SB-SMVN$_{(3)}$	89.72	89.90	88.03	89.45	6.22	37.09	17.90
	SB-SMVN$_{(4)}$	90.11	90.37	88.42	**89.87**	**6.64**	**39.59**	**21.17**
MVA		86.69	86.89	84.98	86.43	—	—	—
MVA	FB-SMVN	88.18	88.60	86.39	87.99	1.56	11.50	—
	SB-SMVN$_{(2)}$	89.27	89.49	87.58	89.02	2.59	19.09	8.58
	SB-SMVN$_{(3)}$	89.69	89.95	88.36	89.52	3.09	22.77	12.74
	SB-SMVN$_{(4)}$	90.27	90.45	88.71	**90.03**	**3.60**	**26.53**	**16.99**
HEQ		86.90	87.73	87.56	87.36	—	—	—
HEQ	FB-SMVN	89.05	89.55	89.26	89.29	1.93	15.27	—
	SB-SMVN$_{(2)}$	89.33	89.81	89.40	89.54	2.18	17.25	2.33
	SB-SMVN$_{(3)}$	89.95	90.35	89.91	90.11	2.75	21.76	7.66
	SB-SMVN$_{(4)}$	90.19	90.59	90.17	**90.34**	**2.98**	**23.58**	**9.80**

表四 SMVN 法結合時域型特徵正規化法之實驗綜合比較表

3. 調變頻譜統計圖等化法結合時域型特徵正規化法之實驗結果

實驗結果討論：

(1) 類似表三、表四，從表五看出，調變頻譜統計圖等化法結合時域型特徵正規化法皆優於個別特徵正規化法。舉例而言：SB-SHE$_{(L=4)}$結合 CMVN 的辨識率爲 90.18%，比起 CMVN 的辨識率 83.23%有相當明顯的改善。但 SB-SHE 結合 CMVN 之實驗結果與 SB-SHE 作在 MFCC 特徵之實驗結果比較，在某些結合方式下，辨識結果會有些微的下降，這可能是實驗的誤差範圍，或是過度正規化的不良效應。

(2) 從表五也可以清楚發現，當切割的子頻段數目越多，越能有效提升語音辨識效能，而 SHE 法結合其他特徵正規化法皆以分割四個頻段的效果最爲優越。其中 SB-SHE$_{(4)}$ 結合 CMN 的 RR$_2$ 爲 6.45%，SB-SHE$_{(4)}$結合 CMVN 的 RR$_2$ 爲 7.97%，SB-SHE$_{(4)}$結合

MVA 的 RR_2 爲 2.60%，SB-$SHE_{(4)}$結合 HEQ 的 RR_2 爲 6.19%。

Method		Set A	Set B	Set C	average	AR	RR_1	RR_2
Baseline		71.98	67.79	78.28	71.56	—	—	—
CMN		80.69	83.41	80.09	81.66	—	—	—
CMN	FB-SHE	89.45	90.08	88.24	89.46	7.80	42.53	—
	SB-$SHE_{(2)}$	89.44	89.97	88.24	89.41	7.75	42.26	-0.47
	SB-$SHE_{(3)}$	89.90	90.29	88.68	89.81	8.15	44.44	3.32
	SB-$SHE_{(4)}$	90.20	90.63	89.06	**90.14**	**8.48**	**46.24**	**6.45**
CMVN		83.55	83.75	81.57	83.23	—	—	—
CMVN	FB-SHE	89.42	89.87	88.07	89.33	6.10	36.37	—
	SB-$SHE_{(2)}$	89.45	89.96	88.23	89.41	6.18	36.85	0.75
	SB-$SHE_{(3)}$	89.74	90.22	88.61	89.70	6.47	38.58	3.47
	SB-$SHE_{(4)}$	90.23	90.67	89.10	**90.18**	**6.95**	**41.44**	**7.97**
MVA		86.69	86.89	84.98	86.43	—	—	—
MVA	FB-SHE	89.97	90.50	88.98	89.99	3.56	26.23	—
	SB-$SHE_{(2)}$	89.98	90.49	88.98	89.98	3.55	26.16	-0.10
	SB-$SHE_{(3)}$	90.25	90.65	89.30	90.22	3.79	27.92	2.30
	SB-$SHE_{(4)}$	90.25	90.76	89.22	**90.25**	**3.82**	**28.15**	**2.60**
HEQ		86.90	87.73	87.56	87.36	—	—	—
HEQ	FB-SHE	89.24	90.09	89.61	89.66	2.30	18.20	—
	SB-$SHE_{(2)}$	89.22	90.08	89.44	89.61	2.25	17.80	-0.48
	SB-$SHE_{(3)}$	89.48	90.30	89.82	89.87	2.51	19.86	2.03
	SB-$SHE_{(4)}$	89.91	90.75	90.17	**90.30**	**2.94**	**23.26**	**6.19**

表五 SHE 法結合時域型特徵正規化法之實驗綜合比較表

四、結論與未來展望

在本論文中，我們提出了一系列分頻段調變頻譜統計正規化的演算法，以不等切的方式切割調變頻譜，再分別針對每個頻段的調變頻譜強度作統計正規化，分析其對語音特徵在雜訊環境下提昇強健性的效果。由實驗結果發現，相對於傳統不切割頻段的方式而言，這些新方法都有明顯的改進效果，我們也發現由於調變頻譜中偏低頻(約 1~16 Hz)的語音成份包含了大多數語音辨識所需的資訊，若我們將此低頻部份切割地越細，進而個別正規化處理，越有效提升語音辨識效能，而在我們所提的各種分頻段調變頻譜正規化法中，皆以切割四個頻段所得的辨識效果最爲優越。另外，我們也將各種分頻段調變頻譜正規化法分別與傳統時間序列域上之特徵統計正規化法作結合；由辨識實驗結果發現，二者組合其辨識精確率皆比使用單一強健性技術所得到的辨識率更好。由此可看出，我們所提的分頻段式之新方法，不僅能有效改善原先全頻段式的方法，更與其他語

音強健性技術有良好的加成性，得以明顯改善雜訊環境下的語音辨識效能。

在未來展望中，我們將進一步研究分頻段調變頻譜統計正規化法中的理論基礎，並希望能藉由更嚴謹的數學分析與推導，求取這些方法中最佳的分頻段數目。此外，我們也希望相關實驗不僅在數字辨識上處理，也擴展至其他較大字彙量的語音辨識，探討這一系列分頻段調變頻譜統計正規化法在不同複雜度之語音辨識系統的效能，或是應用於其他類型的干擾失真環境，進一步驗證這些新方法的效能與實用性，以上各點都是未來能夠嘗試研究發展的方向。期盼將來語音辨識之效能能夠更加提升，並且普遍應用於日常生活，讓人們輕鬆地利用語音與電腦或 3C 產品進行互動，令生活能夠更便利，使語音辨識之發展兼具理論性與實用性。

參考文獻

[1] 王小川, "語音訊號處理", 全華科技圖書, 2004

[2] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. on Acoustics, Speech and Signal Processing*, pp.254-272, 1981

[3] Olli Viikki and Kari Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition ", *Speech Communication*, vol. 25, pp.133-147, 1998

[4] Ángel de la Torre, Antonio M. Peinado, José C. Segura, José L. Pérez-Córdoba, Ma Carmen Benítez, and Antonio J. Rubio, "Histogram equalization of speech representation for robust speech recognition", *IEEE Trans. on Speech and Audio Processing*, pp.355-366, 2005

[5] Liang-che Sun, Chang-wen Hsu and Lin-shan Lee, "Modulation spectrum equalization for robust speech recognition", *in Proc. IEEE Workshop on Automatic Speech Recognition & Understanding* (ASRU), pp.81-86, 2007

[6] Hynek Hermansky and Nelson Morgan, "RASTA processing of speech", IEEE Trans. on Speech and Audio Processing, pp.578-589, 1994

[7] Hynek Hermansky and Petr Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR", *2005 International Conference on Spoken Language Processing* (*Interspeech*), pp.361-364

[8] Noboru Kanedera, Takayuki Arai, Hynek Hermansky, and Misha Pavel, "On the importance of various modulation frequencies for speech recognition", *1997 European Conference on Speech Communication and Technology* (*Eurospeech*)**,** pp. 1079-1082

[9] David Pearce and Hans-Günter Hirsch, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *in Proc. of ISCA IIWR ASR2000*, Paris, France, pp.181-188, 2000

[10] ITU recommendation G.712, "Transmission performance characteristics of pulse code modulation channels", Nov. 1996

[11] Henry Stark, John W. Woods, "Probability and random processes with applications to signal processing", *3rd Edition*, Prentice-Hall, 2002

Web Mining for Unsupervised Classification

戴瑋彥　Wei-Yen Day
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
r97067@csie.ntu.edu.tw

紀均易　Chun-Yi Chi
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
ono.ccy@gmail.com

陳瑞呈　Ruey-Cheng Chen
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
cobain@turing.csie.ntu.edu.tw

鄭卜壬　Pu-Jen Cheng
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
pjcheng@csie.ntu.edu.tw

劉培森　Pei-Sen Liu
資訊工業策進會
Institute for Information Industry
psliu@iii.org.tw

Abstract

Data acquisition is a major concern in text classification. The excessive human efforts required by conventional methods to build up quality training collection might not always be available to research workers. In this paper, we look into possibilities to automatically collect training data by sampling the Web with a set of given class names. The basic idea is to populate appropriate keywords and submit them as queries to search engines for acquiring training data. Two methods are presented in this study: One method is based on sampling the common concepts among the classes, and the other based on sampling the discriminative concepts for each class. A series of experiments were carried out independently on two different datasets, and the result shows that the proposed methods significantly improve classifier performance even without using manually labeled training data. Our strategy for

retrieving Web samples, we find that, is substantially helpful in conventional document classification in terms of accuracy and efficiency.

Keywords: Unsupervised classification, text classification, Web mining

1. Introduction

Document classification has been extensively studied in the fields of data mining and machine learning. Conventionally, document classification is a supervised learning task [1, 2] in which adequately labeled documents should be given so that various classification models, i.e., classifiers, can be learned accordingly. However, such requirement for supervised text classification has its limitations in practice. First, the cost to manually label sufficient amount of training documents can be high. Secondly, the quality of labor works is suspicious, especially when one is unfamiliar with the topics of given classes. Thirdly, in certain applications, such as email spam filtering, prototypes for documents considered as spam might change over time, and the need to access the dynamic training corpora specifically-tailored for this kind of application emerges. Automatic methods for data acquisition, therefore, can be very important in real-world classification work and require further exploration.

Previous works on automatic acquisition of training sets can be divided in two types. One of which focused on augmenting a small number of labeled training documents with a large pool of unlabeled documents. The key idea from these works is to train an initial classifier to label the unlabeled documents and uses the newly-labeled data to retrain the classifier iteratively. Although classifying unlabeled data is efficient, human effort is still involved in the beginning of the training process.

The other type of work focused on collecting training data from the Web. As more data is being put on the Web every day, there is a great potential to exploit the Web and devise algorithms that automatically fetch effective training data for diverse topics. A major challenge for Web-based methods is the way to locate quality training data by sending effective queries, e.g., class names, to search engines. This type of works can be found in [3, 4, 5, 6], which present an approach that assumes the search results initially returned from a class name are relevant to the class. Then the search results are treated as auto-labeled and additional associated terms with the class names are extracted from the labeled data. By sending the class names together with the associated terms, appropriate training documents can be retrieved automatically. Although generating queries is more convenient than manually collecting training data, the quality of the initial search results may not always be good especially when the given classes have multiple concepts. For example, the concepts of class “Apple” include company and fruit. Such a problem can be observed widely in various applications.

The goal of this paper is, given a set of concept classes, to automatically acquire training corpus based merely on the names of the given classes. Similar to our previous attempts, we

employ a technique to produce keywords by expanding the concepts encompassed in the class names, query the search engines, and use the returned snippets as training instances in the subsequent classification tasks. Two issues may arise with this technique. First, the given class names are usually very short and ambiguous, making search results less relevant to the classes. Secondly, the expanded keywords generated from different classes may be very close to each other so that the corresponding search-result snippets have little discrimination power to distinguish one class from the others.

We present two concept expansion methods to deal with these problems, respectively. The first method, expansion by common concepts, aims at alleviating the problem of ambiguous class names. The method utilizes the relations among the classes to discover their common concepts. For example, "company" could be one of the common concepts of classes "Apple" and "Microsoft". Combined with the common concepts, relevant training documents to the given classes can be retrieved. The second method, expansion by discriminative concepts, aims at finding discriminative concepts among the given classes. For example, "iPod" could be one of the unique concepts of class "Apple". Combined with the discriminative concepts, effective training documents that distinguish one class from another can be retrieved.

Our methods are tested under two different experimental setups, the CS papers and Web pages classification tasks. The proposed methods are effective in retrieving quality training data by querying search engines. Moreover, the result shows that the obtained Web training data and manually labeled training data are complementary. Our methods can significantly improve classification accuracy when only a few manually labeled training data is available. Contribution of our work can be addressed as follows. We propose an automatic way to sample the Web and collect the training data with good quality. Apart from the previous work, our methods are fully automatic, reliable, and robust, and achieve an 81% accuracy in text classification tasks. With a little help from a small number of labeled data added into the scene, the classification accuracy can be as high up to 90%. Several experiment results are also revealed to help investigation and realization of automatic Web sampling methods, in which the difficulties encountered are presented in detail.

The sections are organized as follows. In Sections 2 and 3, we present our basic idea and the two methodologies, respectively. The experiments are introduced in Section 4. In Section 5, we discuss the related work of this paper. Finally, in Section 6, we give out discussions and conclusions.

2. The Basic Idea

Suppose that we are given a set of classes $C=(c_1, c_2, ..., c_n)$, where c_i is the name of the i-th class. We plan to generate keywords based on classes C, form a few queries and send them off to search engines so as to collect training instances. Our methods presented in this paper are independent of classification models; that is, any model can be incorporated with our methods.

To carefully examine the possibility of querying search engines for acquiring training data, we did an evaluation with different search engines, search-result types (snippet or document), and the number of search results. 5 CS-related classes were taken into account, including "Architecture", "IR", "Network", "Programming", and "Theory". Each class name c_i was sent to 3 search engines, including Google[1], Yahoo![2], and Live Search[3]. Top 100 snippets were extracted as training data. We also gathered the research papers from the corresponding conferences to the 5 classes as the testing documents. Table 1 shows the performance of different search engines. Querying by the class names can achieve classification accuracy at a range from 0.35 to 0.56. More specifically, the three search engines perform well in "Programming" and "Theory" but poorly in the others on average. This arises from the fact that irrelevant documents may be located for those classes with ambiguous names. The way to query by the class names is not reliable due to the ambiguity of the class names (the first challenge). For example, the word "architecture" is widely used in CS, art and construction. From the results, we select Google as our backend search engine in this paper.

We further explore if the classification performance can be improved by downloading Web pages for training. The result is shown in Table 2. It reveals that Web pages might introduce more noises than snippets do, while the snippets summarize Web pages and capture the concepts of classes C by their context. Moreover, to download Web pages is time-consuming. Our methods, therefore, only retrieve snippets as the training source.

Table 1. Accuracy of different search engines for classification of CS papers.

Engine	Architecture	IR	Network	Programming	Theory	Avg.
Google	0.075	0.382	0.899	0.723	0.762	0.568
Yahoo!	0.112	0.022	0.094	0.863	0.665	0.351
Live Search	0.269	0.006	0.083	0.784	0.815	0.391

Intuitively, collecting more snippets or documents might enhance the performance. Table 3 shows the results of changing training data sizes from 100 to 900. It could be found that classification accuracy does not increase obviously when the numbers of snippets and documents reach 200 and 300, respectively. This is because much relevant information can be retrieved in top ranked search results returned by the search engine. Noises are unavoidably included from longer lists. Hence, simply fetching a large amount of snippets or documents from a single search result cannot achieve satisfactory performance. Even if we expand the queries, i.e., the class names, using pseudo-relevance feedback (PRF) [7, 8, 9], the improvement is still minor since the generated expanded keywords cannot effectively discriminate different classes (our second challenge). The performance comparison between our methods and PRF will be given in Section 4.2.

[1] The Google search engine: http://www.google.com/search
[2] The Yahoo! search engine: http://search.yahoo.com/search
[3] The MSN Live search engine: http://search.live.com/

To collect good training corpora and help classifiers learn more quickly (querying search engines is costly), two methods are proposed in this paper. The first method, expansion by common concepts, aims at alleviating the ambiguity problem. Generally, a short class name easily conveys multiple meanings. For example, class "Apple" may be a fruit or company name. We find that class c_i is context-aware if its context $C - \{c_i\}$ provides relevant information to c_i. For example, if "Apple" and "Microsoft" are put together, "Apple" would be a company. If we are given "apple" and "banana", "apple" could refer to a fruit. Our first method, which will be described in Section 3.1, is trying to discover such common concepts among classes C, i.e., "company" and "fruit", from the Web, and use them as constraints to expand our original queries C.

Table 2. Accuracy of different training types in CS papers classification.

Source	Architecture	IR	Network	Programming	Theory	Avg.
Snippet	0.075	0.382	0.899	0.723	0.762	0.568
Documents	0.272	0.049	0.689	0.505	0.783	0.459

Table 3. Average accuracy of different training sizes in CS papers classification.

# of docs	100	200	300	400	500	600	700	800	900
Snippets	0.568	0.601	0.603	0.601	0.608	0.603	0.604	0.604	0.604
Documents	0.460	0.463	0.507	0.500	0.503	0.504	0.507	0.507	0.508

Common concepts can help us collect more relevant documents to each class but cannot discriminate one class from the others. The latter becomes important because classification is inherently to distinguish different classes. Our second method is focused on the finding of discriminative concepts among classes C. Consider previous example. "PowerPoint" and "iPod" are possible discriminative concepts because "PowerPoint" is only relevant to "Microsoft" while "iPod" is only about "Apple". Different from PRF, our second method, expansion by discriminative concepts, which will be described in Section 3.2, aims at acquiring Web training data not only relevant to each class but also effectively distinguishing one class from another.

3. The Proposed Methods

In this section, we will describe the two training data acquiring methods, sampling the Web by common concepts expansion and discriminative concepts expansion, respectively.

3.1 Expansion by Common Concepts

The goal of our first method is to collect training data via sampling the Web by discovering the common concepts between given classes C. Expanding class names C by their common concepts is helpful in obtaining more suitable training data from search engines. An intuitive way to discover the concepts is to find common concepts from well-known topic hierarchies on the Web such as Open Directory Project[4] (DMOZ), which is one of the largest and comprehensive human-edited directories. To obtain common concepts, we first search

[4] DMOZ Open Directory Project: http://www.dmoz.org/

DMOZ by each class name c_i and get a set of the nodes relevant to c_i in the DMOZ directory. Suppose the set of the nodes is $N(c_i)$. The least common ancestors (LCA) of all of the nodes $N(c_i)$ $(i=1...n)$ are viewed as the common concepts of C. The LCAs are the shared ancestors of $N(c_i)$ $(i=1...n)$ located farthest from the roots of the DMOZ directory. For example, by searching class "Architecture", we get the path "Arts: Architecture" and "Computers: Emulators: Intel x86 Architecture". When search the class "Programming", we find it has the same common concept of "Computers" (from "Computers: Programming") with "Architecture". Thus the concept "Computers" would be the common concepts between the two classes "Architecture" and "Programming".

Although Open Directory Project covers diverse topics and is very precise, sometimes we might get few or even no common concepts among C. The problem is serious for those classes not so popular such as names of person or organizations. For example, if we query the class "Cornell", we only get 5 paths for the class, which contains few candidates of concepts to select and expand. To deal with this problem, we extract terms co-occurring with each class c_i in Web pages, cluster the terms, and treat the representative term for each class as one of the common concepts. More specifically, all of the classes $\{c_1, c_2, \ldots, c_n\} \in C$ are combined into one query "$c_1 + c_2 + \ldots + c_n$", and then submitted to a search engine. After stemming and removing stopwords, we extract 20 high-frequency terms as candidates for common concepts from top 100 snippets returned from the search engine. To group these candidates, we send them separately to the search engine and generate corresponding feature vectors based on their top 100 snippets. Uni- and bi-grams are adopted as feature terms and TF-IDF is used to calculate feature weights. Next, a graph $G = (V, E)$ is constructed, where $v \in V$ represents one candidate term, and $e \in E$ is the cosine similarity between two feature vectors. Finally, we perform the star clustering algorithm [10] to choose the star centers, which are the common concepts among C.

In this paper, we adopt both of LCAs from DMOZ and co-occurring terms from Web pages as our common concepts among C. After common concepts generated, we can either use it to sample the Web and acquire good training data, or utilize them while discriminative concepts are generating.

3.2 Expansion by Discriminative Concepts

A discriminative concept is a concept that can help distinguish one certain concept class of interest (say, c) from all the classes ($c' \neq c$). Such a concept contributes more relevance to one specific class than to the others. Unlike common concepts, which are shared by all the concept classes in C, any discriminative concept has a specific concept class to contribute relevance to, called the *host*. Let f_c be the feature vector of concept class c. Let the similarity between any two concepts x and y be denoted as $\sigma_{x,y} = cos(f_x, f_y)$. An ideal discriminative concept k for concept c must satisfy all the following constraints:

1. Concept k should exhibit high similarity to its host c.

2. The similarity between k and c should be significantly greater than that between k and any one of the other concept classes.

These constraints loosely define the criteria that we can use to discover discriminative concepts, and they also rule out the possibility that a concept k has two or more hosts. Based on the second constraint, a plausible decision criteria for discriminative concept is given below. Let κ_c denote the set of all discriminative concepts hosted by concept class c. We have:

$$k \in \kappa \Leftrightarrow \frac{\sigma_{c,k}}{\sigma_{c',k}} > \theta \; for \; all \; c' \neq c$$

In the criteria, the right-hand side equation needs to be satisfied for all other concept classes c'; in other words, we have a multiple-constraint-satisfaction problem. For every concept-class pair (c, c'), the ratio $\sigma_{c,k}/\sigma_{c',k}$ describes the degree of deviation in similarities exhibited by k toward both classes. When the value is greater than 1, we say that c is more likely to be the host; when it lies between 0 and 1, we say that c' is more likely. The parameter θ determines the tightness of the decision boundary. Since we expect higher similarity between k and c than that between k and other c's, it would normally be defined as a value greater than 1.

Generally, fixed boundary value is easier to train and suitable for general cases, while we also find that this type of setup can cause problems in extreme cases. Suppose we have two classes c_1 and c_2 which are relevant topics or extraordinarily similar to each other (in terms of the similarity between their feature vectors). The number of discriminative concepts for either c_1 or c_2 may drastically decrease because, for most k's, deviation in similarities toward both classes becomes even more subtle and harder to detect by using a simple constant θ. An obvious solution to this problem, we find, is to adopt a function in place of the constant θ, that assigns high threshold value for general cases and low threshold value for the aforementioned extreme cases. In real practice, we use a very simple form of decision criteria to identify discriminative concepts:

$$k \in \kappa \Leftrightarrow \frac{\sigma_{c,k}}{\sigma_{c',k}} > \delta(1 - \sigma_{c,c'}) \text{ for all } c' \neq c$$

where δ is the *discrimination coefficient*. With the new criteria, a number of 5 to 40 discriminative concepts (for each class) can still be discovered even when any two concept classes exhibit high class-to-class similarity to each other.

The next step is to apply this technique to each concept class in C so as to populate new training instances from the Web. Let SR_c be the set of search-result snippets obtained by sending concept c as a query to the search engine. Assume that the training set for concept class c before and after the expansion is denoted as D_c and D'_c, respectively. Given κ_c, we can form a set of new queries by concatenating c with each $k \in \kappa_c$,, send them off to the search engine, and obtain a new set of training instances, i.e., $\{ SR_{c \cup k} \in \kappa_c \; / \; k \in \kappa_c \}$. In formalism, we have:

$$D'_c = D_c \cup_{k \in \kappa_c} SR_{c \cup k}$$

This procedure can be repeated several times for each class so that the total number of discovered training instances can reach our expectation. However, certain changes on definitions and notations required by the adaptation needs to be clarified in advance. First, we can no longer expect that the feature vector for a concept class c remains the same throughout multiple iterations. In each iteration, when new training instances added to the collection, feature vectors for c actually changes. Next, the set of discriminative concepts κ_c discovered in each iteration would vary, as similarity measure suffers from the change as well. Certain modification in definitions should be taken care of so as to seamlessly fit the aforementioned criteria into the framework, while treating notations for a concept and for a concept class differently might clutter up the framework. For simplicity, no explicit treatment will be done to the equations in the following text. Readers should take caution that, when class-to-concept or class-to-class similarity computation is considered in discussions, the feature vector referred to a concept class is in fact derived from its current training set (rather than SR_c). On the other hand, we will use $\kappa_c^{(i)}$ in place of plain κ_c in the rest of the work.

Assume that the algorithm repeats t times. The initial training data for concept class c is denoted as $D_c^{(0)}$, and in each iteration $i \in [1,t]$, a new training set for c is produced and represented by $D_c^{(i)}$. Consider a simple framework as follows. For all $c \in C$, we have:

$$\mathrm{D_c}^{(0)} = \mathrm{SR_c}$$
$$\mathrm{D_c}^{(\mathrm{i})} = \mathrm{D_c}^{(\mathrm{i})} \cup_{\mathrm{k} \in \kappa_c^{(\mathrm{i})}} \mathrm{SR_{c \cup k}}, \qquad \mathrm{i} > 0$$

where $D_c^{(i)}$ and $\kappa_c^{(i)}$ is the set of training instances and the set of discovered discriminative concepts, respectively, for concept class c at iteration i. Eventually, the algorithm stops after the t-th iteration. The content of $D_c^{(t)}$ will serve as the final training data for all concept class c.

Since practically it is infeasible to populate the entire set of κ_c, several heuristics are involved in creation of the set: 1) We look for terms with high discrimination power (specifically, unigram and bigrams) in the context of $D^{(i-1)}$ using commonly-used information-theoretic measures, such as information gain and inverse document-frequency. These candidates are then examined with the decision criteria and disqualified ones are discarded immediately; 2) candidates that survived the test are ranked accordingly by the score function, which is a simple rewrite of the criteria that indicates the average degree of deviation for the candidate k:

$$\frac{1}{|\mathrm{c}| - 1} \sum_{\mathrm{c}' \neq \mathrm{c}} \left[\frac{\sigma_{\mathrm{c,k}}}{\sigma_{\mathrm{c}',\mathrm{k}}} - \delta(1 - \sigma_{\mathrm{c,c}'}) \right]$$

The summands will not cancel out since the score is calculated for the candidates satisfying the criteria. Generally, testing all the candidate terms may result in an extremely inefficient procedure. In practice, we set up a strict threshold on the information gain and idf in light of reducing the number of candidates. We test only the top m concepts selected by the filter in the end. The value m is set to be 15 throughout the work.

4. Experiments

We evaluate our performance in CS papers and Web pages classification.

4.1 Experimental Setup

We conduct experiments on two datasets. First is a set of papers from several CS-related conferences, and there are five classes used for training, including "Architecture", "IR", "Network", "Programming", and "Theory", as shown in Table 4. For the paper dataset, there are about 500 papers in each class. Another dataset is the Web pages collected from four universities, and can be downloaded from the WebKB project[5]. The classes for these universities include "Cornell", "Texas", "Washington", and "Wisconsin". As the original dataset for Web pages is imbalanced, we randomly choose 827 Web pages (i.e. the minimum size of original data among the 4 classes) for classification. Please note that the two datasets are our testing data since the training data are fully collected from the Web. Moreover, the two datasets are quite different in document length and quality. The papers are often longer, well written, and with much useful information about the CS-related classes, while the Web pages might cover more noises and shorter contents.

Our first method of expansion by common concepts is denoted as CM; the second method of expansion by discriminative concepts is denoted as DM; CM+DM is the combination of both methods, where CM is applied first so that search results could be more relevant, and then DM is used to extract discriminative concepts from the relevant search results. With the common concepts extracted from CM, we can use these concepts plus the class name as a whole query and perform DM to iteratively collect the discriminative concepts.

Table 4. The information about the dataset of CS papers for classification.

Class	# papers	From Conferences
Architecture	490	SIGARCH(04-08), DAC(00-07)
IR	484	SIGIR(02-07), CIKM(02-07)
Network	446	SIGCOMM(02-08), IPSN(04-07), MOBICOM(03-07), MOBIHOC(00-07), IMC(01-07)
Programming	505	POPL(02-08), PLDI(00-07), ICFP(02-07), OOPSLA(00-07)
Theory	471	SODA(01-08), STOC(00-07)

To compare our performance with the state-of-the-art Web-based method, LiveClassifier [3], denoted as LC, has been implemented, where the concepts hierarchy are referred from DMOZ. For the CS-related classes, these concepts are "computers", "reference", "business", "software", and "science"; and for the classes of four universities, the concepts are "society", "people", "university", "school", "education", "sports", "United States". The thresholds δ and t used in DM are set to 0.85 and 7 in both dataset, respectively. We use the Rainbow tool[6] and the VSM (Vector Space Model) classification model for all the experiments.

[5] The WebKB project: http://www.cs.cmu.edu/~WebKB/
[6] The Rainbow tool: http://www.cs.cmu.edu/~mccallum/bow/rainbow/

4.2 Text Classification

Table 5 compares the classification accuracy between different methods. For each class, the baseline method is to use only the class name as query, submit to the search engine, and collect snippets as training data. The method from query expansion (QE) is to use pseudo-relevance feedback (PRF) for each class, where the terms with high TF-IDF values in the snippets are selected as expansion terms which used for acquiring training data. LC is the method implemented based on LiveClassifier [3]. With the concept hierarchy of each class, all the concepts are used to combine with the class name as a query, then submit to the search engine and collect the snippets as training data.

From Table 5, we find that merely sending class names as queries cannot retrieve quality training data. It only achieves the accuracy at 0.57 on average. Even if we expand the class names with PRF (the method of QE), the accuracy cannot be improved by QE in this case. This is because PRF is a general solution to the keyword mismatching problem and thus would not be well applied to our classification problem. LC manually labels some concepts of the class names, e.g., the concept of "Architecture" is about computer architecture, so that more relevant training data to original classes can be fetched. LC gets higher accuracy at about 0.71. But such concepts labeled by people are often few. Our CM method can discover more useful common concepts, as shown in Table 6, where keywords "computers", "conference", and "proceeding" are all helpful in searching relevant data for each class when combined with original class name. CM, on the other hand, might introduce noisy keywords often co-occurring with the class names such as keyword "web". Due to the assistance of more common concepts, we collect more quality training data by CM. The accuracy for CM comes to 0.76, which is much better than LC and QE.

In addition, we observe that the performance for class "Network" is originally high in baseline method. To realize what makes the result, we check the process for getting training data in advance. Except for the accurate semantic for the term "Network", we discover that the snippets from search engine are suitable and with good quality for the class "Network". This is related to the characteristic of the Web. To our experience, the documents ranked from search engines might be mostly relevant to the fields about computer or network.

The results of Web pages classifications, as shown in Table 7, are similar to previous experimental results except the average accuracy obtained here is lower in general due to the noisy Web pages that are unreliable for testing. Moreover, sometimes the concepts derived from the Web are not effective in classification (even they are correct). For example, from the Web, DM learns two discriminative concepts of "ut" and "milwaukee" for classes "Texas" and "Wisconsin", respectively. But their impacts on classifying our testing data are futile. Although there are some noises in the Web page dataset and the concepts found are correct but less useful, CM+DM does an improvement while training by the snippets from these concepts, and surpasses the performance of LiveClassifier (LC). By our methods, the quality training data are fetched and the classifier learns better, thus become more robust for text classification.

Table 5. Accuracy of classification in CS papers.
Baseline: Class Names, QE: Query Expansion, LC:LiveClassifier,
CM: Common concept method, and DM: Discriminative concept method.

Method	Architecture	IR	Network	Programming	Theory	Avg.
Baseline	0.76	0.38	0.90	0.72	0.76	0.57
QE	0.03	0.01	0.98	0.71	0.66	0.48
LC	0.88	0.42	0.82	0.54	0.89	0.71
CM	0.80	0.77	0.89	0.56	0.76	0.76
DM	0.04	0.00	0.83	0.66	0.98	0.50
CM+DM	0.76	0.86	0.91	0.71	0.82	0.81

Table 6. Extracted common concepts and discriminative concepts of CS classes.

Method	Architecture	IR	Network	Programming	Theory
CM	By DMOZ: computers By Star Algorithm: conference, proceeding, web				
DM	architecture, architectural, architects, arts, history, contemporary, design	infrared, satellite, visible, image, weather, thermal, cameras	usa, health, action, sports, food, monitor, global	tutorial, java, example, using, oriented, download, linux	graph, mathematical, problems, literary studies, political, number,
CM+DM	isca, energy, oriented, adaptive, aidede, ecaade, architectural	investor, infrafed, retrieval, ecir, trec, sigir, ie	development , wireless, shows, server, first, email ,mail	ferment, oriented, programmers, final, siggraph, graphics, animation	Number, university, critical, math, computational, theoretical, complexity

Table 7. Accuracy of classification in Web pages.
Baseline: Class Names, QE: Query Expansion, LC:LiveClassifier,
CM: Common concept method, and DM: Discriminative concept method.

Method	Cornell	Texas	Washington	Wisconsin	Avg.
Baseline	0.85	0.47	0.68	0.20	0.55
QE	0.87	0.40	0.71	0.41	0.60
LC	0.99	0.14	0.58	0.14	0.46
CM	0.99	0.30	0.72	0.33	0.59
DM	0.47	0.75	0.53	0.66	0.60
CM+DM	0.98	0.43	0.77	0.37	0.64

4.3 Combination of Web Training Data for Text Classification

In this experiment, we want to realize how our methods help conventional supervised text classification. We further conduct an experiment to compare the performance between labeled data plus training data from the Web and the labeled data only.

Both datasets are divided into training and testing data. We combine the training data collected from the Web and sample α % of the training data as a new training corpus. For comparison, we also train a classifier with just those α % of the labeled data only. The α value varies from 1 to 100. α =100 means to use all the labeled data from original divided training set, thus become a supervised learning. 5-fold cross validation is used to evaluate the classification accuracy.

Figure 1 and 2 show the two experimental results, respectively. We find that the Web corpus improves the classifier's accuracy more than 6% when α is small for both datasets. In other words, when the labeled data is insufficient or even with no quality, sampling training data from the Web could substantially complements the manually-labeled data. The performance of both classifiers increases when more labeled data are included. However, when more and more manually-labeled data is given, the improvement by the Web becomes less obvious or even slightly worse. This is because once we add more data from the Web, we also introduce the noises such that the accuracy glows slowly when the quality documents are enough. In CS papers classification, the performance reaches as higher as all training data used when 50% labeled data is added. For Web pages classification, we have the same performance as supervised learning when only use 20% labeled data. The performance even exceeds the result from supervised learning when 40% labeled data is joined. It explains more quality data from the Web is helpful in classification. This result also shows that using all the labeled data is not always as good as expected because some of the labeled data are not in good quality.

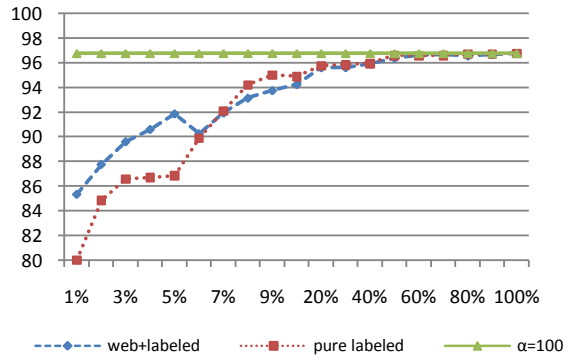


Fig. 1. Accuracy of training data from the Web plus % labeled data for CS papers classification.

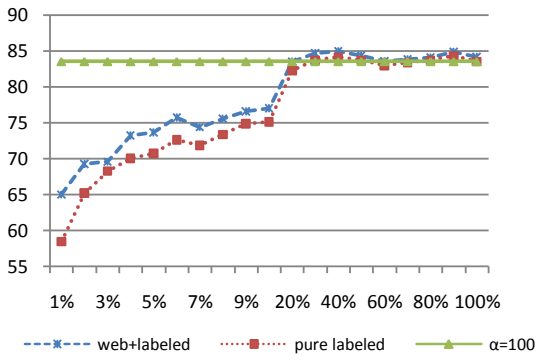


Fig. 2. Accuracy of training data from the Web plus % labeled data for Web pages classification.

This experiment shows us that the training data from the Web does help to improve the text classification. Moreover, we could use a very few number of labeled data, plus the Web corpus our methods collect, to train a desirable classifier. The Web helps the classifiers to learn the unseen concepts which do not exist due to the insufficiency or the unreliable quality of labeled data. It also tells us that the suitable training data might change by time, thus the original labeled data performs worse in new classification tasks. From the result, we believe that our methods can benefit the task of text classification, and other advanced applications.

5. Related Work

Text classification has been extensively studied for a long time in many research fields. Conventionally, supervised learning is usually applied in text classification [1, 2]. Our work focuses on the problem of how to adequately acquire and label documents automatically for classification models.

For the problem of automatic acquiring training data, previous studies discuss in two directions. One is focused on augmenting a small number of labeled training documents with a large pool of unlabeled documents [11, 12, 13, 14, 15, 16, 17, 18]. Such work trains an initial classifier to label the unlabeled documents and uses the newly-labeled data to retrain the classifier iteratively. [11] proposed by Nigam et al. use the EM clustering algorithm and the naive Bayes classifier to learn from labeled and unlabeled documents simultaneously. [12, 13] proposed by Yu et al. efficiently computes an accurate classification boundary of a class from positive and unlabeled data. In [15], Li et al. use positive and unlabeled data to train a classifier and solve the lack of labeled negative documents problem. Fung et al. in [17] study the problem of building a text classifier using positive examples and unlabeled example while the unlabeled examples are mixed with both positive and negative examples. [14] proposed by Nigam et al. starts from a small number of labeled data and employs a bootstrapping method to label the rest data, and then retrain the classifier. In [18], Shen et al. propose a method to use the n-multigram model to help the automatic text classification task. This model could automatically discover the latent semantic sequences contained in the document set of each category. Yu et al. in [16] present a framework, called positive example based learning (PEBL), for Web page classification which eliminates the need for manually collecting negative training examples in preprocessing. Although classifying unlabeled data is efficient, human effort is still involved in the beginning of the training process. In this paper, we propose an acquiring process of training data from the Web, which is fully automatic. The method trains a classifier well for document classification without labeled data, which is the mainly different part from the previous work. Moreover, our experiments show that the Web can help the conventional text classification. The training data acquired from the Web expand the coverage of classifier, which substantially enhance the performance while there is a lack of labeled data, or the quality of labeled data is not well enough.

Another direction is focused on gathering training data by the Web [3, 4, 5, 6]. In [3], Huang et al. propose a system, called “LiveClassifier”, which combines relevant class names as queries based on a user-defined topic hierarchy so that more relevant documents to the classes could be found from the Web. [4, 5, 6] proposed by Hung et al. presents an approach that assumes the search results initially returned from a class name are relevant to the class. These search results are treated as auto-labeled and additional associated terms with the class names are extracted from the labeled data. Although the previous works are similar to our methods, all of them are human-intervened. In this paper, we propose a method which automatically finds the associated concepts for the related classes and train a desirable classifier. The main contribution is that our method utilizes the relationship of classes and samples the Web in an automatic way for key concepts of each class, thus further find the

quality training data from the Web. Without labeled data and associated terms given by human, our methods perform well and classify documents accurately for the text classification problem.

6. Discussions and Conclusions

In this paper, we propose two methods to automatically sample the Web and find quality training data for text classification. We first examine the effects of different search engines, retrieved data types, and sizes of retrieved data. Moreover, from the subset of documents and the method by associated terms, we know that sampling the Web for concepts of classes and fetching training data can substantially improve the performance of classification. It might be hard to distinguish the classes with the ambiguity and close relationship without labeled data. By the discovering of common concepts and discriminative concepts, the ambiguity of class names is eliminated and more relevant concepts are utilized for sampling suitable and quality training data from the Web. Several experiments conducted in this work show that our methods are useful and robust for classifying documents and Web pages. Furthermore, our experiments show that the training data sampled from the Web helps the conventional supervised classification, which need quality and labeled data. The result demonstrates that the quality of labeled data might not always desirable due to the lack of useful key concepts, and we can provide proper training data from the Web to further improve the results of text classification. In additions, two dataset with different characteristics are used for our experiments and the analysis from different dataset is carefully conducted in this paper. Compared to previous works, the advantage of our methods is the fully automatic processes during the concepts expansion and the training data collecting. Our methods are independent of classification models, thus existing models can be incorporated with the proposed methods.

However, our work has some limitations. The classes we choose are related to each other. In other words, the performance would be better while the classes are in the same level in the hierarchy of topic classes. With the relationships between the classes, our methods can perform the context-aware technique among the classes to acquire more relevant documents, making the classifiers robust. To go a step further, there are more challenges to choose the quality documents in the training corpus sampled from the Web. We can also sample good training documents while a pool of unlabeled data is provided. We believe that these challenges are worth studied and would be the research directions in our future work.

References

[1] Y. Yang and X. Liu, “A re-examination of text categorization methods,” in Proceedings of the 22nd Annual International ACM SIGIR conference, 1999, pp. 42–49.

[2] Y. Yang, “An evaluation of statistical approaches to text categorization,” Information Retrieval, vol. 1, pp. 69–90, 1999.

[3] C.-C. Huang, S.-L. Chuang, and L.-F. Chien, “Liveclassifier: Creating hierarchical text classifier through web corpora,” in World Wide Web Conference, 2004.

[4] C.-C. Huang, K.-M. Lin, and L.-F. Chien, “Automatic training corpora acquisition through web mining,” in IEEE/WIC/ACM Conference on Web Intelligence, 2005.

[5] C.-M. Hung and L.-F. Chien, “Text classification using web corpora and em algorithms,” in The Asia Information Retrieval Symposium, 2004, pp. 12–23.

[6] C. M. Hung and L. F. Chien, “Web-based text classication in the absence of manually labeled training documents,” Journal of the American Society for Information Science and Technology, pp. 88–96, 2007.

[7] C. Carpineto, R. D. Mori, G. Romano, and B. Bigi, “An information theoretic approach to automatic query expansion,” ACM Transactions on Information Systems, vol. 19(1), pp. 1–27, 2001.

[8] Y. Qui and H. Frei, “Concept based query expansion,” in Proceedings of the 16th Annual International ACM SIGIR Conference, 1993, pp. 160–169.

[9] J. Xu and W. B. Croft, “Query expansion using local and global document analysis,” in Proceedings of the 19th Annual International ACM SIGIR Conference, 1996, pp. 412–420.

[10] J. Aslam, K. Pelekhov, and D. Rus, “A practical clustering algorithm for static and dynamic information organization,” in In: ACM-SIAM Symposium on Discrete Algorithms. In: ACM-SIAM Symposium on Discrete Algorithms (1999), 1999.

[11] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell, “Text classification from labeled and unlabeled documents using em,” Machine Learning, vol. 39(2/3), pp. 103–134, 2000.

[12] H. Yu, “Svmc: Single-class classification with support vector machines,” in Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, 2003.

[13] H. Yu and C. Zhai, “Text classification from positive and unlabeled documents,” in Proceedings of the 12th Annual International ACM Conference on Information and Knowledge Management, 2003, pp. 232–239.

[14] A. McCallum and K. Nigam, “Text classification by bootstrapping with keywords,” in ACL Workshop for Unsupervised Learning in Natural Language Processing, 1999.

[15] X. Li and B. Liu, “Learning to classify texts using positive and unlabeled data,” in Proceedings of International Joint Conferences on Artificial Intelligence, 2003.

[16] H. Yu, J. Han, and K.-C. Chang, “Pebl: Web page classification without negative examples,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, pp. 70–81, 2004.

[17] G. Fung, J. Yu, H. Lu, and P. Yu, “Text classification without labeled negative documents,” in Proceedings of 21st International Conference on Data Engineering, 2005.

[18] D. Shen, J.-T. Sun, Q. Yang, H. Zhao, and Z. Chen, “Text classification improved through automatically extracted sequences,” in Proceedings of the 22nd International Conference on Data Engineering, 2006.

Query Formulation by Selecting Good Terms

李佳蓉　Chia-Jung Lee,　林怡君　Yi-Chun Lin, 陳瑞呈　Ruey-Cheng Chen
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
{cjlee1010, yi.crystal, rueycheng}@gmail.com

劉培森　Pei-Sen Liu
資訊工業策進會
Institute for Information Industry
psliu@iii.org.tw

鄭卜壬　Pu-Jen Cheng
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
pjcheng@csie.ntu.edu.tw

Abstract

It is difficult for users to formulate appropriate queries for search. In this paper, we propose an approach to query term selection by measuring the effectiveness of a query term in IR systems based on its linguistic and statistical properties in document collections. Two query formulation algorithms are presented for improving IR performance. Experiments on NTCIR-4 and NTCIR-5 ad-hoc IR tasks demonstrate that the algorithms can significantly improve the retrieval performance by 9.2% averagely, compared to the performance of the original queries given in the benchmarks. Experiments also show that our method can be applied to query expansion and works satisfactorily in selection of good expansion terms.

Keywords: Query Formulation, Query Term Selection, Query Expansion.

1. Introduction

Users are often supposed to give effective queries so that the return of an information retrieval (IR) system is anticipated to cater to their information needs. One major challenge they face is what terms should be generated when formulating the queries. The general assumption of previous work [14] is that nouns or noun phrases are more informative than other parts of speech (POS), and longer queries could provide more information about the underlying information need. However, are the query terms that the users believe to be well-performing really effective in IR?

Consider the following description of the information need of a user, which is an example description query in NTCIR-4: Find articles containing the reasons for NBA Star Michael Jordan's retirement and what effect it had on the Chicago Bulls. Removing stop words is a common way to form a query such as “contain, reason, NBA Star, Michael Jordan, retirement, effect, had, Chicago Bulls”, which scores a mean average precision (MAP) of 0.1914. It appears obviously that terms contain and had carry relatively less information about the topic. Thus, we take merely nouns into account and generate another query, “reason, NBA Star,

Michael Jordan, retirement, effect, Chicago Bulls", which achieves a better MAP of 0.2095. When carefully analyzing these terms, one could find that the meaning of Michael Jordan is more precise than that of NBA Star, and hence we improve MAP by 14% by removing NBA Star. Yet interestingly, the performance of removing Michael Jordan is not as worse as we think it would be. This might be resulted from that Michael Jordan is a famous NBA Star in Chicago Bulls. However, what if other terms such as reason and effect are excluded? There is no explicit clue to help users determine what terms are effective in an IR system, especially when they lack experience of searching documents in a specific domain. Without comprehensively understanding the document collection to be retrieved, it is difficult for users to generate appropriate queries. As the effectiveness of a term in IR depends on not only how much information it carries in a query (subjectivity from users) but also what documents there are in a collection (objectivity from corpora), it is, therefore, important to measure the effectiveness of query terms in an automatic way. Such measurement is useful in selection of effective and ineffective query terms, which can benefit many IR applications such as query formulation and query expansion.

Conventional methods of retrieval models, query reformulation and expansion [13] attempt to learn a weight for each query term, which in some sense corresponds to the importance of the query term. Unfortunately, such methods could not explain what properties make a query term effective for search. Our work resembles some previous works with the aim of selecting effective terms. [1,3] focus on discovering key concepts from noun phrases in verbose queries with different weightings. Our work focuses on how to formulate appropriate queries by selecting effective terms or dropping ineffective ones. No weight assignments are needed and thus conventional retrieval models could be easily incorporated. [4] uses a supervised learning method for selecting good expansion terms from a number of candidate terms generated by pseudo-relevance feedback technique. However, we differ in that, (1) [4] selects specific features so as to emphasize more on the relation between original query and expansion terms without consideration of linguistic features, and (2) our approach does not introduce extra terms for query formulation. Similarly, [10] attempts to predict which words in query should be deleted based on query logs. Moreover, a number of works [2,5,6,7,9,15,16,18,19,20] pay attention to predict the quality or difficulty of queries, and [11,12] try to find optimal sub-queries by using maximum spanning tree with mutual information as the weight of each edge. However, their focus is to evaluate performance of a whole query whereas we consider units at the level of terms.

Given a set of possible query terms that a user may use to search documents relevant to a topic, the goal of this paper is to formulate appropriate queries by selecting effective terms from the set. Since exhaustively examining all candidate subsets is not feasible in a large scale, we reduce the problem to a simplified one that iteratively selects effective query terms from the set. We are interested in realizing (1) what characteristic of a query term makes it effective or ineffective in search, and (2) whether or not the effective query terms (if we are able to predict) can improve IR performance. We propose an approach to automatically measure the effectiveness of query terms in IR, wherein a regression model learned from training data is applied to conduct the prediction of term effectiveness of testing data. Based on the measurement, two algorithms are presented, which formulate queries by selecting effective terms and dropping ineffective terms from the given set, respectively.

The merit of our approach is that we consider various aspects that may influence retrieval performance, including linguistic properties of a query term and statistical relationships between terms in a document collection such as co-occurrence and context dependency. Their impacts on IR have been carefully examined. Moreover, we have conducted extensive experiments on NTCIR-4 and NTCIR-5 ad-hoc IR tasks to evaluate the performance of the

proposed approach. Based on term effectiveness prediction and two query formulation algorithms, our method significantly improve MAP by 9.2% on average, compared to the performance of the original queries given in the benchmarks.

In the rest of this paper, we describe the proposed approach to term selection and query formulation in Section 2. The experimental results of retrieval performance are presented in Sections 3. Finally, in Section 4, we give our discussion and conclusions.

2. Term Selection Approach for Query Formulation

2.1 Observation

When a user desires to retrieve information from document repositories to know more about a topic, many possible terms may come into the mind to form various queries. We call such set of the possible terms *query term space* $T=\{t_1, ..., t_n\}$. A query typically consists of a subset of T. Each query term $t_i \in T$ is expected to convey some information about the user information need. It is, therefore, reasonable to assume that each query term will have different degree of effectiveness in retrieving relevant documents. To explore the impact of one query term on retrieval performance, we start the discussion with a degeneration process, which is defined as a mapping function taking the set of terms T as input and producing set $\{T-\{t_1\}, T-\{t_2\}, ..., T-\{t_n\}\}$ as output. Mathematically, the mapping function is defined as:

$$DeGen(T) = \{T - \{x\} | x \in T\}.$$

By applying the degeneration process to the given n terms in T, we can construct a set of n queries $\Delta q = \{\Delta q_1, \Delta q_2, ..., \Delta q_i, ..., \Delta q_n\}$, where $\Delta q_i = \{t_1, ..., t_{i-1}, t_{i+1}, ..., t_n\}$ stands for a query by removing t_i from original terms T.

Suppose query term space T well summaries the description of the user information need. Intuitively, we believe that the removal of a term (especially an important one) from T may result in a loss of information harming retrieval effectiveness. To realize how much such information loss may influence IR performance, we conduct an experiment on NTCIR-4 description queries. For each query, we construct its query term space T by dropping stop words. T is treated as a hypothetical user information need. The remaining terms in the description queries are individually, one at a time, selected to be removed to obtain Δq. Three formulas are used to measure the impact of the removing terms and defined as:

$$g_{\min}(T) = \min_{\Delta q_i \in \Delta q} (pf(\Delta q_i) - pf(T))/pf(T)$$

$$g_{\max}(T) = \max_{\Delta q_i \in \Delta q} (pf(\Delta q_i) - pf(T))/pf(T)$$

$$g_{avg}(T) = \frac{1}{|T|}\sum_i (pf(\Delta q_i) - pf(T))/pf(T)$$

where $pf(x)$ is a performance measurement for query x, $g(T)$ computes the ratio of performance variation, which measures the maximum, minimum and average performance gain due to the removal of one of the terms from T, and $|T|$ is the number of query terms in T.

We use Okapi as the retrieval model and mean average precision (MAP) as our performance measurement for $pf(x)$ in this experiment.

The experimental results are shown in Figure 1. When we remove one term from each of the 50 topics $\{T\}$, in average, 46 topics have negative influence, i.e., $g_{avg}(T)<0$. This means that deleting one term from T mostly leads to a negative impact on MAP, compared to original T. On the other hand, $g_{max}(T)>0$ shows that at least the removal of one term positively improves MAP. By removing such terms we can obtain better performance. The phenomenon appears in 35 out of 50 topics, which is statistically suggestive that there exists noisy terms in most of user-constructed queries. In short, removing different terms from each topic T causes MAP variation in different levels. Some query terms are highly information-bearing, while others might hurt MAP. It is worth mentioned that we conduct the same experiment with the Indri and TFIDF retrieval models using the Lemur toolkit [21]. The results are quite consistent over different models. This characteristic makes it possible for the effectiveness of a query term on IR to be learned and applied to query formulation.

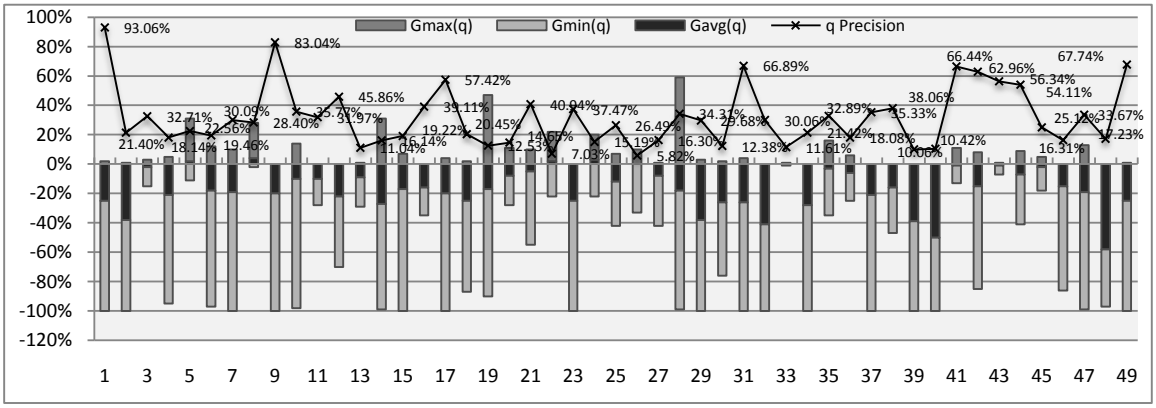


Fig. 1. MAP gain by removing terms from original NTCIR-4 description queries.

2.2 Problem Specification

When a user desires to retrieve information from document repositories to know more about a topic, many possible terms may come into her mind to form various queries. We call such set of the possible terms *query term space* $T=\{t1, ..., tn\}$. A query typically consists of a subset of T. Each query term $ti \in T$ is expected to convey some information about the user's information need. It is, therefore, reasonable to assume that each query term will have different degree of effectiveness in documents retrieval. Suppose Q denotes all subsets of T, that is, $Q=Power\ Set(T)$ and $|Q|=2^n$. The problem is to choose the best subset Δq^* among all candidates Q such that the performance gain between the retrieval performance of T and Δq ($\Delta q \in Q$) is maximized:

$$\Delta q^* = argmax_{\Delta q \in Q}\{(pf(T) - pf(\Delta q))/pf(T)\}. \quad (1)$$

where $pf(x)$ denotes a function measuring retrieval performance with x as the query. The higher the score $pf(x)$ is, the better the retrieval performance can be achieved.

An intuitive way to solve the problem is to exhaustively examine all candidate subset members in Q and design a method to decide which the best Δq^* is. However, since an exhaustive search is not appropriate for applications in a large scale, we reduce the problem

to a simplified one that chooses the most effective query term t_i ($t_i \in T$) such that the performance gain between T and $T-\{t_i\}$ is maximized:

$$t_i^* = argmax_{t_i \in T}\{(pf(T) - pf(T - \{t_i\}))/pf(T)\}. \quad (2)$$

Once the best t_i^* is selected, Δq^* could be approximated by iteratively selecting effective terms from T. Similarly, the simplified problem could be to choose the most ineffective terms from T such that the performance gain is minimized. Then Δq^* will be approximated by iteratively removing ineffective or noisy terms from T.

Our goals are: (1) to find a function $r: T \rightarrow R$, which ranks $\{t_1, ..., t_n\}$ based on their effectiveness in performance gain (MAP is used for the performance measurement in this paper), where the effective terms are selected as candidate query terms, and (2) to formulate a query from the candidates selected by function r.

2.3 Effective Term Selection

To rank term t_i in a given query term space T based on function r, we use a regression model to compute r directly, which predicts a real value from some observed features of ti. The regression function $r: T \rightarrow R$ is generated by learning from each t_i with the examples in form of $<f(t_i), (pf(T) - pf(T - \{t_i\}))/pf(T)>$ for all queries in the training corpus, where $f(t_i)$ is the feature vector of t_i, which will be described in Section 2.5.

The regression model we adopt is Support Vector Regression (SVR), which is a regression analysis technique based on SVM [17]. The aim of SVR is to find the most appropriate hyperplane w which is able to predict the distribution of data points accurately. Thus, r can be interpreted as a function that seeks the least dissimilarity between ground truth $y_i = (\mathrm{pf(T)} - \mathrm{pf(T} - \{\mathrm{t_i}\}))/\mathrm{pf(T)}$ and predicted value $r(t_i)$, and r is required to be in the form of $\mathbf{w}\ f(t_i)+b$. Finding function r is therefore equivalent to solving the convex optimization problem:

$$Min_{w,b,\xi_{i,1},\xi_{i,2}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i(\xi_{i,1} + \xi_{i,2}). \quad (3)$$

subject to:

$$\forall\, t_i \in T \qquad y_i - (\mathbf{w}\ f(t_i)+b) \geq \varepsilon + \xi_{i,1} \quad (4)$$

$$\forall\, i:\ \xi_{i,1}, \xi_{i,2} \geq 0 \qquad (\mathbf{w}\ f(t_i)+b) - y_i \geq \varepsilon + \xi_{i,2}. \quad (5)$$

where C determines the tradeoff between the flatness of r and the amount up to which deviations larger than ε are tolerated, ε is the maximum acceptable difference between the predicted and actual values we wish to maintain, and $\xi_{i,1}$ and $\xi_{i,2}$ are slack variables that cope with otherwise infeasible constraints of the optimization problem. We use the SVR implementation of LIBSVM [8] to solve the optimization problem.

Ranking terms in query term space $T=\{t_1, ..., t_n\}$ according to their effectiveness is then equivalent to applying regression function to each t_i; hence, we are able to sort terms $t_i \in T$ into an ordering sequence of effectiveness or ineffectiveness by $r(t_i)$.

2.4 Generation and Reduction

Algorithms *Generation* and *Reduction*, as shown in Fig. 2, formulate queries by greedily

selecting effective terms or dropping ineffective terms from space T based on function r.

When formulating a query from query term space T, the Generation algorithm computes a measure of effectiveness $r(t_i)$ for each term $t_i \in T$, includes the most effective term t_i^* and repeats the process until k terms are chosen (where k is a empirical value given by users). Note that T is changed during the selection process, and thus statistical features should be re-estimated according to new T. The selection of the best candidate term ensures that the current selected term t_i^* is the most informative one among those that are not selected yet.

Compared to generation, the Reduction algorithm always selects the most ineffective term from current T in each iteration. Since users may introduce noisy terms in query term space T, Reduction aims to remove such ineffective terms and will repeat the process until $|T|-k$ terms are chosen.

Algorithm Generation	Algorithm Reduction
Input: $T=\{t_1,t_2,...,t_n\}$ (query term space)	Input: $T=\{t_1,t_2,...,t_n\}$ (query term space)
k (# of terms to be selected)	k (# of terms to be selected)
$\Delta q \leftarrow \{\ \}$	$\Delta q \leftarrow \{\ t_1,t_2,...,t_n\ \}$
for i = 1 to k do	for i = 1 to n-k do
$t_i^* \leftarrow argmax_{t_i \in T}\{r(t_i)\}$	$t_i^* \leftarrow argmin_{t_i \in T}\{r(t_i)\}$
$\Delta q \leftarrow \Delta q \cup \{t_i^*\}$	$\Delta q \leftarrow \Delta q - \{t_i^*\}$
$T \leftarrow T - \{t_i^*\}$	$T \leftarrow T - \{t_i^*\}$
end	end
Output Δq	Output Δq

Fig. 2. The Generation Algorithm and the Reduction Algorithm

2.5 Features Used for Term Selection

Linguistic and statistical features provide important clues for selection of good query terms from viewpoints of users and collections, and we use them to train function r.

Linguistic Features: Terms with certain linguistic properties are often viewed semantics-bearing and informative for search. Linguistic features of query terms are mainly inclusive of parts of speech (POS) and named entities (NE). In our experiment, the POS features comprise noun, verb, adjective, and adverb, the NE features include person names, locations, organizations, and time, and other linguistic features contain acronym, size (i.e., number of words in a term) and phrase, all of which have shown their importance in many IR applications. The values of these linguistic features are binary except the size feature. POS and NE are labeled manually for high quality of training data, and can be tagged automatically for purpose of efficiency alternatively.

Statistical Features: Statistical features of term t_i refer to the statistical information about the term in a document collection. This information could be about the term itself such as term frequency (TF) and inverse document frequency (IDF), or the relationship between the term and other terms in space T. We present two methods for estimating such term relationship. The first method depends on co-occurrences of terms t_i and t_j ($t_j \in T$, $t_i \neq t_j$) and co-occurrences of terms t_i and $T-\{t_i\}$ in the document collection. The former is called *term-term co-occur feature* while the latter is called term-topic co-occur feature. The second method extracts so-called context vectors as features from the search results of t_i, t_j, and $T-\{t_i\}$, respectively. The *term-term context feature* computes the similarity between the context vectors of t_i and t_j while the *term-topic context feature* computes the similarity

between context vectors of t_i and T-$\{t_i\}$.

Term-term & term-topic co-occur features: The features are used to measure whether query term t_i itself could be replaced with another term t_j (or remaining terms T-$\{t_i\}$) in T and how much the intension is. The term without substitutes is supposed to be important in T.

Point-wise mutual information (PMI), Chi-square statistics (X^2), and log-likelihood ratio (LLR) are used to measure co-occurrences between t_i and Z, which is either t_j or T-$\{t_i\}$ in this paper. Suppose that N is the number of documents in the collection, a is the number of documents containing both t_i and Z, denoted as a = #d(t_i,Z). Similarly, we denote b = #d(t_i,~Z) c = #d(~t_i,Z) and d = #d(~t_i,~Z) i.e., Z=N-a-b-c.

PMI is a measure of how much term t_i tells us about Z.

$$\mathrm{PMI}(\mathrm{t_i}, \mathrm{Z}) = \log[\mathrm{p}(\mathrm{t_i}, \mathrm{Z})/\mathrm{p}(\mathrm{t_i})\mathrm{p}(\mathrm{Z})] \approx \log[\mathrm{a} \times \mathrm{N}/(\mathrm{a}+\mathrm{b})(\mathrm{a}+\mathrm{c})] \tag{6}$$

X^2 compares the observed frequencies with frequencies expected for independence.

$$\chi^2(\mathrm{t_i}, \mathrm{Z}) = [\mathrm{N} \times (\mathrm{a} \times \mathrm{d} - \mathrm{b} \times \mathrm{c})^2]/[(\mathrm{a}+\mathrm{b})(\mathrm{a}+\mathrm{c})(\mathrm{b}+\mathrm{d})(\mathrm{c}+\mathrm{d})] \tag{7}$$

LLR is a statistical test for making a decision between two hypotheses of dependency or independency based on the value of this ratio.

$$\begin{aligned} -2\log \mathrm{LLR}(\mathrm{t_i}, \mathrm{Z}) = \mathrm{a}\log\frac{\mathrm{a} \times \mathrm{N}}{(\mathrm{a}+\mathrm{b})(\mathrm{a}+\mathrm{c})} + \mathrm{b}\log\frac{\mathrm{b} \times \mathrm{N}}{(\mathrm{a}+\mathrm{b})(\mathrm{b}+\mathrm{d})} \\ +\mathrm{c}\log\frac{\mathrm{c} \times \mathrm{N}}{(\mathrm{c}+\mathrm{d})(\mathrm{a}+\mathrm{c})} + \mathrm{d}\log\frac{\mathrm{d} \times \mathrm{N}}{(\mathrm{c}+\mathrm{d})(\mathrm{b}+\mathrm{d})} \end{aligned} \tag{8}$$

We make use of average, minimum, and maximum metrics to diagnose term-term co-occur features over all possible pairs of (t_i,t_j), for any $t_j \neq t_i$:

$$f^X_{avg}(t_i) = \frac{1}{|T|}\sum_{\forall t_j \in T, t_i \neq t_j} X(t_i, t_j), \tag{9}$$

$$f^X_{max}(t_i) = \max_{\forall t_j \in T, t_i \neq t_j} \mathrm{X}(t_i, t_j) \;\&\; f^X_{min}(t_i) = \min_{\forall t_j \in T, t_i \neq t_j} \mathrm{X}(t_i, t_j) \tag{10}$$

where X is *PMI*, *LLR* or X^2. Moreover, given T={t_1, ..., t_n} as a training query term space, we sort all terms t_i according to their $f^X_{avg}(t_i)$, $f^X_{max}(t_i)$, or $f^X_{min}(t_i)$, and their rankings varied from 1 to n are treated the additional features.

The *term-topic co-occur features* are nearly identical to the *term-term co-occur features* with an exception that *term-topic co-occur features* are used in measuring the relationship between t_i and query topic T-$\{t_i\}$. The co-occur features can be quickly computed from the indices of IR systems with caches.

Term-term & term-topic context features: The co-occurrence features are reliable for estimating the relationship between high-frequency query terms. Unfortunately, term t_i is probably not co-occurring with T-$\{t_i\}$ in the document collection at all. The context features are hence helpful for low-frequency query terms that share common contexts in search results.

More specifically, we generate the context vectors from the search results of t_i and t_j (or T-$\{t_i\}$), respectively. The context vector is composed of a list of pairs <document ID, relevance score>, which can be obtained from the search results returned by IR systems. The relationship between t_i and t_j (or T-$\{t_i\}$) is captured by the cosine similarity between their context vectors. Note that to extract the context features, we are required to retrieve documents. The retrieval performance may affect the quality of the context features and the process is time-consuming.

3. Experiments

3.1 Experiment Settings

Table 1. Adopted dataset after data clean. Number of each setting is shown in each row for NTCIR-4 and NTCIR-5

	NTCIR-4	NTCIR-5
	<desc>	<desc>
#(query topics)	58	47
#(distinct terms)	865	623
#(terms/query)	14.9	13.2

Table 2. Number of training instances. (x : y) shows the number of positive and negative MAP gain instances are x and y, respectively

	Indri	TFIDF	Okapi
Original	674(156:518)	702(222:480)	687(224:463
Upsample	1036(518:51	960(480:480)	926(463:463
Train	828(414:414)	768(384:384)	740(370:370
Test	208(104:104)	192(96:96)	186 (93:93)

We conduct extensive experiments on NTCIR-4 and NTCIR-5 English-English ad-hoc IR tasks. Table 1 shows the statistics of the data collections. We evaluate our methods with description queries, whose average length is 14.9 query terms. Both queries and documents are stemmed with the Porter stemmer and stop words are removed. The remaining query terms for each query topic form a query term space T. Three retrieval models, the vector space model (TFIDF), the language model (Indri) and the probabilistic model (Okapi), are constructed using Lemur Toolkit [21], for examining the robustness of our methods across different frameworks. MAP is used as evaluation metric for top 1000 documents retrieved. To ensure the quality of the training dataset, we remove the poorly-performing queries whose average precision is below 0.02. As different retrieval models have different MAP on the same queries, there are different numbers of training and test instances in different models. We up-sample the positive instances by repeating them up to the same number as the negative ones. Table 2 summarizes the settings for training instances.

3.2 Performance of Regression Function

We use 5-fold cross validation for training and testing our regression function r. To avoid inside test due to up-sampling, we ensure that all the instances in the training set are different from those of the test set. The R^2 statistics ($R^2 \in [0, 1]$) is used to evaluate the prediction accuracy of our regression function r:

$$R^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad (11)$$

where R^2 explains the variation between true label $y_i=(pf(T) - pf(T - \{t_i\}))/pf(T)$ and fit value $\hat{y}_i=wf(t_i)+b$ for each testing query term $t_i \in T$, as explained in Section 2.2. $\bar{y}$ is the mean of the ground truth.

Table 3. R^2 of regression model r with multiple combinations of training features. L: linguistic features; C1: co-occurrence features; C2: context features

Performance of Regression Model r		One Group of Features			Two Groups of Features			Three	Four (3+1)		All
		L	C1	C2	L&C1	L&C2	C1&C2	L&C1 &C2	m-Cl	m-SCS	
R^2	Indri	0.120	0.145	0.106	0.752	0.469	0.285	0.975	0.976	0.975	0.976
	TFIDF	0.265	0.525	0.767	0.809	0.857	0.896	0.932	0.932	0.932	0.932
	Okapi	0.217	0.499	0.715	0.780	0.791	0.910	0.925	0.926	0.925	0.926
	Avg.	0.201	0.390	0.529	0.781	0.706	0.697	0.944	0.945	0.944	0.945

Table 3 shows the R^2 values of different combinations of features over different retrieval models, where two other features are taken into account for comparison. Content load (CI) [14] gives unequal importance to words with different POS. Our modified content load (m-Cl) sets weight of a noun as 1 and the weights of adjectives, verbs, and participles as 0.147 for IR. Our m-SCS extends the simplified clarity score (SCS) [9] as a feature by calculating the relative entropy between query terms and collection language models (unigram distributions).

It can be seen that our function r is quite independent of retrieval models. The performance of the statistical features is better than that of the linguistic features because the statistical features reflect the statistical relationship between query terms in the document collections. Combining both outperforms each one, which reveals both features are complementary. The improvement by m-Cl and m-SCS is not clear due to their similarity to the other features. Combining all features achieves the best R^2 value 0.945 in average, which guarantees us a large portion of explainable variation in y and hence our regression model r is reliable.

3.3 Correlation between Feature and MAP

Yet another interesting aspect of this study is to find out a set of key features that play important roles in document retrieval, that is, the set of features that explain most of the variance of function r. This task can usually be done in ways fully-addressed in regression diagnostics and subset selection, each with varying degrees of complexity. One common method is to apply correlation analysis over the response and each predictor, and look for highly-correlated predictor-response pairs.

Three standard correlation coefficients are involved, including Pearson's product-moment

correlation coefficient, Kendall's tau, and Spearman's rho. The results are given in Fig. 3, where x-coordinate denotes features and y-coordinate denotes the value of correlation coefficient. From Fig. 3, two context features, “cosine” and “cosineinc”, are found to be positively- and highly-correlated (ρ >0.5) with MAP, under Pearson's coefficient. The correlation between the term-term context feature (cosine) and MAP even climbs up to 0.8. For any query term, high context feature value indicates high deviation in the result set caused by removal of the term from the query topic. The findings suggest that the drastic changes incurred in document ranking by removal of a term can be a good predictor. The tradeoff is the high cost in feature computation because a retrieval processing is required. The co-occurrence features such as PMI, LLR, and χ^2 also behave obviously correlated to MAP. The minimum value of LLR correlates more strongly to MAP than the maximum one does, which means that the independence between query terms is a useful feature.

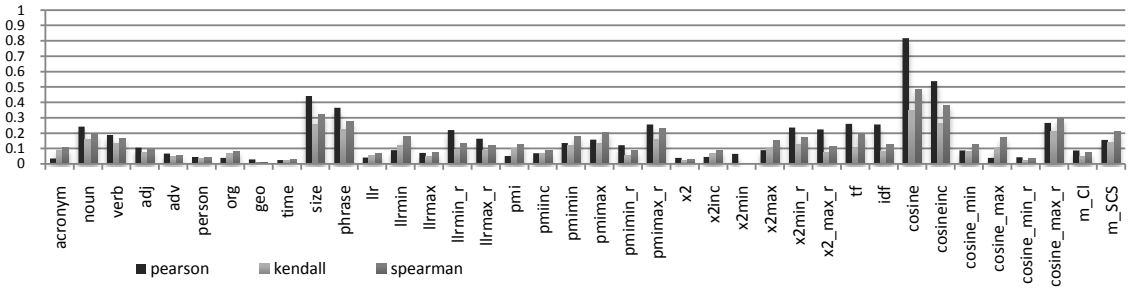


Fig. 3. Three correlation values between features and MAP on Okapi retrieval model

In the linguistic side, we find that two features “size” and “phrase” show positive, medium-degree correlation (0.3< ρ <0.5) with MAP. Intuitively, a longer term might naturally be more useful as a query term than a shorter one is; this may not always be the case, but generally it is believed a shorter term is less informative due to the ambiguity it encompasses. The same rationale also applies to “phrase”, because terms of noun phrases usually refer to a real-world event, such as “911 attack” and “4th of July”, which might turn out to be the key of the topic. We also notice that some features, such as “noun” and “verb”, pose positive influence to MAP than others do, which shows high concordance to a common thought in NLP that nouns and verbs are more informative than other type of words. To our surprises, NE features such as “person”, “geo”, “org” and “time” do not show as high concordance as the others. This might be resulted from that the training data is not sufficient enough. Features “idf” and “m-SCS” whose correlation is highly notable have positive impacts. It supports that the statistical features have higher correlation values than the linguistics ones.

3.4 Evaluation on Information Retrieval

In this section, we devise experiments for testing the proposed query formulation algorithms. The benchmark collections are NTCIR-4 and NTCIR-5. The experiments can be divided into

two parts: the first part is a 5-fold cross-validation on NTCIR-4 dataset, and in the second part we train the models on NTCIR-4 and test them on NTCIR-5. As both parts differ only in assignment of the training/test data, we will stick with the details for the first half (cross-validation) in the following text.

The result is given in Table 4. Evaluation results on NTCIR-4 and NTCIR-5 are presented in the upper- and lower-half of the table, respectively. We offer two baseline methods in the experiments: "BL1" puts together all the query terms into one query string, while "BL2" only consider nouns as query terms since nouns are claimed to be more informative in several previous works. Besides, the upper bound UB is presented in the benchmark: for each topic, we permute all sub queries and discover the sub-query with the highest MAP. As term selection can also be treated as a classification problem, we use the same features of our regression function r to train two SVM classifiers, Gen-C and Red-C. Gen-C selects terms classified as "effective" while Red-C removes terms classified as "ineffective". Gen-R and Red-R denote our Generation and Reduction algorithms, respectively. The retrieval results are presented in terms of MAP. Gain ratios in MAP with respect to the two baseline methods are given in average results. We use two-tailed t-distribution in the significance test for each method (against the BL1) by viewing AP values obtained in all query session as data points, with $p<0.01$ marked ** and $p<0.05$ marked *.

Table 4. MAP of baseline and multiple proposed methods on NTCIR-4 <desc> regression model. (+x, +y) shows the improvement percentage of MAP corresponding to BL1 and BL2. TFIDF and Okapi models have PRF involved, Indri model does not. Best MAP of each retrieval model is marked bold for both collections.

Settings	Metho	Indri	TFIDF	Okapi	Avg.
NTCIR-4 <desc> Queries	UB	0.2233	0.3052	0.3234	0.2839
	BL1	0.1742	0.2660	0.2718	0.2373
	BL2	0.1773	0.2622	0.2603	0.2332
	Gen-C	0.1949	0.2823	*0.2946*	0.2572(+8.38%,+10.2)
	Gen-R	0.1954	*0.2861*	0.2875	0.2563(+8.00%,+9.90)
	Red-C	0.1911*	0.2755	0.2854	0.2506(+5.60%,+7.46)
	Red-R	**0.1974**	0.2773	0.2797	0.2514(+5.94%,+7.80)
NTCIR-5 <desc> Queries	UB	0.1883	0.2245	0.2420	0.2182
	BL1	0.1523	0.1988	0.1997	0.1836
	BL2	0.1543	0.2035	0.1969	0.1849
	Gen-C	0.1699	0.2117*	0.2213	0.2009(+9.42%,+8.65)
	Gen-R	0.1712	*0.2221*	*0.2232*	0.2055(+11.9%,+11.1)
	Red-C	0.1645	0.2194	0.2084	0.1974(+7.51%,+6.76)
	Red-R	**0.1749**	0.2034	0.2160	0.1981(+7.89%,+7.13)

From Table 4, the MAP difference between two baseline methods is small. This might be because some nouns are still noisy for IR. The four generation and reduction methods

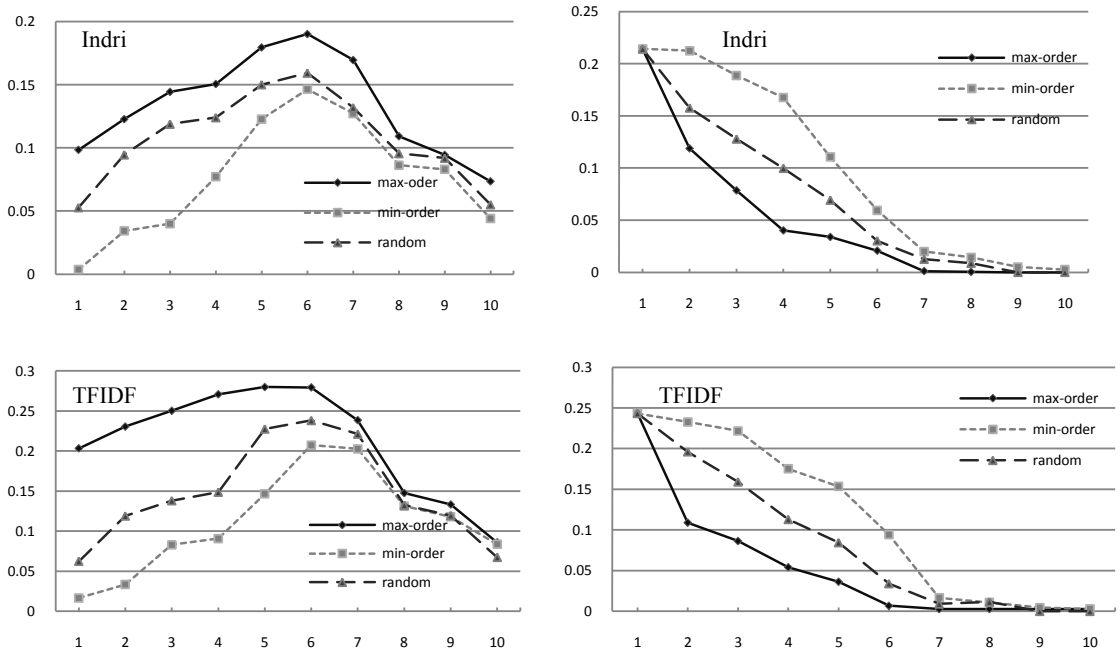
significantly outperform the baseline methods. We improve the baseline methods by 5.60% to 11.9% in the cross-validation runs and on NTCIR-5 data. This result shows the robustness and reliability of the proposed algorithms. Furthermore, all the methods show significant improvements when applied to certain retrieval models, such as Indri and TFIDF; performance gain with Okapi model is less significant on NTCIR-5 data, especially when reduction algorithm is called for. The regression methods generally achieve better MAP than the classification methods. This is because the regression methods always select the most informative terms or drop the most ineffective terms among those that are not selected yet. The encouraging evaluation results show that, despite the additional costs on iterative processing, the performance of the proposed algorithms is effective across different benchmark collections, and based on a query term space T, the algorithms are capable of suggesting better ways to form a query.

[4] proposed a method for selecting Good Expansion Terms (GET) based on an SVM classifier. Our approach is also applicable to selection of query expansion terms. Given the same set of candidate expansion terms which are generated by conventional approaches such as TF and IDF, GET-C runs the Gen-C method whereas GET-R runs the Gen-R on the expansion set (with the NTCIR-4 5-fold cross validation regression model). Table 5 shows the MAP results of the two methods and the baseline method (BL), which adds all expansion terms to original queries. From Table 5, GET-R outperforms GET-C under different retrieval models and data sets, and both methods improve MAP by 1.76% to 3.44% compared to the baseline. Moreover, though extra terms are introduced for query formulation, we can see that certain MAP results in Table 4 still outperform those in Table 5 (marked *italic*). It is therefore inferred that, it is still important to filter out noisy terms in original query even though good expansion terms are selected. Finally, note that we use the NTCIR-4 5-fold cross validation regression model, which is trained to fit the target performance gain in NTCIR-4 dataset, rather than instances in the query expansion terms set. However, results in Table 5 show that this model works satisfactorily in selection of good expansion terms, which ensures that our approach is robust in different environments and applications such as query expansion.

Table 5. MAP of query expansion based on GET-C and GET-R model. (%) shows the improvement percentage of MAP to BL. Significance test is tested against the baseline results.

Settings	Method	Indri	TFIDF	Okapi	Avg.
NTCIR-4	BL	0.2470	0.2642	0.2632	0.2581
<desc>	GET-C	0.2472**	0.2810**	0.2728**	0.2670
	GET-R	*0.2610***	0.2860**	0.2899**	0.2789
NTCIR-5	BL	0.1795	0.1891	0.1913	0.1866
<desc>	GET-C	0.1868	0.1904	0.1927	0.1899
	GET-R	*0.1880**	0.1918*	0.1945*	0.1914

We further investigate the impact of various ranking schemes based on our proposed algorithms. The ranking scheme in the Generation algorithm (or the Reduction algorithm) refers to an internal ranking mechanism that decides which term shall be included in (or discarded away). Three types of ranking schemes are tested based on our regression function *r*. "max-order" always returns the term that is most likely to contribute relevance to a query topic, "min-order" returns the term that is most likely to bring in noise, and "random-order" returns a randomly-chosen term. Figure 4 shows the MAP curve for each scheme by connecting the dots at (1, $MAP^{(1)}$), … , (n, $MAP^{(n)}$), where $MAP^{(i)}$ is the MAP obtained at iteration *i*. It tells that the performance curves in the generation process share an interesting tendency: the curves keep going up in first few iterations, while after the maximum (locally to each method) is reached, they begin to go down rapidly. The findings might informally establish the validity of our assumption that a longer query topic might encompass more noise terms. The same "up-and-down" pattern does not look so obvious in the reduction process; however, if we take the derivative of the curve at each iteration *i* (i.e., the performance gain/loss ratio), we might find it resembles the pattern we have discovered. We may also find that, in the generation process, different ranking schemes come with varying degrees of MAP gains. The ranking scheme "max-order" constantly provides the largest performance boost, as opposed to the other two schemes. In the reduction process, "max-order" also offers the most drastically performance drop than the other two schemes do. Generally, in the generation process, the best MAP value for each setting might take place somewhere between iteration n/2 to 2n/3, given *n* is the size of the query topic.



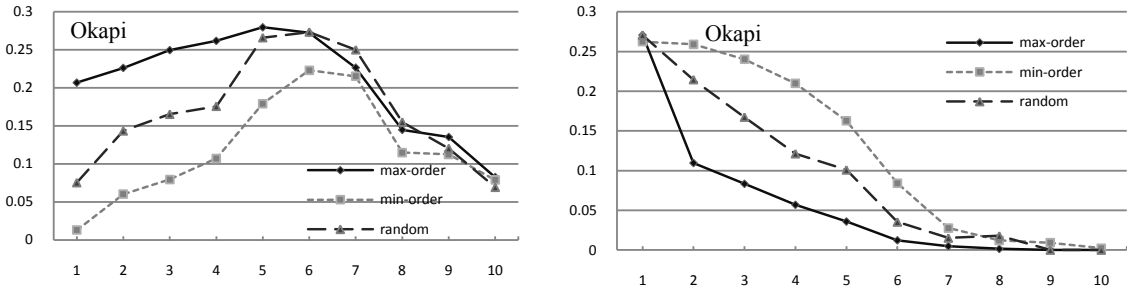


Fig. 4. MAP curves based on regression model for description queries of NTCIR-4 on Indri, TFIDF, and Okapi models, each with three selection order. X coordinate is # of query terms; Y coordinate is MAP.

4. Discussions and Conclusions

In this paper, we propose an approach to measure and predict the impact of query terms, based on the discovery of linguistic, co-occurrence, and contextual features, which are analyzed by their correlation with MAP. Experimental results show that our query formulation approach significantly improves retrieval performance.

The proposed method is robust and the experimental results are consistent on different retrieval models and document collections. In addition, an important aspect of this paper is that we are able to capture certain characteristics of query terms that are highly effective for IR. Aside from intuitive ideas that informative terms are often lengthy and tagged nouns as their POS category, we have found that the statistical features are more likely to decide the effectiveness of query terms than linguistics ones do. We also observe that context features are mostly correlated to MAP and thus are most powerful for term difficulty prediction. However, such post-retrieval features require much higher cost than the pre-retrieval features, in terms of time and space.

The proposed approach actually selects local optimal query term during each iteration of generation or reduction. The reason for this greedy algorithm is that it is inappropriate to exhaustively enumerate all sub-queries for online applications such as search engines. Further, it is challenging to automatically determine the value of parameter k in our algorithms, which is selected to optimize the MAP of each query topic. Also, when applying our approach to web applications, we need web corpus to calculate the statistical features for training models.

5. References

[1] Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Broglio, J., Xu, J., Shu, H.: INQUERY at TREC-5. In: Fifth Text REtrieval Conference (TREC-5), pp. 119--132 (1997)

[2] Amati, G., Carpineto, C., Romano, G.: Query Difficulty, Robustness, and Selective Application of Query Expansion. In: 26th European Conference on IR Research, UK (2004)

[3] Bendersky M., Croft, W. B.: Discovering key concepts in verbose queries. In: 31st annual international ACM SIGIR conference on Research and development in information retrieval,

pp. 491--498 (2008)

[4] Cao, G., Nie, J. Y., Gao, J. F., & Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 243--250 (2008)

[5] Carmel, D., Yom-Tov, E., Soboroff, I.: SIGIR WORKSHOP REPORT: Predicting Query Difficulty - Methods and Applications. WORKSHOP SESSION: SIGIR, pp. 25--28 (2005)

[6] Carmel, D., Yom-Tov, E., Darlow, A., Pelleg, D.: What makes a query difficult? In: 29th annual international ACM SIGIR, pp. 390--397 (2006)

[7] Carmel, D., Farchi, E., Petruschka, Y., Soffer, A.: Automatic Query Refinement using Lexical Affinities with Maximal Information Gain. In: 25th annual international ACM SIGIR, pp. 283--290 (2002)

[8] Chang, C. C., Lin, C. J.: LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001)

[9] He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: 11th International Conference of String Processing and Information Retrieval, pp. 43--54 (2004)

[10] Jones, R., Fain, D. C.: Query Word Deletion Prediction. In: 26th annual international ACM SIGIR, pp. 435--436 (2003)

[11] Kumaran, G., Allan, J.: Effective and efficient user interaction for long queries. In: 31st annual international ACM SIGIR, pp. 11--18 (2008)

[12] Kumaran, G., Allan, J.: Adapting information retrieval systems to user queries. In: Information Processing and Management, pp. 1838-1862 (2008)

[13] Kwok, K., L.: A New Method of Weighting Query Terms for Ad-hoc Retrieval. In: 19th annual international ACM SIGIR, pp. 187--195 (1996)

[14] Lioma, C., Ounis, I.: Examining the Content Load of Part of Speech Blocks for Information Retrieval. In: COLING/ACL 2006 Main Conference Poster Sessions (2006)

[15] Mandl,T., Womser-Hacker, C.: Linguistic and Statistical Analysis of the CLEF Topics. In: Third Workshop of the Cross-Language Evaluation Forum CLEF (2002)

[16] Mothe, J., Tanguy, L: ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications (2005)

[17] Vapnik, V. N.: Statistical Learning Theory. John Wiley & Sons (1998)

[18] Yom-Tov, E., Fine, S., Carmel, D., Darlow, A., Amitay, E.: Juru at TREC 2004: Experiments with Prediction of Query Difficulty. In: 13th Text Retrieval Conference (2004)

[19] Zhou, Y., and Croft, W. B.: Query Performance Prediction in Web Search Environments. In: 30th Annual International ACM SIGIR Conference, pp. 543--550 (2007)

[20] Zhou, Y., Croft, W. B.: Ranking Robustness: A Novel Framework to Predict Query Performance. In: 15th ACM international conference on Information and knowledge management, pp. 567--574 (2006)

[21] The Lemur Toolkit: http://www.lemurproject.org/

中英文專利文書之文句對列[1]

田侃文
國立政治大學
資訊科學系
96753027@nccu.edu.tw

曾元顯
國立臺灣師範大學
資訊中心
samtseng@ntnu.edu.tw

劉昭麟
國立政治大學
資訊科學系
chaolin@nccu.edu.tw

摘要

綜觀今日全球化的趨勢，世界各國皆進行跨語言的專利說明書翻譯工作。在中文專利說明書中譯英方面，為了追求更精確的翻譯品質，蒐集大量且正確的專利文書平行語料，能夠協助輔助式機器翻譯及資訊檢索的研究工作進行。因此本研究便希望利用中英技術名詞對應表，透過統計詞頻調整詞對應權重及計算中英文句子的向量相似度等機制，搭配動態規劃演算法，計算中英專利文書的句對相似度，最後產生對列結果。以精確率、召回率及輔助式機器翻譯系統，評比本對列系統的對列成效。實驗結果顯示本系統不僅在 1：1 對列模式[2]的精確率達到 0.995，且產出的大量中英文句對確實能夠提升輔助式機器翻譯系統的翻譯品質。

關鍵詞：專利說明書、電腦輔助機器翻譯、文句對列、動態規劃演算法

1. 緒論

在這資訊爆炸的時代，科技產業日新月異，因此在開發一項新的產品前，對於專利說明書的熟讀，更顯得格外重要，以避免在研發過程中，侵犯了他人的智慧財產權。近年來，中國大陸的經濟蓬勃發展，外商投資市場急速拓展，因此外語使用者對於中文專利說明書的查詢需求量便大幅增加。於 2007 年 10 月在拉脫維亞舉行的歐洲專利局 (European Patent Office，簡稱 EPO) 專利資訊會議，舉辦的二場主題皆與中文專利文書翻譯密切相關，顯示中文專利文書英譯的議題越來越被重視。

想要有高品質的專利文書翻譯效果，必須倚靠人力進行翻譯，不僅在時間及成本上的花費極高，在數量上也有所限制。為了在產品的生產前置作業上，減少跨語言專利查詢的成本，發展一套專利說明書英譯或是檢索系統，便成了刻不容緩的事情，不僅能方便使用者的查詢，更能間接促進科技產業的發展。

無論是進行專利文書翻譯或檢索研究，大量且正確的平行語料是不可或缺的。在跨語言資訊檢索研究中，針對查詢的問句進行翻譯有許多種策略，如以辭典為本(dictionary-based) 、索引典為本 (thesaurus-based) 等歸類於以知識為本 (knowledge-based) 的策略及語料庫為本 (corpus-based) 等策略[17]。其中以語料庫為本的策略，將大量對列好的雙語文句，透過計算詞彙對譯的強度建構詞彙對應表，利用此表，便能在使用者進行查詢時進行詞彙的翻譯動作[6]。而在輔助式機器翻譯的領域中，能夠利用這些對列好的雙語文句進行翻譯模型的訓練及其它相關的處理，供後續翻譯系統使用。

想要得到平行語料進行研究並非難事，為了保證語料本身的品質，除了利用人工對列的方式進行外，利用簡單且兼顧正確性的方法，如：僅利用雙語辭典進行詞彙比對，訂定保險的篩選門檻，也能夠產生平行的語料。但利用人工的方式進行對列需要大量的人力成本，而僅利用詞彙比對的方法為了兼顧正確性，往往得到數量較少的對列結果產出，仍然不敷成本效益。因此我們希望能夠改進現有的文句對列技術，透過加入專業領域的辭典及利用自然語言處理的方法，發展出一套適用於專利文書文本的文句對列系統。

為了統一詞彙的用法，以下所稱「英漢辭典」指的是本系統採用的英漢辭典；「原中文詞義」指的是英漢辭典未合併中文近義詞表的中文詞義，在中文近義詞表合併至英漢辭典後，我們統稱英漢辭典內的中文詞義及其中文近義詞為「中文詞義」；擷取出英文片語及技術名詞的步驟，簡稱為「詞彙擷取」；「英文詞彙」指的是英漢辭典及「中英技術名詞對應表」的英文詞，

[1] 本論文另有 25 頁的版本，請參照連結：http://www.cs.nccu.edu.tw/~chaolin/papers/rocling_tien.pdf。
[2] 例如：對列模式「1：1」代表一句來源語言句子對應至一句目標語言句子。

或經過詞彙擷取後，英文句子裡的片語、英文技術名詞及句子裡其它利用空白隔開的英文單詞，這些英文詞彙皆由一個以上連續的英文單詞組成，如：「of course」，我們稱為一個英文詞彙，由兩個連續的英文單詞組成，「Of course, I will help you.」，我們稱這英文句子內有五個英文詞彙：「Of course」、「I」、「will」、「help」及「you」；「詞形還原」(lemmatization) 指的是將英文詞彙，還原成其原形的處理步驟，例如：將「ate」還原成「eat」；「對應」一詞，指的是中英文互為翻譯的文章、句子或詞彙，例：「中英文對應句子」指的是中英文互為翻譯的句子；「句對」指的是完成對列後產生的中英文對應句子；「詞對」指的是在對列過程中，中英文句子在詞彙比對時，比對到的中英對應詞彙。而中英文句子在完成對列後，稱為「句對」，該句對包含的一組以上的詞對便形成「詞對集合」。

我們發展的系統（以下簡稱本系統）在前處理步驟中，利用簡單的規則對未經過段落對列處理 (paragraph alignment) 的中英文章進行斷句，透過 StanfordLexParser-1.6 進行英文詞彙的詞形還原，再利用詞彙量豐富的英漢辭典及中英技術名詞對應表，在中文的部份以長詞優先的技術進行斷詞，在英文的部份則進行詞彙擷取及片語保留的動作。為了增加中英文詞彙比對的成功率，透過 HowNet 及中研院現代漢語一詞泛讀系統尋找英漢辭典中的原中文詞義的近義詞彙，拓展英漢辭典的規模，並利用自「國立編譯館學術名詞資訊網」擷取、整理並建構的中英技術名詞對應表，讓本系統能適用於專利文書的文本。在進行各模式中英文句子的詞彙比對部份，我們沿用 Ma[15]於 2006 年提出的詞彙權重計算方法及其採用的動態規劃演算法。在對列的過程中，給予比對的來源句 (source sentence) 及目標句 (target sentence) 詞彙比對的分數，再利用衍生自餘弦相似度 (cosine similarity)[16]的計算原理，計算得到中英文句子之間的向量相似度分數，作為句子相似度的輔助評分權重，之後利用動態規劃演算法，計算得到整體相似度分數最高的對列組合，產生對列的結果。我們同時訂定篩選門檻，以句對平均比對詞數及向量相似度分數作為篩選條件，篩選出「1：1 信心句對」作為平行語料。

2. 文獻探討

跨語言文句對列的技術有很多種，發展至今，已有許多學者提出不同的方法，有針對來源語言 (source language) 及目標語言 (target language) 屬於同一語系(如：英文及法文同屬於印歐語系）的對列技術，也有跨語系的相關對列技術研究，如針對英日翻譯、英漢翻譯等翻譯文章。

近年來利用英漢辭典比對進行文句對列的方法，有[15]提出的「Champollion」這套對列工具。他認為任意兩個句子中的多個對應詞彙，並不應該一律給予相同的權重，於是透過修改 *tf-idf* [18]的權重計算公式，進行詞彙權重的調整，並搭配懲罰機制，依照不同的對列模式及句子長度差異進行扣分的計算。他同樣採用動態規劃演算法進行最佳對列模式的選擇，挑選出整體相似度分數最高的句對組合。

2007 年，Utiyama 與 Isahara[19]提出一套英日的專利平行語料庫，並參加第六屆 NTCIR 專利檢索 (patent retrieval) 的比賽。其語料庫包含了約 199 萬組經自動對列技術得到的英日句對。如此龐大的句對數量，必須倚賴自動對列技術才能蒐集完成。他們在專利說明書中發現「發明說明」段落所包含的兩個小段落：「先前技術」及「實施方法」翻譯較為整齊，因此選擇這兩個小段落進行對列處理。

在對列的方法上，他們利用英日及日英的辭典，搭配動態規劃演算法，對不同對列模式的句子進行相似度分數計算，接著再計算整篇文章的相似度總分及總句數的比例，最後選取在句子層級、文章層級及句數比例最為接近的對列結果，完成對列後約產生 700 萬組句對。接著他們對這些句對進行篩選，首先取出對列模式 1：1 的句對，再依照相似度分數進行排序，只取出以句號結尾的句對，同時過濾重複出現的句對，剩下約 390 萬組。為了保證這些句對的品質，他們進行區段抽樣的實驗，分析後決定取出相似度分數最高的前 200 萬組句對進行最後的篩選，去除句長太長以及長度差異太大的句對後，最後剩下約 199 萬組句對，他們便以這些句對建構專利文書的平行語料庫。除了提出文句對列及句對篩選的方法，他們在詞彙的層級將句對檢驗的評比分為三種等級：句對中的詞彙完全比對成功、50%以上比對成功及 50%以下的詞彙成功比對等三個等級；在語意的層級分成四個等級：句對的語意完全符合、80%以上的語意符合、80%以下的語意符合及語意完全不符合等四個等級。他們將這些專利文書語料依照國際專利分類號 (International Patent Classification，簡稱 IPC) 作分類，對詞彙及句長進行統計，並利用輔助式機器翻譯系統及翻譯指標進行翻譯效果的評比，證明該平行語料庫確實能夠勝任作為輔助式機器翻譯的訓練語料。

表一、語料來源、範圍與文章總數量統計

語料	文章數	範圍	來源
專利文書公開全文	7284	申請號 091132651 至 095145895	經濟部智慧財產局網站
專利文書公告全文	13675	申請號 084111615 至 096222166	經濟部智慧財產局網站
專利文書公告全文 摘要段落	417846	申請號 075102826 至 096213764	經濟部智慧財產局網站
科學人雜誌 中英對照電子書	2065	2002 年 1 月至 2009 年 1 月	國立政治大學圖書館
雙語網站 知識管理平台新聞	737	2005 年 8 月 30 日 至 2007 年 12 月 15 日	官方網站
自由時報 中英對照讀新聞	1553	2005 年 2 月 14 日 至 2009 年 5 月 27 日	官方網站
大考試題	131	2004 年至 2009 年	官方網站

表二、本研究採用訓練語料文章數及句數統計

語料	文章數	語言	總句數
專利文書公開全文	2271	中文	695326
		英文	518482
科學人雜誌中英對照電子書	319	中文	18966
		英文	19282

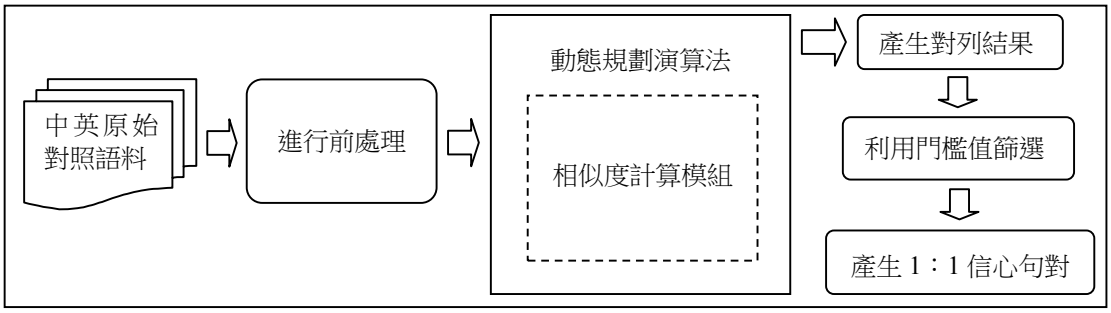
3. 語料來源

在專利說明書方面，本研究所需的語料來源，我們透過對從經濟部智慧財產局[9]擷取的資料整理及過濾，有中英文互為翻譯的「專利文書公開全文」及「專利文書公告全文」共約二萬篇及「專利文書公告全文摘要段落」約 42 萬篇(以下統稱為「專利文書的文本」)；在科普文學方面，我們有「科學人雜誌中英對照電子書」[5]從 2002 年 3 月創刊號至 2009 年 1 月約 2000 篇；在新聞文章方面，我們有「自由時報中英對照讀新聞」約 1500 篇[3]、「雙語網站知識管理平台新聞」[11]約 700 篇及包括了「四技二專統一入學測驗」、「學科能力測驗」及「大學指定科目考試」中「對話測驗」、「綜合測驗」及「閱讀測驗」等多個段落的「大考試題」[10]，共約 130 篇(以下統稱這四種語料為「其它主題的文本」)，本研究詳細的語料來源統計數據如表一所示。由於語料數目相當龐大，在本系統開發時僅採用部份的語料作為訓練用途，如表二所示。

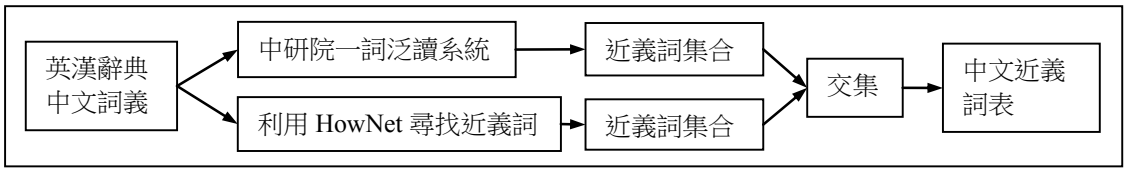
4. 研究方法

4.1 系統架構及對列流程

本系統的架構及處理流程如圖一所示，首先將任意中英互為翻譯、未經過段落對列處理的文章進行斷句、英文詞形還原、斷詞及詞彙擷取等前處理步驟，再透過相似度計算模組計算中英文句子的相似度，利用動態規劃演算法選取整體分數最佳之對列組合，產生對列結果，經過門檻值的篩選，再將高於門檻值條件的 1：1 對列模式句對取出，稱為「1：1 信心句對」，作為平行語料的用途。



圖一、系統架構流程圖



圖二、中文詞義近義詞表製作流程

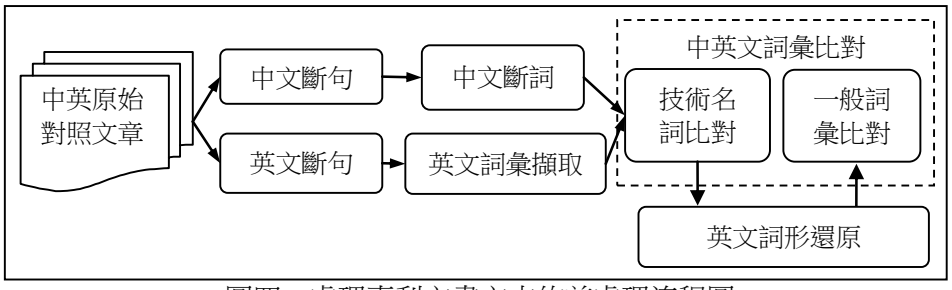
4.2 辭典之建置

在對列處理的過程中，利用英漢辭典進行中英句對的詞彙比對，為了提高詞彙比對的成功率，我們建置「中文近義詞表」，並將其合併至英漢辭典中，擴大英漢辭典的規模。除了建置中文近義詞表外，我們也建置「中英技術名詞對應表」，處理常見於專利說明書中的技術名詞。

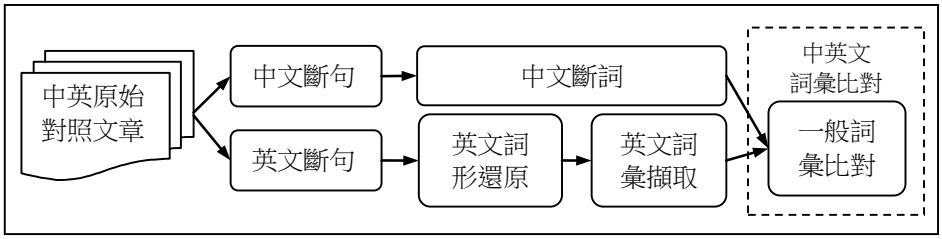
```
==============================
英文單字: chloride shift
氯轉移
鹽分移動
氯轉置
氯離子轉移
==============================
英文單字: chloride silver
氯化銀
==============================
英文單字: chloride stre
氯化物應力
==============================
英文單字: chloride stre corrosion crack
氯化物應力腐蝕裂縫
```

圖三、中英技術名詞對應表

- **英漢辭典**：本系統採用的電子英漢辭典，為國內知名企業英業達股份有限公司開發的 Dr.eye 譯典通線上辭典[12]，其含有 106269 筆英文詞彙及片語，在和牛津辭典比較詞目數量後，我們採用 Dr.eye 譯典通線上辭典作為本系統的英漢辭典。
- **中文近義詞表**：單純倚賴英漢辭典內之中文詞義進行中英詞彙比對雖然可行，但進行中英詞彙比對時的成功率並不高。為了增加中文詞彙比對的成功率，我們沿用呂明欣等學者[4]於 2007 年使用的中文近義詞尋找方法，來建構中文近義詞表，流程如圖二所示。在圖二中，我們將英漢辭典中的原中文詞義，透過 HowNet[13]及中研院一詞泛讀系統[2]尋找其意義相近的中文詞彙，在得到兩種方法產生的中文近義詞集合後，為了去除和原中文詞義偏差較大的中文近義詞，我們將兩個近義詞集合取交集，建構出中文近義詞表。我們將此中文近義詞表合併至英漢辭典中，作為中英文詞彙比對時的依據。
- **中英技術名詞對應表**：我們從「國立編譯館學術名詞資訊網」[6]擷取、整理並建構中英技術名詞對應表，中英文對應的詞目數量為 1660829 筆。在處理專利文書的文本時，除了一般性的英漢辭典外，若能再結合不同專業領域的中英技術名詞資訊，進行中英詞彙比對時，便不會將特殊、罕見的技術名詞視為未知詞，喪失了中英詞彙比對時應得的相似度分數。圖三為中英技術名詞對應表的部份內容，從圖中可以發現，單一英文技術名詞可能會對應至多個中文技術名詞，在中英詞彙比對的過程中，我們會將這些中文技術名詞皆納入比對的考量。
- **中文斷詞辭典**：中文斷詞辭典的詞彙涵蓋範圍，會影響以長詞優先方式進行斷詞的斷詞效果。許多進行斷詞相關研究的學者[14][20]認為較長的詞彙，能夠保留更完整的詞面資訊，因此，一套完整的中文斷詞辭典，在前處理的部份便佔了重要的角色。我們將合併中文近義詞表後的英漢辭典其全部的中文詞義，加上中英技術名詞對應表中全部的中文技術名詞，作為中文斷詞辭典，提供本系統在進行前處理時，進行中文斷詞的依據，中文斷詞辭典總詞目數量為 1835085 筆。



圖四、處理專利文書文本的前處理流程圖



圖五、處理其它主題文本的前處理流程圖

4.3 中英文原始文章前處理步驟

圖四為專利文書前處理流程圖。在圖四中，經過斷句處理後的中英文句子，接著進行中文斷詞及英文詞彙擷取，完成後即進入相似度計算模組中的中英文詞彙比對階段。在詞彙比對的過程中，首先利用中英技術名詞對應表進行技術名詞的比對，接著才進行英文詞形還原步驟。在處理專利文書的文本時，我們以「查表」的方式進行英文詞形還原，在完成英文詞形還原後，才利用英漢辭典進行一般性中英文詞彙的比對。

圖五為其它主題文本的前處理流程圖。可以發現圖五和圖四的最大不同點在於：完成斷句處理後的英文句子，在進行英文詞彙擷取前便進行了英文詞形的還原步驟，接著才進行詞彙擷取的處理及中英文詞彙的比對，這是因為在處理其它主題的文本時，並未進行技術名詞的比對。在處理其它主題的文本時，我們是以 StanfordLexParser-1.6 進行英文詞形的還原。

4.3.1 斷句

在中英文句子的斷句方面，我們僅利用「問號」、「驚嘆號」及「句號」三種符號進行斷句處理，而不將句子切分成太細的單位。我們利用簡單的規則避免英文句號與小數點混淆，並簡單地作人名及地名等名詞縮寫判斷，以避免斷句錯誤的情況發生。

4.3.2 英文詞形還原

StanfordLexParser-1.6 能夠在建立該英文句子的剖析樹 (parse tree) 後，根據該英文詞彙的詞性進行詞形還原處理。以圖六為例子，首先利用 StanfordLexParser-1.6 產生附帶著詞性的英文原句剖析樹，再利用 StanfordLexParser-1.6 的詞形還原套件對該剖析樹進行處理，使句中三個英文詞彙「plays」、「his」及「friends」進行詞形還原，產生這三個英文詞彙的原形「play」、「he」及「friend」。

StanfordLexParser-1.6 為 Java 語言所撰寫的套件，隨著電腦記憶體大小的不同，所能夠處理的英文句長也有限制。在專利說明書中，由於翻譯者的翻譯風格的不同，在文章中偶爾會出現極長的英文句子，這會造成剖析樹過深導致對列處理中斷，這樣的現象我們無法預期何時會發生。因此我們在對列其它主題的文本時，使用 StanfordLexParser-1.6 進行英文詞形還原的處理，而在對列專利文書的文本時，我們在完成斷句及斷詞後，在中英文詞彙比對的階段，完成技術名詞的比對之後才利用查表的方式進行英文詞形還原。我們沿用[15]開發的對列工具 Champollion 的英文詞形還原對應表，共有 136390 筆英文詞目及原形。有關中英詞彙比對的方式及步驟，會在 4.4.1 節作介紹及說明。

4.3.3 中文長詞優先斷詞、英文詞彙擷取

在完成中英文斷句處理後，在中文的部分首先對已經完成斷句的中文句子進行斷詞處理，在英文的部份，擷取出英文片語及技術名詞後再利用空白將其它的英文單詞分開，以利後續中英文翻譯詞彙的比對，我們以英漢辭典的英文詞彙及中英技術名詞對應表的英文技術名詞作為英文詞彙擷取的依據，詞目數量為 1151130 筆。

英文原句：Jim always plays baseball with his friends.
其剖析樹：(S(NP(NNP Jim))(ADVP(RB always))(VP(VBZ plays)(NP(NN baseball))(PP(IN with)(NP(PRP$ his)(NNS friends))))(. .))
詞形還原：Jim always play baseball with he friend.

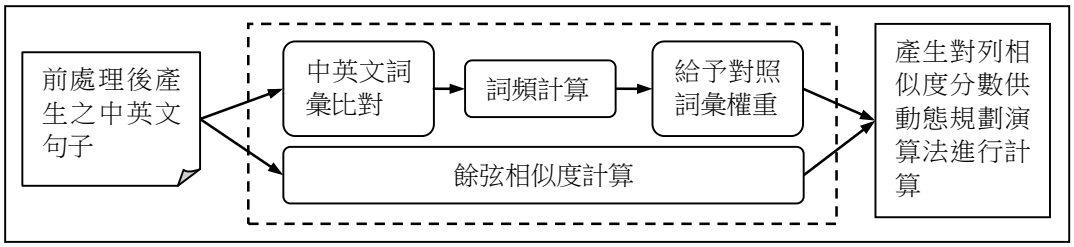
圖六、利用 StanfordLexParser-1.6 進行英文詞形還原

原中文句子：本發明提供一種形成半導體電子裝置的閘極堆疊的方法，其藉由晶圓接合含有高介電常數介電材料之至少一結構。

中研院斷詞：本 發明 提供 一 種 形成 半導體 **電子** **裝置** 的 **閘** **極** 堆疊 的 方法 ，其 藉由 **晶圓** **接合** 含有 高 **介電** **常數** **介電** **材料** 之 至少 一 結構 。

本系統斷詞：本 發明 提供 一種 形成 半導體 **電子裝置** 的 **閘極** 堆疊 的 方法 ，其 藉 由 **晶圓接合** 含有 高 **介電常數** **介電材料** 之至 少 一 結構 。

圖七、中研院中文斷詞及本系統斷詞結果比較



圖八、相似度計算模組（以虛線表示）

目前常見的中文斷詞工具為中研院的中文斷詞系統[1]，我們能透過網頁介面或撰寫程式進行查詢，提供中研院一段中文句子或文章，中文斷詞系統會進行詞性的標記及斷詞處理，並將查詢結果回傳。雖然中研院的中文斷詞系統為目前廣泛使用的斷詞工具，但為了處理專利說明書中常見的技術名詞，即時且大量地進行對列工作，我們不採用中研院的中文斷詞系統進行斷詞，而是利用中文斷詞辭典，以長詞優先的斷詞方式進行中文句子的斷詞處理。

圖七為利用中研院中文斷詞系統及中文長詞優先斷詞結果的比較，在技術名詞的斷詞差異處我們以粗體加上底線的字體表示。在圖七中，中研院的中文斷詞系統將「電子裝置」、「閘極」、「晶圓結合」、「介電常數」及「介電材料」等五個中文技術名詞錯誤斷詞，而利用長詞優先斷詞能夠正確地保留住這些中文詞彙。以圖中的中文詞「晶圓接合」為例，其對應的英文詞彙為「chip connection」，由於我們以「完全比對」的方式進行中英技術名詞的比對（也就是當中文詞必須完全相同才算比對成功），若「晶圓接合」被錯誤斷成「晶圓」及「接合」二詞，則在技術名詞比對的步驟便無法成功比對，而英漢辭典中沒有「chip connection」此英文片語，在中英文詞彙比對的步驟上，「chip connection」及「晶圓接合」這組中英文詞彙便失去了應得的詞彙比對分數。在圖七中的例子，共計有五組這樣的中英文對應詞彙被斷詞錯誤，累計起來失去的相似度分數便相當多，這也是我們採用長詞優先的方式進行中文句子的斷詞，並且將英文句子的片語及技術名詞擷取出來的主要原因。

4.4 相似度計算模組

相似度計算模組如圖八所示。相似度計算模組中，有兩個部份同時進行。模組中的第一個部份為中英文句子詞彙的比對及給分。第二個部份以計算向量之間相似度的原理，將中英文句子視為兩組向量，計算其之間的向量相似度。此計算方法不需要句長統計的數據，即可將中英文句子間的句長差異納入扣分考量。除此之外，對於句長相似但擁有不同權重詞彙之中英文句子（即不應該為翻譯對應之中英文句子），也能進行懲罰，使其得到較低的分數。將第一部份中英詞彙比對得到的分數，乘上第二部份之向量相似度作為扣分權重，最後得到該對列模式下的中英文句子相似度總分，這項分數接著提供給動態規劃演算法作為挑選最佳對列模式組合的依據。

4.4.1 翻譯詞彙的搜尋及比對方式

在中英文句詞彙比對的過程中，中英技術名詞對應表的參照順序在英漢辭典之前。進行中英文句子的詞彙比對時，以由左至右的順序將完成詞彙擷取處理的英文句子詞彙取出，並至中英技術名詞對應表進行搜尋，以完全比對的方式比對中文詞彙。

我們接著以同樣的方式，將英文句子中尚未比對到的詞彙取出，至英漢辭典搜尋。若能在英漢辭典中找到該英文詞彙，則將其中文詞義取出，至完成斷詞處理後的中文句子比對中文詞彙。為了兼顧比對的成功率及正確性，本系統以一字詞完全比對、二字詞以上部份比對的方式進行比對的動作，也就是英文詞彙及中文詞彙的最長共同子序列 (longest common subsequence)

為二字詞以上，也視為比對成功。當完成一組中英文句子的詞彙比對後，這些成功比對到的中英文詞對便形成詞對集合，我們予以記錄，供後續計算使用。

4.4.2 詞彙權重的計算

我們沿用[15]提出的方法，依照詞對集合內，各詞對的英文詞彙重要程度不同，給予比對到的詞對不同分數。

在資訊檢索的領域中，*tf-idf* (term frequency - inverse document frequency)[18]為一常見的計算公式，能夠計算文件中的詞彙權重值，*tf* 值為文件中某詞彙的出現次數，如公式(1)所示。而 *idf* 值代表該詞彙在所有語料的文件中的重要程度，*idf* 值的計算方式如公式(2)所示。在公式(2)中，N 為文件總數，w 為詞彙，$nd(w)$代表含有詞彙 w 的文件數量。詞彙的 *tf-idf* 權重值計算方式如公式(3)所示。

$$tf(w) = \text{詞彙 } w\text{在文件中出現的次數} \quad (1) \qquad idf(w) = \log\left(\frac{N}{nd(w)}\right) \quad (2)$$

$$tfidf(w) = tf(w) \times idf(w) \quad (3)$$

[15]將計算文件中詞彙權重值的概念，應用到計算句子中的詞彙權重值。他認為中英文詞彙的比對，不應該將所有成功比對到的詞彙分數視為相同，太常出現的英文字，其強度代表性不如鮮少出現的英文詞彙。他將 *tf* 值的計算方式進行修改，並將名稱改寫為 *stf* (segment-wide term frequency) ，代表某中英文對應詞彙在該中英文句子中出現的次數如公式(4)所示，在公式(4)中，e 及 c 分別為某詞對集合中對應詞對的英文詞彙及中文詞彙，若進行中英文句子詞彙比對時，在英文句子中 e 出現三次，在中文句子中 c 出現三次，則在進行中英文詞彙比對的過程中，英文詞彙 e 及中文詞彙 c 會比對到三次，則其 *stf* 值為 3；[15]將 *idf* 值的計算方式進行修改，並將名稱改寫為 *idtf* (inverse document term frequency) ，代表句子中的詞彙在整篇文章中的重要程度，*idtf* 值的計算方式如公式(5)所示。在公式(5)中，T 代表文章的總詞頻（在此將總詞頻定義為完成「詞彙擷取」前處理步驟後的英文文章，統計其包含的英文詞彙出現的總次數，而非不同英文詞彙的數量，例如在某英文文章中只有三個英文詞彙：「really」、「really」及「good」，則該篇文章的英文總詞頻數為 3），e 代表詞對集合中詞對的英文詞彙，$O(e)$代表該英文詞彙 e 在英文文章中出現的次數。如公式(6)所示，將 *stf* 值及 *idtf* 值進行相乘，可以得到該中英文詞對的 *stf-idtf* 值。由於[15]在公式(6)中是利用英文詞彙 e 進行 *stf-idtf* 值的計算，我們在此予以沿用。

$$stf(e,c) = \text{詞對集合中的中英文對應詞彙在該中英文句子出現的次數} \quad (4)$$

$$idtf(e) = \frac{T}{O(e)} \quad (5) \qquad stfidtf(e,c) = stf(e,c) \times idtf(e) \quad (6)$$

$$\text{中英文句子基礎相似度分數} = \sum_{i=1}^{k} \log(stfidtf(e,c)) \quad (7)$$

如公式(7)所示，假設某中英文句子間其比對到的詞對數共有 k 組（詞對集合中有 k 組詞對），在得到各詞對的 *stf-idtf* 值後，再取對數函數（以 10 為底數）將該 k 組詞對的分數進行加總，得到該組中英文句子的「基礎相似度分數」（在此稱為基礎相似度分數，是因為該組中英文句子還需經過向量相似度的計算，最後才會得到該組中英文句子的相似度分數）。

4.4.3 中英句子向量相似度的計算

我們在這篇論文中提出衍生自計算向量之間相似度的方法（餘弦相似度），作為中英文句子相似度的輔助計算，我們稱作「中英文句子間的向量相似度」，這也是本系統相似度計算模組中的第二個部份，常見的餘弦相似度的計算方式如公式(8)所示。

$$\cos\theta = \frac{V_1 \cdot V_2}{\|V_1\| \, \|V_2\|} \quad (8)$$

在公式(8)中，餘弦函數值介於 0 至 1 之間，當兩個向量的夾角越小，則其餘弦函數值越接近 1，代表這兩個向量越相似，反之則越接近 0。

$$ctf(w) = \frac{SO(w)}{L} \quad (9) \qquad idcf(w) = \log\left(\frac{SN}{n(w)}\right) \quad (10)$$

$$ctfidcf(w) = ctf(w) \times idcf(w) \quad (11)$$

我們利用[15]提出的 *stf-idtf* 概念，將 *tf-idf* 的公式，從文件的層級轉化為句子的層級，但不同於[15]，我們將 *tf* 值的計算方法作修改，並將名稱改寫成 *ctf*，如公式(9)所示。在公式(9)中，*w* 代表詞彙，*SO*(*w*)代表詞彙 *w* 出現在該句子中的頻率，*L* 則為該句子的詞數，即句長。我們將 *idf* 值的計算方式作修改，並將名稱改寫成 *idcf*，如公式(10)所示。在公式(10)中，*SN* 為一篇文章中的句子數量，*w* 代表詞彙，*n*(*w*)代表在文章中含有詞彙 *w* 的句子數量。我們將此方法應用於計算中英文句子之間的相似程度，視英文句及中文句各為一個向量，向量中屬性的值為每一個詞彙的 *ctf-idcf* 值，*ctf-idcf* 值的計算方式如公式(11)所示。

我們將 *tf-idf* 值的計算公式進行修改，因為 *tf-idf* 值代表的是某詞彙對其所在文件的權重，我們希望將詞彙對所屬文件的權重，轉化為詞彙對所屬句子的權重。原先存在文件之間的某詞彙，經過其 *tf-idf* 值的計算，我們可以知道其對其所屬文件的重要程度，轉化為句子的層級後，我們就可以知道某詞彙對其所屬句子的重要程度。和[15]提出的 *stf-idtf* 值計算原理不同，他所提出的 *idtf* 值計算方式雖然同樣計算句子中的詞彙重要程度，不過 *stf-idtf* 值乘上 *stf* 值的目的是為了要將 *idtf* 的值加倍，也就是當一組中英文句子間某一個中英文詞彙出現多次，則這一組中英文句子的相關性就更強，而我們提出的 *ctf-idcf* 值計算目的，則是為了要得到句子中的詞彙對於文件中該句子的權重，作為計算向量相似度時使用。

我們的構想為：當一篇中文文章和英文文章若互為翻譯，在正確地中文斷詞及英文詞彙擷取的情況下，英文句子中的每一個英文詞彙都應該對應至中文句子的一個中文翻譯詞彙。我們將中英文句子各視為一個向量，向量中屬性的值為每一個詞彙的 *ctf-idcf* 值，若在英文句子中去除 stop words 的前提下，和未去除 stop words 的英文句子相比較，有去除 stop words 的英文句子在進行中英文句子的向量相似度計算時，其值會較趨近於 1（本系統並未對英文文章進行去除 stop words 的前處理，因此僅將向量相似度計算得到的分數作為輔助分數）。但單純地利用向量相似度計算仍會碰到數個困難點，如：正常情況下，中英文句子長度通常不會相同，導致向量的維度不同而無法計算；中英互為翻譯的句子，因為語言本身的特性，常具有翻譯詞序不同的現象，這將導致在計算向量內積時，會對應到錯誤的的 *ctf-idcf* 值。

基於以上可能會碰到的問題，我們在進行相似度計算前，先對向量內的 *ctf-idcf* 值由小到大進行排序，以解決中英文詞序可能不同的問題。由於中英文句長可能不同，我們將維度較小的向量，進行補足維度的動作，也就是將維度較小的向量，補上 *ctf-idcf* 值為 0 的值，直到中英文句子的維度相同，得以進行相似度的計算。補足維度的方式，同時作為針對中英文句長差異的

E = { [Increasing][LED][directionality][makes][LEDs][more][attractive][for certain]
0.18 0.18 0.13 0.15 0.13 0.15 0.18 0.18
[applications][such as][projectors] }
0.18 0.15 0.18

C = { [增加] [發光二極體] [的方向性] [可以] [使發光] [二極體] [對於] [例如] [投影機] [之]
0.10 0.10 0.13 0.06 0.13 0.13 0.11 0.09 0.13 0.01
[特定] [應用] [變得] [更] [吸引] }
0.11 0.13 0.10 0.11 0.13

⇩ 將向量內的 *ctf-idcf* 值進行補齊、排序

V_E = { 0, 0, 0, 0, 0.13, 0.13, 0.15, 0.15, 0.15, 0.18, 0.18, 0.18, 0.18, 0.18, 0.18 }

V_C = { 0.01, 0.06, 0.09, 0.10, 0.10, 0.10, 0.11, 0.11, 0.11, 0.13, 0.13, 0.13, 0.13, 0.13, 0.13 }

句對餘弦相似度分數：0.935816

圖九、向量相似度計算範例

計分，即中英文句子之間的長度差異越大，則其向量相似度的分數，便會越小而越趨近於 0。利用此計算方法的優點在於，互為翻譯的中英文句子長度雖然有差異，但不全然會因為長度的差異而得到過低的向量相似度分數，當中英文句子的向量具有數量及數值差異較小的 *ctf-idcf* 值時，便能夠得到較高的向量相似度分數。由於可能發生權重值相同及句長相同的誤判情況，因此我們僅將向量相似度分數作為輔助扣分的依據。若向量相似度分數越趨近於 1，則在和第一部份的基礎相似度分數（中英文詞彙比對得到的相似度分數）相乘之後，仍保留住分數，反之，相乘之後則形同扣分的作用。

圖九為本系統進行向量相似度計算的範例。在圖九中，E 及 C 分別代表英文及中文的句子，V_E 及 V_C 分別為代表英文及中文句子的向量。我們將經過斷詞之後的中英文詞彙以括號予以區隔，而在中英文詞彙的下方，我們標記著該詞彙的 *ctf-idcf* 值。由於範例中的中英文句子句長不相同，在英文句子較短的情況下，我們將代表英文句子的向量補上四個值為 0 的 *ctf-idcf* 值，使代表中英文句子的向量維度相同，接著再對向量內的 *ctf-idcf* 值進行排序，最後進行向量相似度的計算，得到中英文句子的向量相似度分數。將此分數與相似度計算模組中的第一部份計算得到的相似度分數予以相乘便得到中英文句子的相似度分數。

4.4.4 對列模式與動態規劃演算法

考慮專利文書文本及其它主題文本的語料特性，我們僅採用 9 種對列模式，分別為：「1：0」、「0：1」、「1：1」、「1：2」、「2：1」、「1：3」、「3：1」、「1：4」及「4：1」。

$$S(i,j) = max\begin{cases} S(i-1,j) + sim(Seg_{i,i}, \emptyset) \\ S(i,j-1) + sim(\emptyset, Seg_{j,j}) \\ S(i-1,j-1) + sim(Seg_{i,i}, Seg_{j,j}) \\ S(i-1,j-2) + sim(Seg_{i,i}, Seg_{j-1,j}) \\ S(i-2,j-1) + sim(Seg_{i-1,i}, Seg_{j,j}) \\ S(i-1,j-3) + sim(Seg_{i,i}, Seg_{j-2,j}) \\ S(i-3,j-1) + sim(Seg_{i-2,i}, Seg_{j,j}) \\ S(i-1,j-4) + sim(Seg_{i,i}, Seg_{j-3,j}) \\ S(i-4,j-1) + sim(Seg_{i-3,i}, Seg_{j,j}) \end{cases} \quad (12)$$

中英文的句子依照對列模式的不同而進行組合，例如：對列模式「1：3」，則本系統會將三句中文句子進行組合成為一句，再將一句英文句子和這組合後的中文句子進行句子相似度的計算，經過相似度計算模組的計算，會得到一個相似度分數，我們沿用[15]採用的最佳對列模式組合的動態規劃演算法，如公式(12)所示，來挑選整篇中英文文章的最佳對列組合。

在公式(12)中，$S(i,j)$代表來源語言文章的 i 個句子及目標語言文章的 j 個句子之間的相似度分數，$Seg_{a,b}$代表中英文章內句子編號 a 至 b 的句子，$sim(Seg_{a,b}, Seg_{c,d})$ 則為兩組中英文句子 $Seg_{a,b}$ 及 $Seg_{c,d}$ 之間的相似度分數，從中英文文章的第一句計算至最後一句，在過程中每一次都記錄該回合最高分的對列模式，最後得到整篇中英文文章的對列總分 $S(i,j)$，沿著計算得到$S(i,j)$的路徑往回推算可以得到每一回合最高的對列分數及對列模式，最後將這些回溯得到的分數及模式列出，可以得到中文文章中第一句至第 i 句之間的句子和英文文章中第一句至第 j 句之間的句子對列結果。

4.4.5 「1：1 信心句對」篩選門檻

參考[19]的作法，我們將擷取平行語料的目標放在 1：1 模式的句對上。我們設計在本系統最後產生對列結果後，能夠依照向量相似度的分數及句對平均比對到的詞數這兩項門檻值進行篩選，讓使用者能夠得到正確率較高的 1：1 句對。句對平均比對詞數計算方式如公式(13)所示。在公式(13)中，將中英文句對比對到的中英文詞對數量除以英文句長，可以得到該句平均比對到的詞對數量，當平均比對到的詞對數量越多，表示本系統在該中英文句對之間比對到越多中英文詞彙，這也表示此句對的對應正確性就越高，越有可能是互為翻譯的中英文句子。

$$\text{句對平均比對詞數} = \frac{\text{句對比對到的詞對數}}{\text{英文句句長}} \quad (13)$$

表三、訂定篩選門檻值之測試數據

語料	語言	句數	總詞彙個數	平均句長	1：1 對列數	1:1 對列錯誤數	精確率
專利公開全文 50 篇	中文	13775	263051	19.1	7131	33	0.995
	英文	10847	146342	13.5			

另一項門檻值則為句對的向量相似度分數，因為我們觀察到具有較低向量相似度分數的中英文句對，通常不會是正確的對列結果，因此採用向量相似度分數作為另一項門檻值條件，利用兩項門檻值條件篩選出 1：1 的句對。為了能夠訂定較為客觀且正確的門檻值，在系統開發的過程中，我們對訓練語料以隨機產生檔案序號的方式抽選 50 篇「專利公開全文摘要」進行對列測試。在本系統完成對列程序後，共產生 7131 組 1：1 句對。我們針對這些句對進行檢查，得到 42 組錯誤的對列結果，在將對列過程中比對成功詞對數為 0 的 1：1 對列結果去除後（在有設定門檻值的情況下，本系統原本就不會對這些沒有比對到詞數的句對作篩選的處理），最後得到錯誤的組數為 33 組，詳細的測試數據，如表三所示。

我們針對這些錯誤的 1：1 結果進行分析，發現這些錯誤的對列結果其平均向量相似度分數為 0.835，句對平均比對到的詞數為 0.218。而這 33 組錯誤結果中，有 28 組的向量相似度分數在 0.94 以下，佔了錯誤組數量的 84.8%；有 32 組的句對比對詞數在 0.34 以下，佔了錯誤組數量的 97%，因此我們訂定「向量相似度分數 0.94」及「句對比對詞數 0.34」作為本系統進行對列結果測試的預設門檻值。

5. 系統效果評估

5.1 實驗語料來源

本實驗採用的對列實驗語料，主要分為專利文書的文本及其它主題的文本。而在輔助式機器翻譯系統翻譯品質評估方面，則依照 5.2 節的實驗設計，也將本系統產生的「1：1 信心句對」分成專利文書文本及其它主題文本兩種，並加以組合，分別進行測試。

本實驗的第一個部份為對列結果的評估，詳細的實驗語料統計數據如表四所示。我們以電腦產生隨機亂數序號代表進行測試的檔案，並以人工的方式作檔案抽取的動作。在專利文書方面隨機抽選的語料，共計有申請號範圍從 091132651 至 095121449「專利公開全文摘要」4998 篇、申請號範圍從 091132651 至 094101510 的「專利公開全文敘述」200 篇（包括了技術領域、先前技術、發明內容及實施方法等四個段落）。在其它主題文本方面，新聞文章有「雙語網站知識管理平台新聞」從 2005 年 8 月 30 日 至 2007 年 12 月 15 日共計 737 篇及「自由時報中英對照讀新聞」從 2005 年 2 月 14 日至 2006 年 12 月 31 日共計 686 篇；科普文章有「科學人雜誌中英對照電子書」從 2003 年 1 月至 2009 年 1 月共計 1745 篇文章及包括了「四技二專統一入學測驗」、「學科能力測驗」及「大學指定科目考試」中「對話測驗」、「綜合測驗」及「閱讀測驗」等多個段落的「大考試題」，共約 130 篇。

在本實驗的第二個部份我們為了進行「1：1 信心句對」的效果評估，將進行實驗的語料產生的「1：1 信心句對」分成專利文書及其它主題文本的句對各為 26401 句及 42010 句。

表四、對列實驗語料統計

語料	語言	文章數	總句數	總詞彙個數	文章平均句數	平均句長
專利公開全文摘要	中文	4998	19899	520475	3.98	26.2
	英文		18968	452150	3.80	23.8
專利公開全文敘述	中文	200	47985	1127704	239.93	23.5
	英文		42072	1016088	210.36	24.2
科學人雜誌中英對照電子書	中文	1745	112649	1871576	64.56	16.6
	英文		117785	2376440	67.50	20.2
雙語網站知識管理平台新聞	中文	737	9272	207580	12.58	22.4
	英文		9408	191051	12.77	20.3
自由時報中英對照讀新聞	中文	686	5523	123803	8.05	22.4
	英文		5594	104699	8.16	18.7
大考試題	中文	131	1534	27937	11.71	17.1
	英文		1604	24152	12.24	15.1

表五、TIMSS2003 實驗組別

八年級 2003 M 組	八年級 2003 S 組	四年級 2003 M 組	四年級 2003 S 組	八年級 2003 MS 組	四年級 2003 MS 組
國中數學 領域試題	國中科學 領域試題	國小數學 領域試題	國小科學 領域試題	國中數學及科學 領域試題	國小數學及科學 領域試題

5.2 實驗設計

表六、翻譯評比實驗組別

組別	建立範例樹語料
A	Chang 科學人
B	專利 1：1
C	一般 1：1
D	專利 1：1 + 一般 1：1
E	Chang 科學人 + 專利 1：1
F	Chang 科學人 + 一般 1：1
G	Chang 科學人 + 專利 1：1 + 一般 1：1

5.2.1 對列測試及對列結果隨機抽驗

在實驗的第一個部份，我們對實驗語料進行對列測試，再以精確率 (precision) 及召回率 (recall) 分別針對專利文書文本及其它主題的文本作評估。參考[19]的作法，我們以隨機抽樣的方式進行檢驗，利用電腦產生隨機檔案序號，並以人工選取這些檔案進行對列結果抽樣檢測及評比。

在對列結果檢測方面，我們的操作方法為：在事先沒有正確對列答案的情況下，我們首先將完成對列的檔案取出（此檔案為本系統對列結果），並且複製一份同樣的檔案進行句對的檢測，若該句對的對列結果是正確的，則註記本系統答對（即對列正確）；若該句對的對列結果是錯誤的，則予以修正至正確的對列結果。當完成此複製檔案的註記後，我們便得到了正確的對列結果及本系統答對的對列結果，加上原先產生的對列結果，便可以進行精確率及召回率的評比。

依照文本的不同，我們將對列結果的評估分為專利文書的對列結果、其它主題文本的對列結果及綜合對列結果三種進行評比。為了比較向量相似度計算機制的效果，特別將該機制移除後，同樣對這些隨機抽樣的檔案進行對列實驗。除了評比本系統的計算機制，我們也以同樣的方式，利用[15]提出的對列工具 Champollion 進行對列實驗，和其比較對列的效果。

5.2.2 利用輔助式機器翻譯系統進行翻譯

在實驗的第二個部份，透過張智傑及劉昭麟[8]於 2008 提出的輔助式機器翻譯系統，對 2003 年國際數學與科學教育成就趨勢調查 (Trends in International Mathematics and Science Study，簡稱 TIMSS) 的試題（以下簡稱 TIMSS2003 試題）進行翻譯效果的評估，以檢視本系統產生的大量平行語料，是否能夠提升現有輔助式機器翻譯系統的翻譯品質。TIMSS2003 試題的實驗組別如表五所示。

參考[8]採用的系統方法及流程，他們的作法分為建構範例樹及翻譯模組兩個部份。在第一個部份，他們將中英文句對作為語料，並利用 StanfordLexParser-1.6 產生英文句子的剖析樹，在中英文詞彙對應後，將中英文句子中對應詞彙順序有前後調換現象的句對，記錄其英文剖析樹及其子樹的樹葉詞彙順序編號及詞性，建構成為範例樹資料庫，在進行翻譯處理時，則將剖析後的目標英文句，依照其詞性進行範例樹資料庫的搜尋比對，若有比對到範例樹，則能夠依照其詞彙順序的編號進行翻譯時詞序的調動，由於僅記錄有詞序調換現象句子的剖析樹，因此大量的平行句對可能在建構範例樹資料庫的過程中便篩選剩下較少的句對。他們的系統第二個部份為翻譯模組，針對英文句子進行中譯的動作，第一個部份的正確詞序調動能夠幫助他們的系統在第二個部份作更正確的選詞。

我們將產生的「1：1 信心句對」視為平行語料，這些「1：1 信心句對」在透過建構範例樹之詞彙對列的篩選機制進行篩選後，專利文書文本及其它主題文本的句對分別剩下 556 句及 608 句。我們同時沿用[8]當時採用的實驗語料以進行比較：在建立範例樹的語料方面，共計有「科學人雜誌中英對照電子書」從 2002 年至 2006 年共 110 篇文章，經過詞彙對列篩選後共計有 30 組句對作為建構範例樹的語料[3]。而在訓練選詞機率模型方面，則有「科學人雜誌中英對照電子書」從 2002 年至 2006 年共 2685 個句對[4]及「自由時報中英對照讀新聞」從 2005 年至 2007 年共 4248 個句對。他們的語料並未和我們進行翻譯品質評估的語料重複。

[3] 該論文作者於之後碩士論文提出翻譯評比分數較高之範例樹語料組合，故於本實驗中予以沿用。

[4] 該論文作者於之後碩士論文提出更新後的數據。

表七、抽樣對列目標統計數據

		專利文書文本		其它主題文本				總和
		專利公開全文摘要	專利公開全文敘述	科學人雜誌中英對照電子書	雙語網站知識管理平台新聞	自由時報中英對照讀新聞	大考試題	
英文句		192	695	340	228	142	231	1828
中文句		177	845	326	228	151	228	1955
1：0 及 0：1	對列數	10	150	9	7	6	4	186
	比例	5.85%	20.6%	3%	3.3%	4.3%	1.8%	10.5%
1：1	對列數	138	425	210	159	115	179	1226
	比例	80.7%	58.4%	70%	76.1%	81.6%	82.5%	69.4%
1：2 及 2：1	對列數	18	81	70	41	16	31	257
	比例	10.5%	11.1%	23.3%	19.6%	11.3%	14.3%	14.6%
1：3 及 3：1	對列數	0	38	6	0	3	3	50
	比例	0%	5.2%	2%	0%	2.1%	1.4%	2.8%
1：4 及 4：1	對列數	2	5	2	0	0	0	9
	比例	1.2%	0.7%	0.7%	0%	0%	0%	0.5%
其它	對列數	3	29	3	2	1	0	38
	比例	1.8%	4.0%	1%	1.0%	0.7%	0%	2.2%
總和		171	728	728	300	209	141	217

1:1 的句對依照文本的不同分為專利文書文本及其它主題文本（以下分別簡稱「專利 1：1」及「一般 1：1」），並搭配[8]當時採用的實驗語料（以下簡稱「Chang 科學人」）共設計出 7 種組合，分別以代號 A 至 G 表示，如表六所示，並和 Google 的翻譯結果進行比較。

5.3 實驗結果與分析

5.3.1 對列效果分析

依照實驗的設計，我們對進行對列測試的語料檔案進行隨機抽樣評比，各文本的句數及對列模式統計數據如表七所示。在表七中可以觀察到，各文本皆以 1：1 的對列模式佔了最多的數量比例。而在「專利公開全文敘述」的抽樣語料中，「1：0 及 0：1」的對列模式佔了約 20%的比例，這顯示專利說明書的內文敘述有許多沒有完整的中英對應。觀察表七，這些語料的「其它」對列模式數量佔了 2.2%，這些多句對應的對列模式並不在我們對列處理的考量內。

表八、移除「向量相似度計算」機制後的抽樣對列結果數據

語料	專利文書文本		其它主題文本		全部文本	
模式	精確率	召回率	精確率	召回率	精確率	召回率
1：0 及 0：1	0.561	0.460	0.333	1.000	0.519	0.491
1：1	0.978	0.868	0.958	0.802	0.967	0.832
1：2 及 2：1	0.545	0.692	0.629	0.897	0.598	0.815
1：3 及 3：1	0.416	0.933	0.412	0.875	0.415	0.918
1：4 及 4：1	0.238	0.833	0.5	1.000	0.280	0.875

表九、加入「向量相似度計算」機制後的抽樣對列結果數據

語料	專利文書文本		其它主題文本		全部文本	
模式	精確率	召回率	精確率	召回率	精確率	召回率
1：0 及 0：1	0.579	0.619	0.393	0.923	0.530	0.661
1：1	0.989	0.913	1.000	0.908	0.995	0.910
1：2 及 2：1	0.652	0.929	0.812	0.981	0.744	0.961
1：3 及 3：1	0.413	0.868	0.588	0.833	0.443	0.860
1：4 及 4：1	0.333	0.429	1.000	0.500	0.400	0.444

表十、Champollion 抽樣對列結果數據

語料	專利文書文本		其它主題文本		全部文本	
模式	精確率	召回率	精確率	召回率	精確率	召回率
1：0 及 0：1	0.485	0.463	0.675	1.000	0.529	0.552
1：1	0.951	0.982	0.991	0.973	0.972	0.977
1：2 及 2：1	0.563	0.831	0.897	0.906	0.739	0.877
1：3 及 3：1	0.489	0.742	0.667	0.667	0.509	0.730
1：4 及 4：1	0.238	1.000	1.000	1.000	0.304	1.000

我們將向量相似度計算機制移除後，進行對列的結果數據如表八所示。從表八中可以發現，僅倚賴中英文詞彙比對的方式進行對列，在全部文本的對列表現，以 1：1 對列模式的精確率最高，但其餘對列模式的精確率皆不高，未達 0.6。而在召回率的部份，觀察數量最多的 1：1 對列模式在全部文本的表現，召回率未超過 0.9，這也表示僅倚賴中英文詞彙比對的方式進行對列，

確實有改進的空間，也因此在研究的過程中，我們加入了向量相似度計算的輔助機制，期望提升整體的對列效果。

本系統在加入向量相似度計算機制後的對列結果如表九所示。在表九中，在專利文書 1：1 對列模式的部份，精確率高達 0.989，這表示產生的 1:1 句對近乎完全正確，召回率也達到 0.913，表示能夠找出大部份正確的 1：1 句對正確。而在 1：2 及 2：1 對列模式的部份，雖然召回率很高，但三種組合在精確率的部份和 1：1 模式相較之下則下降許多，分析其原因，有許多較短的句子會因為詞彙數較少，在中英詞彙比對時並未成功比對到詞彙，但在動態規劃演算法整體計分下，最後會認為這樣的組合總分最高，而產生錯誤的對列結果。在其它的「一對多」對列模式下，也有相同的錯誤原因。從其它主題的文本對列結果可以觀察到，在精確率的部份，以 1：1 及「1：4 及 4：1」對列模式最高，在 1：2 及 2：1 對列模式方面也有 0.812 的精確率，表示本對列系統在 1：1 的對列模式下，不僅在專利文書的文本，在其它主題的文本同樣能有高精確率。在召回率的部份，表現則以 1：2 及 2：1 對列模式最佳，而「1：0 及 0：1」及 1：1 兩種模式的召回率有 0.9 以上，1：3 及 3：1 對列模式有 0.8 以上。和專利文書文本對列結果的精確率及召回率進行比較，我們可以發現在其它主題文本的表現都較佳，最大的因素為其它主題文本的翻譯品質較專利文書整齊，在對列的挑戰上，其它主題的文本較專利文書簡單許多。觀察全部文本的對列結果，在精確率的部份，以 1：1 對列模式表現最佳，高達 0.995，這也表示在近 1200 組 1：1 的模式下，本系統產生的對列答案近乎全對，其召回率也高達 0.910，這也是我們以 1：1 對列結果作為翻譯語料的原因之一。

我們用 Champollion 以同樣的方式進行對列，得到的對列結果如表十所示，與表九的本系統對列結果作比較，可以發現在其它主題文本的對列表現上，在精確率的部份，對列模式「1：0 及 0：1」、「1：2 及 2：1」及「1：3 及 3：1」的結果 Champollion 皆較本系統佳，而在專利文書文本的表現上，僅有對列模式 1：3 及 3：1 的表現優於本系統，這表示在此實驗中，本系統在專利文書文本的表現較優於 Champollion。從全部文本的對列結果進行綜合比較，Champollion 也僅有對列模式 1：3 及 3：1 的表現優於本系統，不過可以觀察到在 1：1 對列模式的精確率及召回率部份，Champollion 雖然精確率較本系統低，但其召回率卻較高，在 1：1 對列模式的部份，本系統與 Champollion 確實皆有很好的表現。

5.3.2 輔助式機器翻譯效果評估

在本實驗的第二個部份，我們利用 BLEU 及 NIST 指標對 TIMSS2003 進行翻譯評測，各實驗組別評比得到的分數如表十一所示。在表十一中，我們以斜粗體字表示各試題組別中除了 Google

表十一、各種組合之 BLEU 及 NIST 評比分數[5]

組別	八年級 2003 M 組		八年級 2003 S 組		四年級 2003 M 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU
A	4.7956	0.1506	4.4854	0.1454	4.1865	***0.1501***
B	4.7911	0.1510	4.4983	0.1467	4.1817	0.1461
C	4.7713	0.1490	4.4941	0.1470	4.1841	0.1464
D	4.7920	0.1518	4.5091	0.1485	4.1866	0.1464
E	4.7893	0.1512	4.4967	0.1466	4.1703	0.1459
F	4.7946	0.1508	4.5003	0.1484	***4.2759***	0.1473
G	***4.7986***	***0.1529***	***4.5111***	***0.1498***	4.1708	0.1459
Google	4.5930	0.1482	5.0538	0.1898	3.7682	0.1046
組別	四年級 2003 S 組		八年級 2003 MS 組		四年級 2003 MS 組	
指標	NIST	BLEU	NIST	BLEU	NIST	BLEU
A	4.2023	0.1072	4.9619	0.1487	4.5040	***0.1251***
B	***4.2262***	0.1075	4.9655	0.1494	4.5167	0.1235
C	4.2074	0.1075	4.9493	0.1483	4.5037	0.1237
D	4.2261	***0.1076***	4.9698	0.1506	4.5188	0.1238
E	4.2252	0.1074	4.9635	0.1495	4.5120	0.1235
F	4.2064	0.1075	4.9668	0.1500	***4.5352***	0.1239
G	4.2251	***0.1076***	***4.9747***	***0.1518***	4.5120	0.1236
Google	4.8162	0.1655	5.0646	0.1637	4.7315	0.1428

[5] 我們於 2008 年 6 月得到 Google 的翻譯結果，同樣的題目，Google 分數較 2007 年時來得高。

外，最高的 BLEU 及 NIST 分數。我們可以觀察得到最高分數的範例樹組合，除了 Google 之外，多數皆落在結合所有範例樹語料的 G 組上，證明了在產生之大量且正確的「1：1 信心句對」後，增加了範例樹的數量，能夠幫助輔助式機器翻譯系統在翻譯時能夠正確地進行詞序的調動，讓產生的 TIMSS2003 中文翻譯句更為通順。

6. 結論

在專利說明書文本中，中英文翻譯並不如其它主題的文本整齊，往往因為主題的不同，文句複雜程度也改變極大，對文句對列的任務而言為極大的挑戰。事實上在專利說明書方面，我們缺乏中英文翻譯對照的標準答案，在利用人工進行對列檢驗的前提下，極耗費時間及人力，而我們需要對更多的語料進行對列實驗的評比，以取得更客觀公正的數據。我們利用精確率及召回率進行對列實驗的檢驗，代表在 1：1 的對列模式下，確實有很好的對列效果；藉由輔助式機器翻譯系統，利用產生的平行語料進行翻譯實驗及比較，證明大量且正確的語料能夠增進 TIMSS2003 翻譯的品質。我們也期望在未來能夠針對[15]提出的詞彙權重計算方式，如：***stf-idtf*** 值的計算，進行比較的實驗，並針對專利文書的文本，利用專利文書的平行語料進行中英翻譯的實驗，檢視本系統在專利文書翻譯方面的實質效果。

本中英文句對列系統之建置，透過自然語言處理等技術，以豐富的專業領域資源如：中英技術名詞對應表，改進現有文句對列的工具，藉由方法的改良，使系統能適用於不同文本主題的中英文句對列任務，能夠產生大量、正確且適合作為平行語料的中英文句對。

我們將現有對列工具進行改進，發展出適合於不同主題文本的文句對列技術，不僅止於專利說明書的文本，甚至在語文翻譯與學習的領域上，皆能夠透過本工具獲得豐富的中英文文句對列資源。

致謝

本研究承蒙國科會研究計畫 NSC-97-2221-E-004-007-MY2 的部份補助僅此致謝。我們感謝匿名評審對於本文初稿的各項指正與指導。雖然我們已經在從事相關的部份研究議題，不過限於篇幅因此不能在本文中全面交代相關細節。

參考文獻

[1] 中央研究院中文斷詞系統。http://ckipsvr.iis.sinica.edu.tw/。最後造訪：2009 年 8 月 8 日。

[2] 中央研究院現代漢語一詞泛讀系統。http://elearning.ling.sinica.edu.tw/CWordframe.html。最後造訪：2009 年 8 月 8 日。

[3] 自由時報中英對照讀新聞。http://iservice.libertytimes.com.tw/Service/english/。最後造訪：2009 年 8 月 8 日。

[4] 呂明欣、劉昭麟、高照明及張俊彥。針對數學與科學教育領域之電腦輔助英中試題翻譯系統，*第十九屆自然語言與語音處理研討會論文集*，407–421，2007。

[5] 科學人雜誌中英對照電子書。http://edu2.wordpedia.com/taipei_sa/。最後造訪：2009 年 8 月 8 日。

[6] 陳光華。超越資訊檢索的語言藩籬，*大學圖書館第二卷第一期*，87–99，1998。

[7] 國立編譯館學術名詞資訊網。http://terms.nict.gov.tw/。最後造訪：2009 年 8 月 8 日。

[8] 張智傑及劉昭麟。以範例為基礎之英漢 TIMSS 試題輔助翻譯，*第二十屆自然語言與語音處理研討會論文集*，308–322，2008。

[9] 經濟部智慧財產局。http://www.tipo.gov.tw/ch/。最後造訪：2009 年 8 月 8 日。

[10] 遠東高中・高職英文網站 - 歷年大考試題。http://www.hsenglish.com.tw/2009/teach/resource/exam_paper.asp。最後造訪：2009 年 8 月 8 日。

[11] 雙語網站知識管理平台新聞。
http://design.taiwannews.com.tw/demosite/2005/rdec/ver10/htm/se-learning01.htm。最後造訪：2009 年 8 月 8 日。

[12] 譯典通線上辭典。
www.dreye.com/tw/dict/dict.phtml。最後造訪：2009 年 8 月 8 日。

[13] HowNet。
http://www.keenage.com/。Last visited on 8 August 2009。

[14] Y. Liu, Q. Tan and K. X. Shen, *Modern Chinese Word Segmentation Specification and Automatic Segmentation Methods for Information Processing (in Chinese)* , Beijing: Qinghua University and Nanning: Guangxi Science and Technology Press, 1994.

[15] X. Ma, Champollion: A Robust Parallel Text Sentence Aligner, *Proceedings of the Fifth International Conference of the Language Resources and Evaluation*, 489–492, 2006.

[16] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[17] D. W. Oard, Alternative Approaches for Cross-Language Text Retrieval, *Working Notes of the American Association for Artificial Intelligence Spring Symposiums on Cross-Language Text and Speech Retrieval*, 131–139, 1997.

[18] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1986.

[19] M. Utiyama and H. Isahara, A Japanese-English Patent Parallel Corpus, *Proceedings of the Eleventh Machine Translation Summit*, 475–482, 2007.

[20] P. K. Wong and C. Chan, Chinese Word Segmentation based on Maximum Matching and Word Binding Force, *Proceedings of the Sixteenth International Conference of the Computational Linguistics*, 200–203, 1996.

意見持有者辨識之研究

A Study on Identification of Opinion Holders

李佳穎　古倫維　陳信希
國立臺灣大學資訊工程學系
{cylee, lwku}@nlg.csie.ntu.edu.tw , hhchen@csie.ntu.edu.tw

摘要

意見持有者辨識是從意見句中擷取出表述意見的人或組織，本研究將意見持有者辨識分爲作者意見辨識及意見持有者標記兩部分，作者意見辨識使用支援向量機處理，意見持有者標記使用條件隨機域處理。本研究提出的方法應用在 NTCIR7 MOAT 繁體中文語料的效能達到 F 值 0.734，是採取機器學習方法的參賽隊伍中效能最佳者，也相當接近目前最佳系統的效能。對於意見持有者辨識語料中標記歧異的情形，本研究加以分析，並提出使用此語料來訓練模型的方法。

Abstract

The identification of opinion holders aims to extract entities that express opinions in opinion sentences. In this paper, the task of opinion holder identification is divided into two subtasks: the identification of author's opinions and the labeling of opinion holders. Support vector machine is adopted to identify author's opinions, and conditional random field model (CRF) is utilized to label opinion holders. The proposed method achieves an F-score 0.734 in NTCIR7 MOAT task at traditional Chinese side. The proposed method achieves the best performance among participants who adopted machine learning methods, and also this performance was close to the best performance in this task. In addition, the ambiguous markings of opinion holders are analyzed, and the best way to utilize the training instances with ambiguous markings is proposed.

關鍵詞：意見持有者辨識，意見探勘，條件隨機域，支援向量機

Keywords: opinion holders identification, opinion mining, CRF, SVM.

一、緒論

意見代表人們對某個議題的主觀想法，人們常透過文章表述意見。隨著 Web2.0 的崛起，網路上出現大量、免費與即時的資料，使用者對文章中的意見很感興趣，但卻無法大量閱讀數以千萬計的資料。意見探勘 (opinion mining) 的技術可以幫助使用者自動分析文章中的意見， Kim 和 Hovy[1] 在 2004 年提出意見中包括意見傾向 (opinion polarity)、意見強度 (opinion strength)、意見持有者 (opinion holder) 及評論目標 (opinion target) 四個要素。意見傾向描述此意見是正面、中立或負面，意見強度描述此意見的語氣強弱，表述此意見的人或組織稱爲意見持有者，而討論的主題則稱爲評論目標。以例句 1 爲例，此句的意見傾向爲正面、意見強度爲強烈、意見持有者爲王建民、評論

目標爲打棒球。意見持有者通常會以一或多個詞的形式出現在意見句中，我們將這些詞稱爲意見持有者的代表詞，但有時意見持有者不會以詞的形式出現在意見句中，例如例句 2 是作者根據例句 1「王建民」的意見推論的意見，例句 2 的意見持有者爲文章作者。

例句 1：王建民非常喜歡打棒球

例句 2：王建民應該也喜歡打網球

在意見探勘中，意見持有者辨識的技術對於了解有哪些人或組織在表述意見、某個人或組織在哪些議題中發表過意見及兩個人或組織發表過的意見是否相似等相關資訊特別重要。意見持有者辨識可應用於社群網路分析中，找出社群網路中是否存在著一些意見領袖，他們的意見常被引用，也會影響其他使用者的意見。意見持有者辨識也可以應用在意見問答系統中，找出某些意見是由哪些意見持有者提出的，並進而藉由意見持有者的權威性與可靠度來輔助判斷答案的權威性與可靠度。

意見持有者辨識主要有三大挑戰：同指涉解析、巢狀結構及處理歧異的標記。意見持有者有時會以代詞 (Anaphor) 的形式出現在文句中，並指涉到前面的先行詞 (Antecedent) ，例如例句 3 中的「雙方」即是指涉到「美國」與「中共」。

例句 3：

據媒體報導，美國在與中共討論簽署停止以核武相互瞄準協議的問題，貝肯說，雙方過去就曾討論此事，前任國防部長裴利在中共國防部長遲浩田於一九九六年十二月訪美時就曾提起，後來雙方在其他會議中也曾討論。

意見句有時會有巢狀結構 (nested structure) ，以例句 3 爲例，文章作者引述「媒體報導」的內容，「媒體報導」的內容又引述「貝肯」的發言，意見持有者常會是子句的主詞，判斷意見持有者是哪一層結構中的主詞也是意見持有者辨識的一個重要議題。

標記意見持有者辨識所用的語料時，有時會出現標記歧異，不同的標記者可能會認爲文句的意見持有者爲不同的實體。以例句 3 爲例，一位標記者認爲意見持有者爲「國防部發言人貝肯/貝肯」，另一位標記者卻認爲意見持有者爲『美「中」/雙方』，從文句內容來看，意見持有者爲「國防部發言人貝肯/貝肯」，但深究背後的意義，貝肯是轉述『美「中」/雙方』的意見，兩種說法都沒錯，端看標記者的認知，也因此意見持有者可能被多個標記者標記出不同的答案。，如何利用標記歧異的語料也是意見持有者辨識的一大挑戰。

二、相關研究

Pang 和 Lee[2] 整理出意見探勘領域中重要的研究，意見持有者辨識的研究剛開始起步，研究團隊使用的方法主要可分爲以經驗法則 (heuristic rule) 爲基礎與以機器學習爲基礎兩種。

(一)、以經驗法則爲基礎的方法

以經驗法則爲基礎的方法中，Yohei 等人[3] 先使用名詞片語與語法特徵值，透過支援向量機，將意見持有者分爲文章作者與非文章作者，接著再透過語法規則，選出最有可能的具名實體，做爲答案的意見持有者，他們主要專注於處理英文與日文的語料。Xu

和 Wong[4] 提出的方法是先解決同指涉問題，再使用經驗法則擷取出意見持有者，使用的規則與標點符號、連接詞、字首 (prefix) 、字尾 (suffix) 與表述關鍵字相關，Xu 和 Wong 的方法是目前中文意見持有者辨識中效能最佳的，在 NTCIR7 多語意見分析評比項目的繁體中文語料上， F 值可達到 0.825。

(二)、以機器學習爲基礎的方法

以機器學習爲基礎的方法中，許多研究團隊使用最大熵法 (maximum entropy) 、支援向量機演算法 (support vector machine algorithm) 與條件隨機域模型 (conditional random field model) 等分類器解決此問題。

Kim 和 Hovy[5] 以最大熵法從新聞語料的文句擷取出意見持有者與意見評論目標，他們先找出意見詞 (opinion words) 與進行語意角色標注 (semantic role labeling) ，再找出代表意見持有者與意見評論目標的語意角色。

使用支援向量機演算法的研究團隊中， Kim 等人[6] [7] 先將意見持有者分爲文章作者、有同指涉情形與沒有同指涉情形三種，再使用詞彙與語法特徵值 (syntactic features) ，透過支援向量機，選出最有可能的意見持有者，Kim 等人的方法是目前英文意見持有者辨識中效能最佳的，應用在 NTCIR7 多語意見分析評比項目的英文語料上，F 值可達到 0.346。 Wu 等人[8] 則使用詞彙與詞性特徵值，透過 L2-norm 線性核心支援向量機，以類似具名實體辨識的方法解決中文的意見持有者辨識問題。

使用條件隨機域模型的研究團隊中， Breck 與 Choi 等人[9] [10] 使用詞彙、語法、字典 (dictionary-based) 及依存關係 (dependency relation) 特徵值，透過條件隨機域模型標記出最有可能的意見持有者。相形之下， Meng 和 Wang[11] 使用詞彙、詞性及表述關鍵字 (operator) 特徵值，而 Liu 和 Zhao[12] 則使用詞性、語意、依存關係、位置 (position) 及前後文 (contextual) 特徵值，透過條件隨機域模型標記出最有可能的意見持有者。

三、意見持有者辨識方法

本研究將意見持有者辨識分爲作者意見辨識及意見持有者標記兩個主要工作。本研究提出的流程包括前置處理程序、作者意見辨識程序、意見持有者標記程序、後置處理程序及結果合併程序五個部份。

(一)、針對斷詞與詞性標記的特殊處理

前處理程序包括斷詞、詞性標記、具名實體辨識及特徵值擷取。本實驗使用的是羅[13]研發的斷詞及詞性標記系統，爲了能夠更準確的斷出與意見持有者相關的具名實體，我們修改斷詞系統的人名模組並引入字典資訊。我們發現外國人名容易出現斷詞錯誤，所以我們著手修改斷詞及詞性標記系統的人名模組，來處理日文姓名長度與中文姓名長度不同的問題，我們在原本人名模組使用的姓氏列表中加入日本常見姓氏，名字長度的限制也從兩個字放寬爲三個字，使得系統能正確斷出如「高村正彥」、「兒玉源太郎」、「鈴木」等日本人名。

我們另外加入了職稱名、職業名、日本常見姓氏及台灣公營企業列表等字典。本研究的具名實體辨識是使用查詢詞典的方法，將人名、地名及組織名分別標上標籤。

(二)、作者意見辨識

作者意見辨識的目的是辨識意見句之意見持有者是否爲文章作者。本研究把作者意見辨識的問題視爲二元分類問題 (binary classification problem)，使用支援向量機來處理，實作上使用的套裝軟體是 Chang 和 Lin[14] 開發的 LIBSVM 。

表一、作者意見辨識使用的特徵值

特徵值類別	特徵值代號	特徵值描述
詞彙相關資訊	fHasI	本句有沒有「我」
	fHasWe	本句有沒有「我們」
	fNumI	本句有幾個「我」
	fNumWe	本句有幾個「我們」
詞性相關資訊	fHasPronoun	本句有沒有代名詞
	fHasManPronoun	本句有沒有人稱代名詞
	fNumPronoun	本句有幾個代名詞
	fNumManPronoun	本句有幾個人稱代名詞
具名實體資訊	fHasPer	本句有沒有人名詞
	fHasLoc	本句有沒有地名詞
	fHasOrg	本句有沒有組織名詞
	fHasNa	本句有沒有普通名詞
	fHasNb	本句有沒有專有名詞
	fHasNc	本句有沒有地方名詞
	fNumLoc	本句有幾個地名詞
	fNumOrg	本句有幾個組織名詞
	fNumPer	本句有幾個人名詞
	fNumNa	本句有幾個普通名詞
	fNumNb	本句有幾個專有名詞
	fNumNc	本句有幾個地方名詞
標點符號資訊	fHasExclamation	本句有沒有驚嘆號，例如：「！」或「!」
	fHasQuestion	本句有沒有問號，例如：「？」或「?」
	fHasColon	本句有沒有冒號，例如：「：」或「:」
	fHasLeftQuotation	本句有沒有上引號，例如：『「』或「【」
	fHasRightQuotation	本句有沒有下引號，例如：『」』 或「】」
文句組成資訊	fNumChar	本句有幾個字
	fNumWord	本句有幾個詞
	fNumSubsen	本句有幾個子句
意見相關資訊	fOperator	本句有沒有某個表述關鍵字

表二、意見持有者標記使用的特徵值

特徵值類別	特徵值代號	特徵值描述
詞彙相關資訊	fWord	本詞
詞性相關資訊	fPOS	本詞的詞性
	fIsPronoun	本詞是不是代名詞
	fIsNoun	本詞是不是名詞
具名實體資訊	fIsPer	本詞是不是人名
	fIsLoc	本詞是不是地名
	fIsOrg	本詞是不是組織名
標點符號資訊	fAfterParen	本詞是否在下引號之後兩詞，例如:『 」』 或「】」
	fBeforeColon	本詞是否在冒號之前兩詞，例如：「 ： 」或「:」
文句組成資訊	fNearSenStart	本詞是否靠近句首
	fSenLen	本詞所在句中的詞數
	fWordOrder	本詞在句中的詞序
	fWordPerc	本詞在句中詞序的百分比
前後文相關資訊	fNearVerb	同句中最靠近本詞的動詞
	fNearVerbPOS	同句中最靠近本詞的動詞詞性
	fDistNearVerb	同句中本詞到動詞的最短距離
意見相關資訊	fHasOpKW	同句中有沒有表述關鍵字
	fHasPosKW	同句中有沒有正面意見詞
	fHasNegKW	同句中有沒有負面意見詞
	fHasNeuKW	同句中有沒有中立意見詞
	fNearOpKW	同句中最靠近本詞的表述關鍵字
	fNearPosKW	同句中最靠近本詞的正面意見詞
	fNearNegKW	同句中最靠近本詞的負面意見詞
	fNearNeuKW	同句中最靠近本詞的中立意見詞
	fNearOpKWPOS	同句中最靠近本詞的表述關鍵字的詞性
	fNearPosKWPOS	同句中最靠近本詞的正面意見詞的詞性
	fNearNegKWPOS	同句中最靠近本詞的負面意見詞的詞性
	fNearNeuKWPOS	同句中最靠近本詞的中立意見詞的詞性
	fDistOpKW	同句中本詞到表述關鍵字的最短距離
	fDistPosKW	同句中本詞到正面意見詞的最短距離
	fDistNegKW	同句中本詞到負面意見詞的最短距離
	fDistNeuKW	同句中本詞到中立意見詞的最短距離

作者意見辨識使用的特徵值主要可分爲詞彙、詞性、具名實體、標點符號、文句組成及

意見相關資訊六種類別，表一列出作者意見辨識所有使用的特徵值，其中詞性、文句組成、意見相關資訊及標點符號中的驚歎號相關特徵值爲本研究首先提出的。

文句組成相關特徵值包括作者意見的文句在文句長度上的資訊。意見相關資訊則包含表述關鍵字 (operator) ，希望了解作者發表的意見中是否較常使用特定的表述關鍵字。表述關鍵字是用來表達意見的詞，通常爲動詞，例如：「說」、「報導」及「主張」等。標點符號中的驚嘆號常用來表達個人情緒的反應。

(三)、意見持有者標記

意見持有者標記的目的是辨識出意見持有者的代表詞，本研究將意見持有者標記問題視爲二元分類問題，試著使用決策樹演算法 (Decision Tree Algorithm) 解決，實作上使用的套裝軟體是 Mierswa 等人[15] 開發的 RapidMiner 中的 CHAID 決策樹演算法，CHAID 爲使用卡方檢定 (CHI Square Test) 的剪枝決策樹 (Pruned Decision Tree) 。本研究也將意見持有者標記的問題視爲序列標記問題 (sequential labeling problem) ，使用 Lafferty 等人[16] 提出的條件隨機域模型來標記出意見詞有者所涵蓋的詞彙，實作上使用的套裝軟體是 Kudo[17] 開發的 CRF++ 。

意見持有者標記使用的特徵值主要可分爲主要可分爲詞彙、詞性、具名實體、標點符號、文句組成、前後文及意見相關資訊七種類別，表二列出意見持有者標記所有使用的特徵值，其中前後文、意見相關資訊及詞性中的本詞是不是代名詞或名詞特徵值爲本研究首先提出的。

前後文相關資訊的特徵值考慮意見持有者的代表詞是否會較常與某些動詞搭配使用。意見相關資訊的特徵值則包含句中與意見關鍵字：表述關鍵字、正面意見詞、負面意見詞及中立意見詞相關的特徵值。正面意見詞爲表達正面意見立場的詞：如「同意」、「相信」、「成功」等。負面意見詞爲表達負面意見立場的詞：如「不會」、「反對」、「指控」等。中立意見詞爲表達中立意見立場的詞：如「未置評」、「兩難」、「可能」等。

NTCIR7 多語意見分析評估項目的訓練集較小，我們引入 Blum 和 Mitchell[18] 提出的協同訓練 (co-training) 來改善效能。協同訓練是半監督式機器學習方法 (semi-supervised learning method) ，能結合標記資料與未標記資料一起訓練模型。本研究挑選 CRF 預測信心值較高的實例，以文句爲單位回饋到訓練語料中，藉此提升系統效能。

(四)、後置處理

後置處理包含意見持有者爲詞組時之特殊處理及具名實體修復兩個部份。意見持有者標記會標示出本詞是不是意見持有者的一部分，但意見持有者常會由多個詞組成，因此需要根據標記結果將他們組合起來。本研究使用五種標籤來標記意見持有者：意見持有者的首詞 (H) 、尾詞 (T) 、中間詞 (I) 、本身爲意見持有者的單詞 (S) 及非意見持有者 (O) ， 因此 CRF 標籤集由 HITSO 五種標籤排列組合而成。

CHAID 分類器產生的結果爲 YES 與 NO 標籤，代表本詞是不是意見持有者的一部分，如果 CHAID 分類器產生的結果全部爲 NO 標籤時，系統會將意見持有者設爲文章作者。本系統使用下列兩條規則將這些詞組合成詞組：

規則 1：將連續名詞組合起來。
例如：「印度 (Nc) 總統 (Na) 瓦希德 (Nb) 」將組合成「印度總統瓦希德」

規則 2：使用連接詞、「的」字及頓號「 、 」將連續名詞組合起來。
例如：「瓦希德 (N) 、 (PAUSECATEGORY)柯林頓 (N) 與 (Caa) 小淵惠三 (N) 」將組合成「瓦希德、柯林頓與小淵惠三」

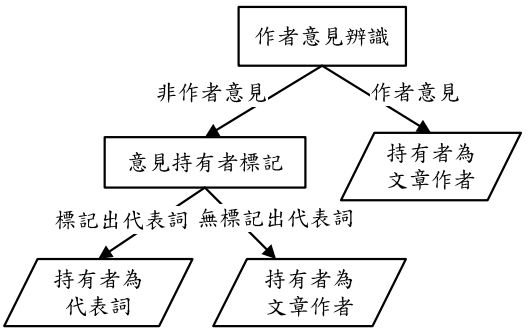
根據 CRF 分類器產生的結果標籤來組合意見持有者，組合規則爲先找到信心值最高的 H 標籤，再找到後面連續多個 I 標籤，最後再找到 T 標籤將這些詞組合起來。如果 CRF 分類器產生的結果全部爲 O 標籤時，系統會提報文章作者爲組合結果。

外國譯名的具名實體容易在斷詞時出現錯誤，這樣的錯誤可能造成辨識出不完整的意見持有者，因此我們可能需要修復意見持有者中的具名實體。本系統假設不完整的具名實體在文章中出現的頻率與完整的具名實體相同，所以本系統會將意見持有者標記結果跟前後字串接，測試組合成的新詞與原詞在文章中出現的頻率是否相同，相同則以新詞取代原詞。例句 4 爲具名實體修復過程的一例，原本意見持有者標記的結果是「蘇哈」，括號內的數字代表該詞在文章中出現的次數，修復後可輸出「蘇哈托」。透過這樣的具名實體修復方法，本系統可以將斷詞錯誤的具名實體修復爲完整的具名實體。

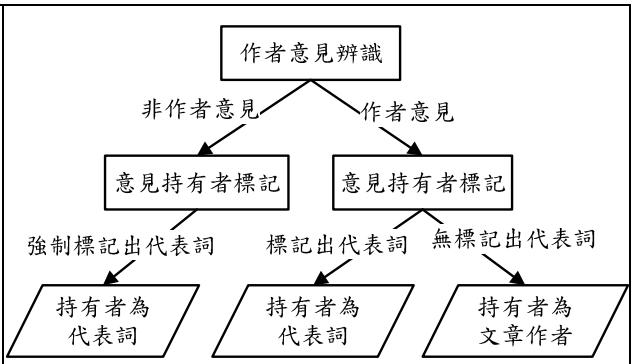
例句 4：印尼強人蘇哈托統治印尼卅二年
執行過程：蘇哈(16) → 蘇哈托(16) 頻率相同往後繼續
→ 蘇哈托統(1) 頻率不同改往前 → 人蘇哈托(3) 頻率不同結束

(五)、合併作者意見辨識與意見持有者標記之結果

作者意見辨識將意見句區分爲作者意見及非作者意見兩類，意見持有者標記則可標記出意見持有者的代表詞，本研究將此兩部分的結果合併，產生最後提報之意見持有者。本研究提出兩種結果合併策略，圖二、三爲結果合併策略 A 與 B 的示意圖。



圖一、結果合併策略 A



圖二、結果合併策略 B

結果合併策略 A 中，我們相信作者意見辨識判斷出的作者意見，作者意見辨識判斷出非作者意見的文句則透過意見持有者標記標出意見持有者的位置。結果合併策略 B 中，我們相信作者意見辨識判斷出的非作者意見，作者意見辨識判斷出是作者意見的部份，因爲不夠確定，再透過意見持有者標記二次檢查是否爲作者意見：如果沒有標出意

見持有者的代表詞，才提報該句的意見持有者是作者。作者意見辨識判斷出的非作者意見並沒有標記出意見持有者的代表詞爲何，所以我們透過意見持有者標記程序強制標記出代表詞，也就是從所有詞中找出最有可能代表意見持有者的詞。

透過前置處理、作者意見辨識、意見持有者標記、後置處理及結果合併五個程序，我們就可以提報最後辨識出之意見持有者。

四、實驗與討論

本節將介紹本實驗使用的語料與資源，並討論作者意見辨識實驗、意見持有者標記實驗及意見持有者辨識整體實驗的結果。

(一)、NTCIR 7 多語意見分析評比項目介紹

本實驗使用的語料爲 NTCIR 7 多語意見分析評比項目中繁體中文的語料庫， NTCIR (NII Test Collection for IR Systems) 是日本國家資訊研究所 (National Institute of Informatics，NII) 所策劃主辦的國際評比會議，是世界三大資訊檢索會議之一。多語意見分析評比項目 (Multilingual Opinion Analysis Task, MOAT) 是其中一個評比項目，Seki 等人[19] 對此評比項目有詳細的介紹。

多語意見分析評比項目提供英文、日文、繁體中文與簡體中文的語料庫，語料庫中提供相關句、意見句、意見傾向、意見強度、意見持有者與評論目標的標記。語料庫分爲訓練集與測試集， NTCIR7 訓練集包括 3 個主題、1,509 個文句、944 個意見句，NTCIR7 測試集包括 14 個主題、4,665 個文句、2,174 個意見句，參賽者們會以文句爲單位進行意見分析。因爲 NTCIR7 訓練集較小，本實驗的訓練語料加入 NTCIR6 意見分析試驗評比項目 (Opinion Analysis Pilot Task) 繁體中文的測試集，NTCIR6 意見分析試驗評比項目是 NTCIR 7 多語意見分析評估項目的前身，語料庫中提供相關句、意見句、意見傾向、意見持有者的標記。 NTCIR6 測試集包括 29 個主題、9,240 個文句、5,453 個意見句，我們利用語料庫中關於意見句與意見持有者的標記當作我們的實驗語料。

(二)、實驗資源

本研究使用的實驗資源包括意見詞詞典與具名實體詞典。意見詞詞典的部份包含標記者從 NTCIR 7 多語意見分析評比項目的訓練集中標記出表述關鍵字、正面意見詞、負面意見詞及中立意見詞等意見詞，也使用 Ku 和 Chen[20] 開發的台大意見詞詞典 (NTUSD) ，本研究將這些意見詞詞典應用於特徵值擷取。

本研究引入人名詞典、地名詞典及組織名詞典三類具名實體詞典，使用的詞典包含百萬人名字典、中文詞庫、中央社譯名檔、國立編譯館專業字典、日本常見七千個姓氏、教育部地名譯名詞典、外國地名譯名及台灣公營企業列表。這些具名實體詞典則應用於具名實體辨識。

(三)、作者意見辨識實驗

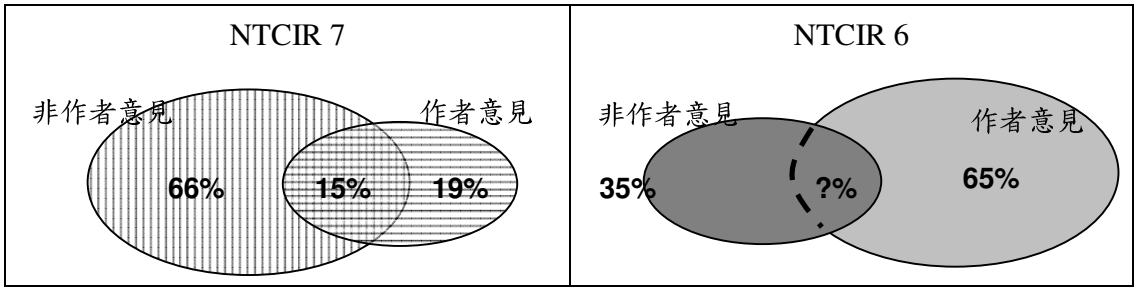
本實驗的目的是辨識意見句之意見持有者是否爲文章作者，本實驗使用的訓練語料是 NTCIR7 訓練集及 NTCIR6 測試集，測試語料是 NTCIR7 測試集中的寬鬆意見句。

NTCIR7 測試集中的意見句判定分爲嚴格意見句與寬鬆意見句，嚴格意見句的條件是三位標記者都將此句標記爲意見句，寬鬆意見句的條件則是三位標記者中，有兩位以上將此句標記爲意見句。

本實驗判斷此句的意見持有者是不是文章作者，也就是此句是不是作者意見，我們使用精確率 (Precision) 、召回率 (Recall) 、F 值 (F-score) 及正確率 (Accuracy) 來評估系統效能。我們研究語料之後發現部分文句會有作者意見標記歧異的情形，例如例句 5 的第二句，一種說法認爲意見持有者爲「臺灣」，另一種說法認爲意見持有者標記爲文章作者。

例句 5：
第一句：臺灣官方已排除核武選擇，並將最終安全託付於美國，
第二句：因爲臺灣了解，核武之路將根本破壞與美國之間的關係。

我們將標記的情形加以分析。圖三是語料庫中作者意見比例示意圖，我們發現在 NTCIR7 的語料庫中，被部分標記者標記爲作者意見，其他標記者標記爲非作者意見的標記歧異語料佔 15%，與被所有標記者標記爲作者意見的比例 (19%) 相距不遠。



圖三、語料庫中作者意見比例示意圖

在 NTCIR6 的語料庫中，不只是意見句，所有的文句都會被標上發表者，代表表示這個說法的人或組織， NTCIR6 的意見持有者標記方法是如果可以從文章中標記出意見持有者就標記，如果不能，則意見持有者爲文章作者，所以無法得知 NTCIR6 的語料庫中標記出現歧異的文句占全部文句的比例。

我們從實驗中發現使用意見句當訓練語料的 F 值爲 73.24%，比使用全部文句當訓練語料的 F 值高了 8.04%。使用 NTCIR 6 加 NTCIR 7 訓練集的 F 值爲 79.98%，也比單獨使用其中一個訓練集爲高，所以本實驗使用 NTCIR 6 加 NTCIR 7 訓練集中的意見句作爲訓練語料。

表三、標記歧異的語料對作者意見辨識效能的影響

設定	精確率	召回率	F 值	正確率
視爲作者意見	***69.68%***	***93.85%***	***79.98%***	***83.49%***
視爲非作者意見	64.87%	95.94%	77.40%	80.31%
不列入訓練語料	50.52%	91.53%	65.10%	77.28%

本實驗的實驗設定共有三種：將標記歧異的語料視爲作者意見、將標記歧異的語料視爲非作者意見、及標記歧異的語料不列入訓練語料。表三顯示在這三種實驗設定下，標記歧異的語料對作者意見辨識效能的影響。

實驗結果顯示，視爲作者意見的設定效能最佳： F 值爲 79.98%，其次爲視爲非作者意見的設定： F 值爲 77.40%，效能最差的是不使用標記歧異的語料： F 值爲 65.10%。從本實驗結果可以發現，使用標記歧異的語料能提升系統效能，將標記歧異的語料視爲作者意見加入訓練效能最佳。

(四)、意見持有者標記實驗

本實驗的目的是辨識出意見持有者的代表詞，以供後置處理程序組合成意見持有者。本實驗使用的訓練語料是 NTCIR7 訓練集，測試語料是 NTCIR7 測試集中的嚴格意見句與寬鬆意見句。本實驗將意見持有者標記的結果組合出最後的意見持有者，再使用正確數、錯誤數、 set F 值 (set F-score) 來評估系統效能。 NTCIR7 的評估方式會先評估每個參賽隊伍正確提報的意見句數，再評估每個參賽隊伍正確提報的意見持有者，參賽隊伍提報的意見句數等於提報的意見持有者數，也等於答案的意見持有者數，此時精確率、召回率、 F 值與正確率會相等，因此我們不使用精確率、召回率、 F 值，而使用 set F 值來評估效能， set F 值的定義如下：

$$setF\text{值} = \frac{\text{系統正確提報意見持有者的意見句數}}{\text{系統正確提報的意見句數}} \quad (1)$$

實驗比較兩種分類演算法：用來解決二元分類問題的決策樹演算法 CHAID ，及用來解決序列標記問題的條件隨機域模型 CRF 。 CHAID 使用的意見持有者標記標籤是 YES 與 NO 標籤， CRF 使用的意見持有者標記標籤是意見持有者的首詞 (H) 、意見持有者的中間詞 (I) 及非意見持有者組成詞 (O) 。表四顯示使用不同分類演算法對作者意見辨識效能的影響。

表四、分類演算法對意見持有者標記效能的影響

	分類演算法	正確數	錯誤數	set F 值
嚴格意見句	CHAID	564	605	48.16%
	CRF	817	351	69.89%
	CRF+CHAID	825	344	70.57%
寬鬆意見句	CHAID	981	967	50.31%
	CRF	1317	631	67.57%
	CRF+CHAID	1322	627	67.83%

嚴格意見句部份的評估中， CRF 的 set F 值爲 69.89%，比 CHAID 高 21.73%，效能好很多，寬鬆意見句部份的評估也可以得到類似的結果。接著我們將 CHAID 預測出來的結果當作 CRF 的一個特徵值再重新訓練模型，也可以小幅提升系統效能。原因可能因爲 CRF 使用的意見持有者標記標籤較多，增加的 H 標籤有助提升系統效能，也可能因爲 根據 CRF 標籤結合詞組的效能比 根據 CHAID 的結果再用規則連接的效能

好。

寬鬆意見句部份的評估中，本研究能達到的最佳效能是 set F 值 67.83%。在標記結果錯誤的實例中，有 13% (254 句) 是正確答案的意見持有者爲單詞或詞組，但系統標記出之單詞或詞組與正確答案不符，以下將分析標記結果錯誤的實例，並提出解決的方法。我們根據正確答案與系統標記出之答案的位置分析，將主要的標記錯誤分爲 6 類：

1.答案無關聯

無法判斷正確答案與系統標記出之答案之間的關聯，此類佔標記錯誤的 29.1%。

2.多擷取前後一詞

系統標記出之答案包含正確答案，但卻又多將前後一詞判斷爲意見持有者的一部分，例如：也許**蘇哈托**、**魯斯曼**日前、**他們**可以。範例中，**粗體字**代表正確答案，標底線的字代表系統標記出之答案，此類佔標記錯誤的 18.1%。

3.擷取出頭銜但未擷取出其後的人名

系統標記出之答案包含正確答案中之頭銜，但卻沒有擷取出頭銜後的人名，例如：**科索伏著名塞裔領袖特拉伊科維契**，此類佔標記錯誤的 8.3%。

4. 擷取出形容詞但未擷取出其後的普通名詞

系統標記出之答案包含正確答案中前面的形容詞但未擷取出其後的普通名詞，例如：**該裁決**、**國際停火觀察團**、**美國全國公共廣播電台**，此類佔標記錯誤的 7.5%。

5. 額外擷取出其他非答案詞

系統標記出之答案包含正確答案，但卻又額外擷取出其他的詞，例如：**狄蘭**在記者會、**他**祝賀巴勒斯坦的科學家，此類佔標記錯誤結果的 5.5%。

6. 具名實體擷取不完整

正確答案包含系統標記出之答案，但具名實體部份擷取得不完整，例如：**德州農工大學的複製專家韋斯特休生**、**車燈廠堤維西**，此類佔標記錯誤結果的 4.7%。

根據這些標記結果錯誤的類別，我們提出幾種方法來改善系統效能。大部分的類別都有意見持有者詞組中首詞、尾詞不明確的問題，所以我們提出增加意見持有者標籤的方法。從實驗中發現，使用 HIO 標籤集在嚴格意見句部份的評估中可得到最佳的 set F 值 70.57%。針對第 6 類具名實體擷取不完整的問題，我們也提出具名實體修復方法來解決這個問題，從實驗中發現，加入協同訓練與具名實體修復在嚴格意見句部份的評估中可得到最佳的 set F 值 72.03%，效能提升了 1.46%。

(五)、意見持有者辨識整體實驗

本實驗的目的是探討使用不同結果合併策略對意見持有者辨識效能的影響，實驗設定爲結果合併策略 A 與結果合併策略 B ，表五顯示使用不同結果合併策略的系統效能。

嚴格意見句部份的評估中策略 B 的 set F 值爲 73.40%，比策略 A 高了 2.48%，寬鬆意見句部份的評估也可以得到類似的結果。結果顯示作者意見辨識程序較擅長判斷非作者意見，可能與我們使用較多與非作者意見相關的特徵值有關，實驗結果也顯示策略 B 可以達到最佳效能。

表五、結果合併策略對意見持有者辨識效能的影響

	結果合併策略	正確數	錯誤數	set F 值
嚴格意見句	策略 A	829	340	70.92%
	策略 B	***858***	***310***	***73.40%***
寬鬆意見句	策略 A	1338	611	68.65%
	策略 B	***1372***	***576***	***70.40%***

(六)、與 NTCIR 7 參賽隊伍比較

我們將本系統效能與 NTCIR7 參賽隊伍的效能比較， NTCIR7 的評估方式分爲兩種，一種評估系統提報正確的意見句，也就是我們在意見持有者標記中使用的評估方式，另一種則評估 NTCIR7 語料中所有的意見句。

表六顯示本系統與 NTCIR 7 參賽隊伍的效能比較。 NTCIR7 意見持有者擷取評比項目的參賽隊伍包含香港中文大學、北京大學、龍捲風科技及台灣大學四隊，香港中文大學使用經驗法則方法、北京大學使用條件隨機域模型、龍捲風科技使用支援向量機、台灣大學使用決策樹演算法。

嚴格意見句部份，以系統提報正確的意見句評估，本系統的最佳效能爲 F 值 73.40%。香港中文大學的效能最佳： F 值爲 82.30%，比本系統高了 8.90%，但本系統的效能也比其他使用機器學習方法的隊伍高了 15.09%以上。以所有意見句數評估中，本系統與香港中文大學的效能差距拉近到 5.16%。比較嚴格意見句與寬鬆意見句的評估，可以發現本系統與其他系統不同，較擅長於辨識嚴格意見句的意見持有者，換句話說，本系統擅長於辨識出較無爭議的意見持有者，也就是較爲可靠的意見持有者。

表六、意見持有者辨識整體效能—與 NTCIR 7 參賽隊伍比較

	參賽隊伍	猜對意見句數	以系統猜對意見句數評估			以所有意見句評估		
			精確率	召回率	F 值	精確率	召回率	F 值
嚴格意見句	香港中文大學	757	***82.30%***	***82.30%***	***82.30%***	***19.88%***	***49.52%***	***28.38%***
	北京大學	880	57.84%	57.84%	57.84%	13.03%	40.53%	19.72%
	龍捲風科技	1213	54.91%	54.91%	54.91%	8.22%	52.95%	14.23%
	台灣大學	1169	48.16%	48.16%	48.16%	8.14%	44.90%	13.78%
	本系統	1169	**73.40%**	**73.40%**	**73.40%**	**12.38%**	**68.31%**	**20.97%**
寬鬆意見句	香港中文大學	1134	***82.54%***	***82.54%***	***82.54%***	***29.92%***	***43.05%***	***35.31%***
	北京大學	1364	58.72%	58.72%	58.72%	20.51%	36.84%	26.35%
	龍捲風科技	2070	56.47%	56.47%	56.47%	16.78%	40.02%	23.65%
	台灣大學	1948	50.31%	50.31%	50.31%	14.43%	53.73%	22.75%
	本系統	1948	**70.40%**	**70.40%**	**70.40%**	**19.80%**	**63.11%**	**30.15%**

五、結論與未來展望

本研究提出一個以機器學習方法爲基礎的意見持有者辨識方法，並且依照此方法實作出一套意見持有者辨識系統。

本研究根據意見持有者的分類將意見持有者辨識分爲作者意見辨識及意見持有者標記兩部分。在作者意見辨識中，本研究提出詞彙相關資訊、詞性相關資訊、具名實體資訊、關鍵符號資訊、文句組成資訊及意見相關資訊等特徵值，並使用支援向量機來解決此問題。在意見持有者標記中，本研究提出詞彙相關資訊、詞性相關資訊、具名實體資訊、關鍵符號資訊、文句組成資訊、前後文相關資訊及意見相關資訊等特徵值，並使用條件隨機域模型並提出不同的標記方式來解決此問題。本研究提出協同訓練來解決訓練語料過少的問題，並提出結果合併策略以提升意見持有者辨識效能。

本研究實作出一套意見持有者辨識系統。本系統在 NTCIR7 多語意見分析評比項目繁體中文語料中可以達到 F 值爲 0.734 的效能，是使用機器學習方法中效能最佳的，也相當接近目前最佳系統的效能。目前效能最佳的方法是使用經驗法則解決本問題，經驗法則較難重製與驗證，但研究者很容易就可以重製與驗證本研究提出的機器學習方法。本研究分析意見持有者辨識訓練語料中標記歧異的情形，並提出最佳的應用方式，本研究也分析系統辨識之錯誤結果，並以具名實體修復及意見持有者尾詞標記的方法來改善錯誤情況。

最終我們希望能將意見持有者辨識的結果與其它意見探勘的結果結合，整合成一套能自動擷取出意見句的意見傾向、意見強度、意見持有者及意見評論目標的意見探勘系統，以提供使用者更有用的資訊。

參考文獻

[1] S. M. Kim and E. Hovy. “Determining the sentiment of opinions.” Proceedings of the COLING conference, pp.1367-1374 , 2004

[2] B. Pang and L. Lee. “Opinion mining and sentiment analysis” Foundations and Trends in Information Retrieval, Vol. 2, pp. 1-135, 2008S. M. Kim and E. Hovy. “Determining the sentiment of opinions.” Proceedings of the COLING conference, pp.1367-1374 , 2004

[3] Y. Seki, N. Kando and M. Aono. “Multilingual opinion holder identification using author and authority viewpoints.” Journal of Information Processing and Management, pp. 189-199, 2009 [co-training]

[4] R. Xu and K. F. Wong. “Coarse-Fine opinion mining – WIA in NTCIR-7 MOAT task.” Proceedings of the Seventh NTCIR Workshop, pp. 307-313, 2008

[5] S. M. Kim and E. Hovy. “Extracting opinions, opinion holders, and topics expressed in online news media text.” Proceedings of the Workshop on Sentiment and Subjectivity in Text at the joint COLING-ACL conference, pp. 1-8, 2006

[6] Y. Kim, Y. Jung and S. H. Myaeng. “Identifying opinion holders in opinion text from online newspapers.” International Conference on Granular Computing, pp. 699-702, 2007

[7] Y. Kim, S. Kim and S. H. Myaeng. “Extracting topic-related opinions and their targets in NTCIR-7.” Proceedings of the Seventh NTCIR Workshop, pp. 247-254, 2008

[8] Y. C. Wu, L. W. Yang, J. Y. Shen, L. Y. Chen and S. T. Wu. “Tornado in multilingual opinion analysis: a transductive learning approach for Chinese sentimental polarity recognition.” Proceedings of the Seventh NTCIR Workshop, pp. 301-306, 2008

[9] E. Breck and Y. Choi and C. Cardie. “Identifying expressions of opinion in context.” Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 2683-2688, 2007

[10] Y. Choi, C. Cardie, E. Riloff and S. Patwardhan. “Identifying sources of opinions with conditional random fields and extraction patterns.” Proceedings of EMNLP conference, pp. 355-362, 2005

[11] X. Meng and H. Wang. “Detecting opinionated sentences by extracting context information.” Proceedings of the Seventh NTCIR Workshop, pp. 268-271, 2008

[12] K. Liu and J. Zhao. “NLPR at Multilingual Opinion Analysis Task in NTCIR7.” Proceedings of the Seventh NTCIR Workshop, pp. 226-231, 2008

[13] 羅永聖, “結合多類型字典與條件隨機域之中文斷詞與詞性標記系統研究” 國立台灣大學碩士論文, 2008

[14] C. C. Chang and C. J. Lin. “LIBSVM: a library for support vector machines”, 2001

[15] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz and T. Euler. “YALE: rapid prototyping for complex data mining tasks” In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 935-940, 2006

[16] J. Lafferty, A. McCallum, F. Pereira. “Conditional random fields: probabilistic models for segmenting and labeling sequence data” In Proceedings of International Conference on Machine Learning, pp. 282-289, 2001

[17] T. Kudo, “CRF++: yet another CRF toolkit.” http://crfpp.sourceforge.net/ , 2003

[18] A. Blum and T. Mitchell. “Combining labeled and unlabeled data with co-training.” Conference on Computational Learning Theory, pp. 92-100, 1998

[19] Y. Seki, D. K. Evans, L. W. Ku, L. Sun, H. H. Chen and N. Kando. ”Overview of multilingual opinion analysis task at NTCIR-7.” Proceedings of the Seventh NTCIR Workshop, pp.185-203, 2008

[20] L. W. Ku and H. H. Chen. “Mining opinions from the web: beyond relevance retrieval.” Journal of American Society for Information Science and Technology, pp. 1838-1850, 2007

聲調對嗓音起始時間的影響：以國語和客語爲研究對象
Tonal effects on voice onset time: Stops in Mandarin and Hakka

彭瑞鳳 Jui-Feng Peng
陳麗美 Li-mei Chen
林依雲 Yi-Yun Lin

國立成功大學外國語文學系（所）
Department of Foreign Languages & Literature
National Cheng Kung University
leemay@mail.ncku.edu.tw

Abstract

This study examines the influence of lexical tone upon voice onset time (VOT) in Mandarin and Hakka. Examination of VOT values for Mandarin and Hakka word-initial stops /p, t, k, p^h, t^h, k^h/ followed by three vowels /i, u, a/ in different lexical tones revealed that lexical tone has a significant influence on the VOTs. The result is important because it suggests that future studies should take its influence into account when studying VOT values for stops in tonal languages. In Mandarin, stops' VOTs, ordering from the longest to the shortest, are in Tone 2, Tone 3, Tone 1, and Tone 4: this sequence is the same as Liu, Ng, Wan, Wang, and Zhang's (2008) [1] results. However, later it was found that the sequence results from the existence of non-words. Because in order to produce non-words correctly, participants tended to pronounce them at a lower speed, especially those in Tone 2. Therefore, we further examined the data without non-words, in which no clear sequence had been found. For Hakka, Post hoc tests (Scheffe) show that aspirated stops in Tones 4 and 8 have significantly shorter VOT values than they have in other tones.

Keywords: Voice onset time, Mandarin tones, Hakka stops, Mandarin stops

1. Introduction

The aim of this paper is to explore whether lexical tones influence the VOT values for word-initial stops. This issue is important because VOT is considered as a reliable phonetic feature to differentiate consonant stops ([2], [3], [4], [5], [6], [7]) and recently it has been used to study the language production of patients with language deficits or disorders ([8], [9]). Among the languages being investigated, some are tone languages, i.e. Mandarin, Cantonese, and Taiwanese. In a tonal language, the duration of each lexical tone is slightly different. Consequently, it is possible that lexical tone will affect stop's voice onset time. However, few studies have taken this factor into consideration while studying tone languages. It is hoped that with data from Mandarin and Hakka, we can establish the groundwork for future studies related to VOTs in tonal languages. If lexical tone does have an influence on the VOT, it should be taken into account when creating stimulus words in future studies for tonal languages, thereby rendering studies more valid and reliable.

1.1 Voice onset time

Lisker and Abramson (1964) [2] have defined voice onset time (VOT) as the

temporal interval from the release of an initial stop to the onset of glottal pulsing for a following vowel. It has been considered as a reliable phonetic cue to categorizing the stop consonants, i.e. voiced vs. voiceless or unaspirated vs. aspirated, in various languages ([2], [3], [4], [5], [6], [7], [10]). Additionally, by comparing VOT values for stops produced by native and non-native speakers for specific languages, researchers have provided some suggestions for language learning and teaching ([6], [11], [12]). Moreover, recently researchers have studied aphasia, apraxia and stuttering patients' production deficits by observing their VOT values for stops ([8], [9]).

1.2 Factors affecting voice onset time

When investigating stops, researchers found that the VOT values for stops varied in relation to the place of articulation. Cho and Ladefoged (1999) [4], sorted out researchers' findings, have claimed that the further back the closure, the longer the VOT ([2], [4], [6], [13]). That is velar stops have the longest VOT values, alveolar stops the intermediate values, and bilabial stops have the shortest values. However, there are some exceptions. Alveolar stops in Tamil, Cantonese, Eastern Armenian, Hungarian, Japanese, and Mandarin, have shorter VOTs than bilabial stops ([2], [3], [5], [7], [12], [14]).

Liu et al. (2008) [1] speculated that the VOT durations may be affected by tone, because different tones have different fundamental frequencies and pitch levels, which are determined mainly by the tension of the vibrating structure. In order to achieve different levels of tension, different amounts of time might be needed. Consequently, the VOT values may vary when they are in different lexical tone. Only a few studies have tried to examine whether lexical tone influences VOT values. For example, Liu et al. (2008) [1] studied the effect of tonal changes on VOTs between normal laryngeal and superior esophageal speakers of Mandarin Chinese, and reported that for normal laryngeal speakers there are significant differences of VOT values caused by lexical tones. In addition, stops in Tone 4 have significantly shorter mean VOT values than stops in Tones 2 and 3. The study by Liu et al. [1] is a pioneering piece of work in this field, but more evidence is still needed. Therefore, by carrying out a systematic study with respect to the influence of lexical tone for stop's VOT using two tonal languages, i.e. Mandarin and Hakka, we try to verify previous findings in order to provide references for future linguistic studies on tonal languages.

1.3 The features of Mandarin and Hakka

Mandarin Chinese and Hakka are tonal languages, in which a word's meaning can be changed by the tone in which it is pronounced. Chao (1967) [15] suggested a numerical notation for lexical tones: dividing a speaker's pitch range into four equal intervals by five points: 1 low, 2 half-low, 3 middle, 4 half-high, and 5 high. The numerical notation indicates how the pitches of a lexical tone change. For example, the numerical notation for Tone 2 in Mandarin is 35, which represents that the pitch will go from middle to high. Table 1 reveals the numerical notation for each lexical tone in Mandarin and Hakka. In Mandarin, there are four contrasting lexical tones, Tone 1 (high–level), Tone 2 (mid-rising), Tone 3 (falling-rising), and Tone 4 (high-falling). Sixian Hakka has six contrasted lexical tones, Tone 1 (24), Tone 2 (31), Tone3 (55), Tone 4 (32), Tone 5 (11), and Tone 8 (55). The pitch values for Tone 3 and Tone 7 are the same, therefore Tone 7 has been omitted. Although there are regional differences for Hakka, Sixian Hakka was chosen as it is the most widely used Hakka dialect in Taiwan.

Table 1. The numerical notations for lexical tones in Mandarin [15] and Hakka [16].

Lexical Tone		1	2	3	4	5	7	8
Numerical notation	Mandarin	55	35	214	51			
	Hakka	24	31	55	32	11	(55)	55

Note: Those which are underlined represent pitches that are short and rapid.

Mandarin Chinese and Hakka have their specific tone sandhi rules and one example from each language is listed below. In Mandarin, Tone 3, which has the longest duration among the four lexical tones, will become Tone 2 while it is followed by another Tone 3 [17]. The tone sandhi rule for Sixian Hakka is as follows: Tone 1 will become Tone 5, when it precedes Tone 1, Tone 3, or Tone 8 [18]. Therefore, tone sandhi rules are taken into consideration when making stimulus words, and the combinations that might cause tonal change will be avoided.

Mandarin
Tone 3 →Tone 2 / _____ { Tone 3 }

Sixian Hakka
Tone 1→Tone 5 / _____ { Tone 1, Tone 3, Tone 8 }

2. Methodology

Mandarin and Hakka word-initial stops, unaspirated /p, t, k/ and aspirated /p^h, t^h, k^h/, in combination with three vowels /i, u, a/ were studied. Except for participants and stimulus words, the methodology employed for both languages was the same.

2.1 Participants

Mandarin and Hakka participants were different. For Mandarin, there were fifteen male and fifteen female native speakers recruited from college students and staff from an elementary school in Tainan City. All the participants grew up in Taiwan, with no hearing and speech defects. Their ages ranged from 23 to 33 years (mean = 27.2 years). As for the Hakka, there were twenty-one participants, eleven men and ten women, from Miaoli, Pingtung, and Taoyuan County. Their average age was fifty-one, the oldest being eighty, and the youngest thirty-six. As it was not easy to find fluent Hakka speakers their age range was quite wide.

2.2 Stimuli and procedure

The speech stimuli in both language were combination of six stops /p, t, k, p^h, t^h, k^h/ and three vowels /i, u, a/, i.e., 18 combinations. They were /pi/, /pu/, /pa/, /ti/, /tu/, /ta/, /ki/, /ku/, /ka/, /p^hi/, /p^hu/, /p^ha/, /t^hi/, /t^hu/, /t^ha/, /k^hi/, /k^hu/, and /k^ha/. For Mandarin there were four contrasting lexical tones, thus 72 monosyllabic words were created in total. Among them, 18 combinations do not exist in Mandarin. As for Hakka, there were six contrasting lexical tones in Sixian Hakka, hence there were 108 monosyllabic words obtained. Among these stimulus words, 12 words do not actually exist in Hakka. Chen et al. (2007) [14] has claimed that disyllabic words can create a more natural-like context for participants. Therefore, in order to make speakers produce the words more naturally, all the stimulus words were followed by another word and would become meaningful disyllables. For example, Mandarin word, /pi/, was followed by another word, /p^huo/ to become the existing disyllable, /pi p^huo/

(force). Some stimulus words in Hakka were tri-syllabic, due to the fact that no meaningful disyllables were found.

The stimulus words were arranged randomly, and the participants were asked to read it out loud at a normal speed. After finishing, the participants were asked to read out the words for the second time. Therefore, two groups of data were gathered for each participant. All the speech was recorded by a 24 bit WAV recorder, connected with a AKG head-worn cardioid condenser vocal microphone positioned of approximately 10~15 cm from the participant's mouth in a quiet room.

2.3 Data Measurement and analysis

After recording, data were edited into individual files and analyzed using the Praat software. VOT, measured in milliseconds (ms), was obtained by measuring the temporal interval between the beginning of the release burst and the onset of the following vowel as shown in Figure 1. The values of both the waveform and spectrogram were recorded, but the VOTs were determined primarily through waveform analysis; the values in the spectrogram were provided as references. If the values in waveform differed from the values in the spectrogram by more than five milliseconds, the data were re-measured to verify accuracy.

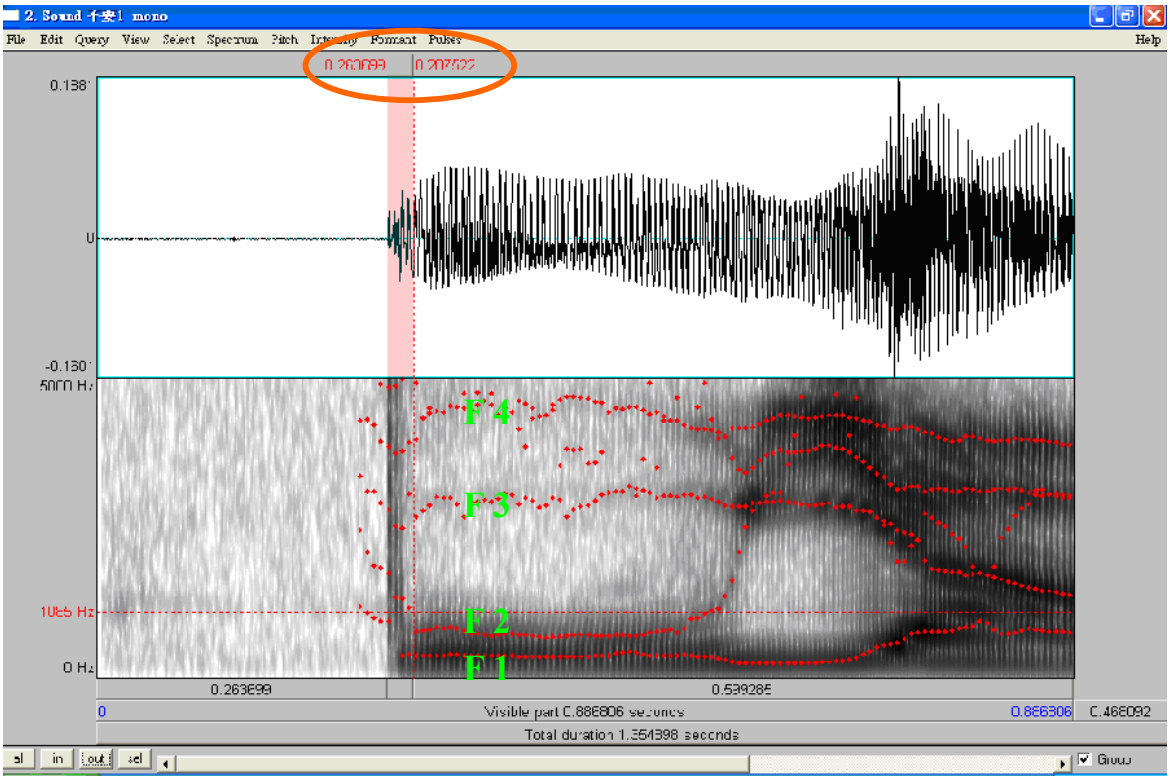


Figure 1. The spectrogram and waveform for the Mandarin word /pu iau/ 'don't want'. The values in the circle are the starting and endpoints of the VOT in the spectrogram.

When analyzing the data, the VOT values for the mispronounced words were excluded, and the data for Hakka /pi/ in Tone 8 were not analyzed because of wrong word-choosing. ANOVA test was used to examine whether or not there is a significant influence on stop's voice onset time. In addition, the differences between the

examined targets were analyzed by Post Hoc tests (Scheffe). The measurements of stops' VOT values were made by the same investigator. Furthermore, randomly selected 10% of each recording were re-measured by another investigator to verify the reliability of the results. Therefore, 7 Mandarin words and 11 Hakka words for each recording were re-measured. The inter-rater reliability was then examined by Pearson's product-moment correlations.

3. Results

Pearson's product-moment correlations indicated high inter-rater agreement for both the Mandarin and Hakka data (Mandarin: r = .995, p<.001; Hakka: r = .978, p<.001). This indicates that the measurements were reliable throughout. It was found that the mean VOTs for Mandarin stops get longer due to the existence of non-words. Therefore, the data excluding non-words was further examined to verify the results. For Hakka, there is no clear difference because most of the non-words were pronounced incorrectly. Therefore, most of the values of Hakka non-words are not included in the analysis.

3.1 Lexical tone and VOT in Mandarin

Mandarin stops' mean VOT values and standard deviations in each lexical tone are shown in Table 2. ANOVA test reveals that lexical tones have significant influences on the VOTs for stops (F(3,1040)=2.681, p<.05 for unaspirated stops; F(3,1040)=8.934, p <.001 for aspirated stops). When examining the data with non-words, it is shown that for both unaspirated and aspirated, stops in Tone 2 have the longest mean VOTs and stops in Tone 4 have the shortest mean VOTs. Stops' VOT values ordering from the longest to the shortest are in Tone 2, Tone 3, Tone 1, and Tone4. Post hoc tests revealed that aspirated stops in Tone 4 have significantly shorter mean VOTs than stops in Tone 2 and Tone 3 (p< .05).

Table 2. Mandarin stops' mean VOT values in individual lexical tones. All measurements are in milliseconds (ms).

	With non-words				Without non-words			
	unaspirated stops		aspirated stops		unaspirated stops		aspirated stops	
	mean	*SD*	mean	*SD*	mean	*SD*	mean	*SD*
Tone 1	20.20	(11.90)	92.72	(25.53)	17.71	(9.95)	88.69	(20.4)
Tone 2	21.10	(12.68)	101.02	(30.21)	13.99	(6.03)	89.47	(23.31)
Tone 3	20.89	(13.35)	97.03	(27.75)	17.00	(10.98)	92.30	(23.49)
Tone 4	18.42	(9.94)	89.4	(25.72)	16.32	(9.07)	85.62	(24.18)

The results were verified by examining the data without non-words. In Figures 2 and 3, it is noted that the values for the data without non-words are shorter than the values for the data with non-words. It additionally shows that unaspirated stops in Tone 1 have the longest mean VOT values and unaspirated stops in Tone 2 have the shortest. Aspirated stops in Tone 3 have the longest mean VOTs, while those in Tone 4 have the shortest. ANOVA tests still indicate that lexical tone has significant influences on the VOT values for stops (F(3,692)=4.800, p<.01 for unaspirated stops; F(3,779)=2.953, p<.05 for aspirated stops). Furthermore, Post Hoc tests show that unaspirated stops in Tone 2 have significantly shorter mean VOTs than stops in Tone

1 and Tone 3 (p<.05), and aspirated stops in Tone 4 have significantly shorter mean VOTs than stops in Tone 3 (p<.05).

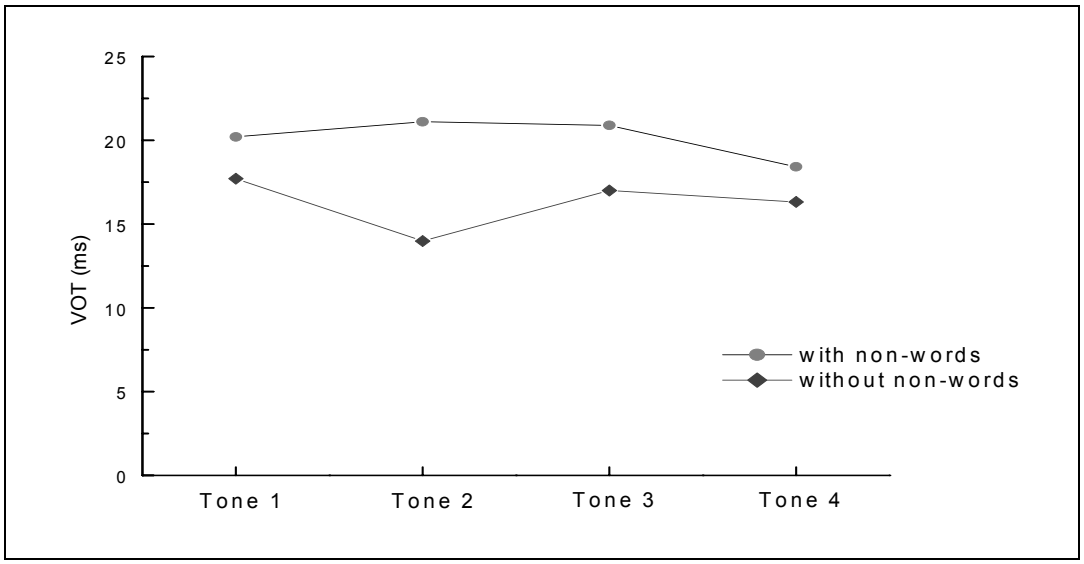


Figure 2. The mean VOTs for Mandarin unaspirated stops in individual lexical tones.

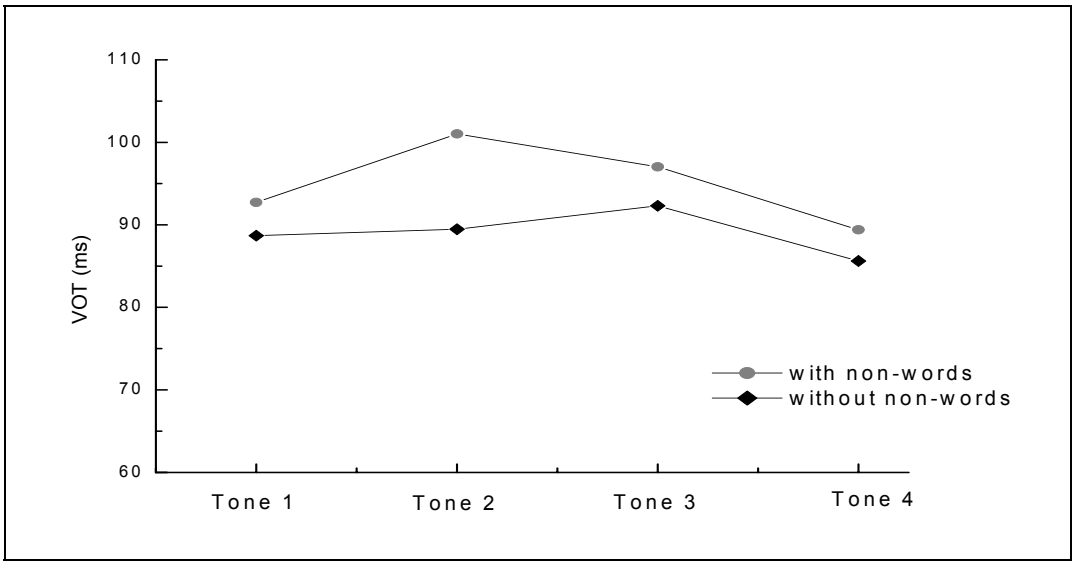


Figure 3. The mean VOTs for Mandarin aspirated stops in individual lexical tones.

3.2 Lexical tone and VOT in Hakka

The mean VOT values and standard deviations for Hakka stops in each lexical tone are shown in Table 3. ANOVA tests show that lexical tones have a significant influence on stop's VOTs ($F(5,943)=3.521$, p<.01 for unaspirated stops; $F(5,900)=37.365$, p<.001 for aspirated stops). In Figures 4 and 5, it is shown that unaspirated and aspirated stops in Tone 1 and Tone 5 have longer mean VOTs than stops in other tones. And the shortest mean VOTs for both unaspirated and aspirated stops are in Tone 8. Post hoc tests revealed that aspirated stops in Tone 4 and Tone 8 have significantly shorter mean VOTs than in Tone 1, Tone 2, Tone 3, and Tone 5 (p<.001).

Table 3. Hakka stops' mean VOT values in individual lexical tones. All measurements are in milliseconds (ms).

	Unaspirated stops		Aspirated stops	
	mean	(*SD*)	mean	(*SD*)
Tone 1	20	(11.56)	86.83	(25.8)
Tone 2	16.94	(8)	84.67	(26.56)
Tone 3	18.88	(11.02)	81.32	(23.73)
Tone 4	17.19	(9.44)	62.93	(18.36)
Tone 5	19.4	(11.43)	90.08	(27.08)
Tone 8	16.11	(7.98)	61.53	(20.36)

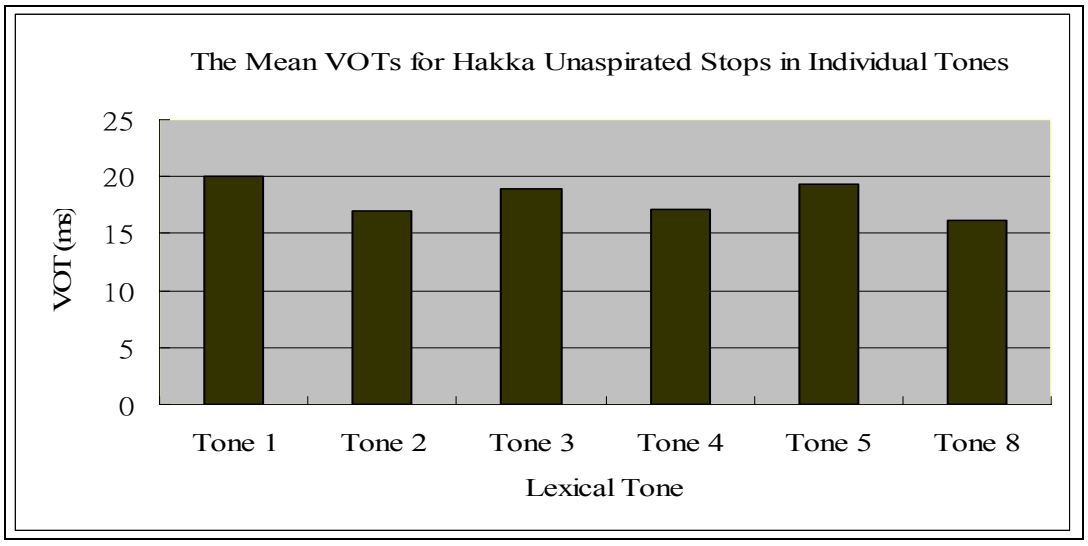


Figure 4. The mean VOTs for Hakka unaspirated stops in individual lexical tones.

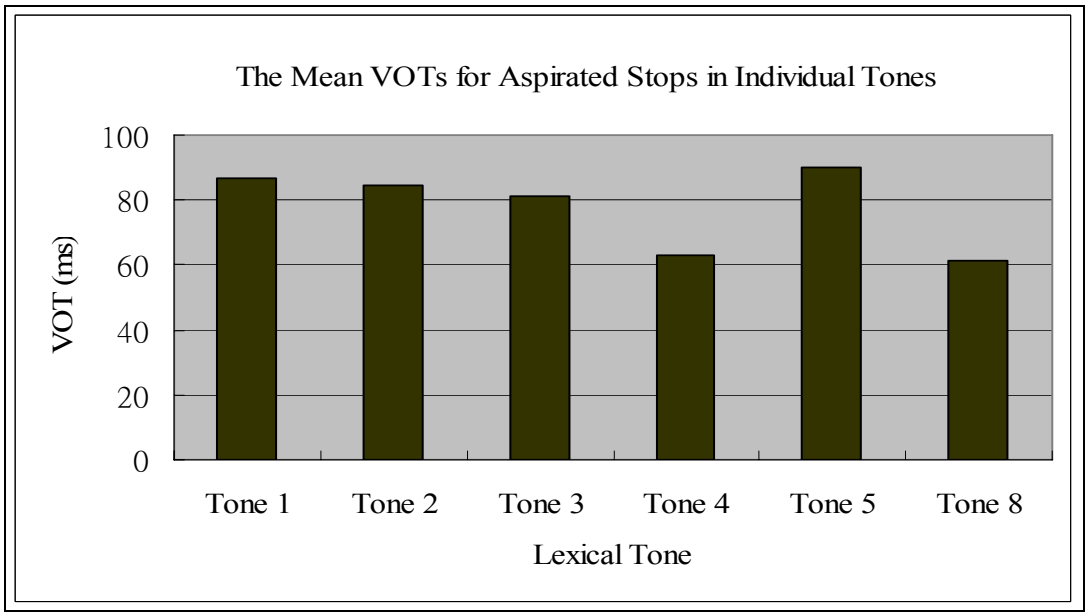


Figure 5. The mean VOTs for Hakka aspirated stops in individual lexical tones.

4. Discussion and Conclusion

In the current study, ANOVA tests reveal that lexical tone has a significant influence on the VOT values for Mandarin and Hakka stops. In Mandarin, the VOTs for both unaspirated and aspirated stops in Tone 2 have the longest mean VOT values, and in Tone 4 have the shortest mean VOT values. Stops' VOTs, ordering from the longest to the shortest, are in Tone 2, Tone 3, Tone 1, and Tone 4. This sequence is the same as Liu et al.'s (2008) [1] results. But it is worth noting that in both studies, some of the stimulus words are non-words. Later, it was found that the sequence results from the existence of non-words because in order to produce non-words correctly, participants tended to pronounce them at a lower speed, especially those in Tone 2. Therefore, we examined the data without non-words, in which no clear sequence had been found. In general, ANOVA tests revealed that lexical tones have significant influences on stops' VOTs. Moreover, Post hoc tests show that unaspirated stops in Tone 2 have significantly shorter mean VOTs than in Tones 1 and 3; while aspirated stops in Tone 4 have significantly shorter mean VOTs than in Tone 3. As for Hakka stops, the existence of non-words does not have a significant impact. Post hoc tests show that aspirated stops in Tones 4 and 8 have significantly shorter VOT values than stops in other tones. Hakka words in Tones 4 and 8 have similar phonetic characteristics, which are short, rapid and ended by a stop. This may explain why Hakka stops in Tones 4 and 8 are shorter than stops in other tones. The results in this study indicate that lexical tone has significant influence. Therefore, it is suggested that future studies should take the effects of lexical tone into consideration in creating the stimulus words of tonal languages when analyzing the VOT values for stops, in order to reduce the risk of introducing experimental errors. However, in what way tone will affect the VOT values for stops, further studies are needed.

References

[1] H. Liu, M. L. Ng, M. Wan, S. Wang, and Y. Zhang, "The effect of tonal changes on voice onset time in Mandarin esophageal speech," *Journal of Voice*, 2008, vol. 22, no. 2, pp. 210-218.

[2] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: acoustical measurements," *Word*, 1964, vol. 20, pp. 384-422.

[3] B. L. Rochet and Y. Fei, "Effect of consonant and vowel context on Mandarin Chinese VOT: production and perception," *Canadian Acoustics*, 1991, vol. 19, no. 4, pp. 105-106.

[4] T. Cho and P. Ladefoged, "Variation and universals in VOT: evidence from 18 languages," *Journal of Phonetics*, 1999, vol. 27, pp. 207–229.

[5] M. Gósy, "The VOT of the Hungarian voiceless plosives in words and in spontaneous Speech," *International Journal of Speech Technology*, 2001, vol. 4, pp. 75–85.

[6] X.-R. Zheng and Y.-H. Li, "A contrastive study of VOT of English and Korean stops," *Journal of Yanbian University*, 2005, vol. 38, no.4, pp. 99-102.

[7] T. J. Riney, N. Takagi, K. Ota, and Y. Uchida, "The intermediate degree of VOT in Japanese initial voiceless stops," *Journal of Phonetics*, 2007, vol. 35, pp. 439-443.

[8] L. Jäncke, "Variability and duration of voice onset time and phonation in stuttering and nonstuttering adults," *Journal of Fluency Disorders*, 1994, vol. 19, no. 1, pp. 21-37.

[9] P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: a review," *Clinical Linguistics & Phonetics*, 2000, vol. 14, no. 2, pp. 131-150.
[10] P. A. Keating, W. Linker, and M. Huffman, "Patterns in allophone distribution for voiced and voiceless stops," *Journal of Phonetics*, 1983, vol. 11, pp. 277–90.
[11] T. J. Riney and N. Takagi, "Global foreign accent and voice onset time among Japanese EFL speakers," *Language Learning*, 1999, vol, 49, no. 2, pp. 275–302.
[12] S. J. Liao, "Interlanguage production of English stop consonants: A VOT analysis," M. A. thesis, National Kaohsiung Normal University, Kaohsiung, Taiwan, 2005.
[13] B. S. Rosner, L. E. López- Bascuas, J. E. García-Albea, and R. P. Fahey, "Voice-onset times for Castilian Spanish initial stops," *Journal of Phonetics*, 2000, vol. 28, pp. 217-224.
[14] L.-M. Chen, K.-Y. Chao, and J.-F., Peng, "VOT productions of word-initial stops in Mandarin and English: a cross-language study," *Proceedings of the 19th Conference on Computational Linguistics and Speech Processing*, 2007, pp. 303-317.
[15] Y. R. Chao, Mandarin Primer, Cambridge: Harvard University Press. 1967.
[16] S.-S. He, "A contrastive study of Taiwan Hakka and Mandarin phoneme," In G.-S. Gu,（Ed.）. Introduction to Taiwan Hakka, pp. 163-192, Taipei: Wu-Nan Book Inc. 2005.
[17] C.-C. Cheng, A Synchronic Phonology of Mandarin Chinese, The Hague: Mouton. 1973.
[18] R.-F. Chung, An introduction to Taiwan Hakka phonology, Taipei: Wu-Nan Book Inc. 2004.
[19] P. Boersma and D. Weenink, Praat: doing phonetics by computer (Version 5.1.12), 2009. [Computer program]. Available: http://www.praat.org/.

Latent Prosody Model-Assisted Mandarin Accent Identification

Yuan-Fu Liao[1], Shuan-Chen Yeh[2], Ming-Feng Tsai[3],

Wei-Hsiung Ting[4], and Sen-Chia Chang[5]

[1,2,3,4]Department of Electronic Engineering, National Taipei University of Technology
[5]Advanced Technology Center, Information and Communications Research Laboratories,
Industrial Technology Research Institute
[1,2,3,4]yfliao@ntut.edu.tw, [5]chang@itri.org.tw

Abstract

A two-stage latent prosody model-language model (LPM-LM)-based approach is proposed to identify two Mandarin accent types spoken by native speakers in Mainland China and Taiwan. The frontend LPM tokenizes and jointly models the affections of speaker, tone and prosody state of an utterance. The backend LM takes the decoded prosody state sequences and builds n-grams to model the prosodic differences of the two accent types. Experimental results on a mixed TRSC and MAT database showed that fusion of the proposed LPM-LM with a SDC/GMM+PPR-LM+UPR-LM baseline system could further reduced the average accent identification error rate from 20.7% to 16.2%. Therefore, the proposed LPM-LM method is a promising approach.

Keywords: Accent recognition, latent prosody model, Mandarin, Taiwan

1. Introduction

Over the past decades, many approaches have been proposed to deal with language identification (LID) tasks. They tried to capture the specific characteristics of different languages. These characteristics roughly fall into three categories: the phonetic repertoire, the phonotactics, and the prosody. The mainstream system (as shown in NIST language recognition evaluation (LRE) 2007) [1] is usually based on the fusion of multiple acoustic and phonotactic systems.

Although LID is extensively studied, less works have been done on accent identification (AID), especially for native speakers, such as American and Indian English, Mainland China and Taiwan Mandarin, Hindi and Urdu Hindustani and Caribbean and non-Caribbean Spanish. Comparing with LID task, AID of native speakers is more challenging because, (1) some linguistic knowledge, such as syllable structure, may be of little use since native speakers seldom make such mistakes; (2) difference among those speakers is relatively smaller than

that among foreign (non-native) speakers. In other words, the capacities of the popular acoustic and phonotactic approaches may be limited in this case.

Many approaches have been proposed to model the prosodic differences between languages, dialects or accents [2], recently. Most of them are based on direct modeling of surface prosodic features, i.e., the raw prosodic features. For example, frame-level pitch flux features and GMMs were proposed in [3]; segmental-level pitch features were extracted using Legendre polynomials and modeled by ergodic Markov model in [4]; and supra-segment-level prosodic features were captured by n-gram in [5].

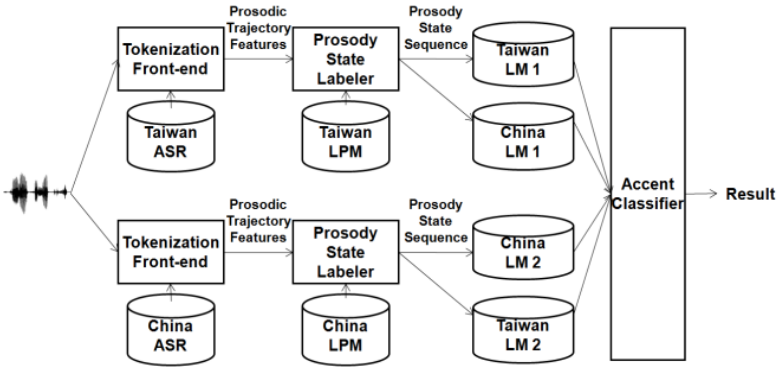


Figure 1. The block diagram of the proposed LPM-LM-based Mandarin accent identification system.

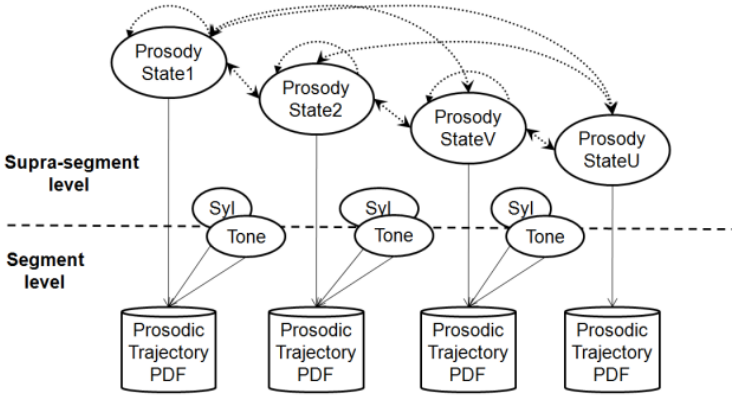


Figure 2. The block diagram of the proposed LPM framework (speaker factor is omitted to simply this figure).

However, surface prosodic features are often affected by many other non-prosodic latent factors, such as channel, speaker, phonetic context, and so on. Therefore, it is necessary to apply some feature normalization methods [6] to alleviate the unwanted affections. To absorb those unwanted affections, in this study a two-stage latent prosody model-language model (LPM-LM)-based approach as shown in Fig. 1 and 2 is proposed. The aim is to discriminate two Mandarin accent types spoken by native speakers in Mainland China and Taiwan.

In this approach, the frontend LPM [7] tokenizes (with the help of automatic speech recognizers (ASRs)) an input utterance into smaller prosodic units (sub-syllable in our case) and artificially introduces latent prosody states to represent the prosodic status of each token in an utterance. It then jointly models the affections of speaker, tone and prosody state on surface prosodic features in order to decode more precise prosody state sequences of the utterance. The backend LM then takes the decoded prosody state sequences and builds an n-gram to model the supra-segmental prosodic charactistics of each accent type.

In more detail, LPM as shown in Fig. 2 (1) introduces a two-level hierarchical structure of speech prosody [8] with prosodic states and state transition probabilities and (2) describes the joint affections of latent factors in a state by a variable-parameter probability density function whose parameters varies as a function of those latent factor-dependent parameters. The purpose is to explain the variant due to speaker, phonetic context and, especially, tone factors.

It is worth noting that (1) the proposed LPM-LM framework is similar to the popular parallel phone recognizer (PPR)-LM approach. However, the phone recognizers are replaced by automatic prosodic state tokenizers/labelers and, especially, (2) the LPM module could be trained in an unsupervised way to avoid any human annotation efforts.

This paper is organized as follows. Section 2 reviews the LPM framework. Section 3 discusses the application of LPM-LM on Mandarin AID. Section 4 reports the experimental results on a Mainland China and Taiwan Mandarin corpus. Some conclusions are given in the last section.

2. Latent Prosody Model of Speech Prosody

Based on the proposed LPM framework shown in Fig. 2, an input training utterance is first tokenized into a sequence of smaller prosodic units (sub-syllable in this case) including voiced and unvoiced segments. For each token, a segment-level prosodic feature vector $\mathbf{x}_n$ is extracted (coefficients of log-pitch and log-energy trajectories and the duration of the segment). Here, the coefficients of trajectories are computed using Legendre polynomial function from the raw log-pitch and log-energy contours. The speech prosody of an input utterance is thus represented by a sequence of segment-level prosodic feature vectors, i.e., $\mathbf{X}=\{\mathbf{x}_n, n=1,...,N\}$.

To well explain the variant of the observed prosodic feature vector sequence $\mathbf{X}$ of the utterance, several latent factors are introduced including speaker s, tone $\mathbf{T}=\{t_n, n=1,...,N\}$ (or major/minor stress in toneless language) and prosody state sequence $\mathbf{Q}=\{q_n, n=1,...,N\}$ (phonetic context is ignored in this study). The probability of $\mathbf{X}$ is defined as follows:

$$p(\mathbf{X})=\sum_{s,\mathbf{Q},\mathbf{T}} p(\mathbf{X}|s,\mathbf{T},\mathbf{Q})\,p(s,\mathbf{T},\mathbf{Q}) \tag{1}$$

Assume that each observed $\mathbf{x}_n$ is dependent only on local prosodic state q_n and tone t_n (and the speaker s), the first term in the right hand side of Eq. (1) is approximated as follows:

$$p(\mathbf{X}|s,\mathbf{T},\mathbf{Q})=\prod_{n=1}^{N} p(\mathbf{x}_n|s,t_n,q_n) \tag{2}$$

Assume that speaker, prosodic state and tone sequences are all independent variables and the probabilities of speaker s and tone sequence $\mathbf{T}$ are uniform distributions, the last term in the right hand side of Eq. (1) is approximated as follows:

$$p(s,\mathbf{T},\mathbf{Q}) \propto p(q_1)\prod_{n=2}^{N} p(q_n|q_{n-1}) \tag{3}$$

Finally, the distribution of the surface prosodic feature vector $\mathbf{x}_n$ is modeled by the following linearly additive [9] formulation:

$$\mathbf{x}_n = \mathbf{y}_n + \boldsymbol{\mu}_s + \boldsymbol{\mu}_{t_n} + \boldsymbol{\mu}_{q_n} \tag{4}$$

where $\mathbf{y}_n$ are prosodic feature vectors representing the normalized prosodic contours of the n-th syllable in an utterance; $\boldsymbol{\mu}_s$, $\boldsymbol{\mu}_{t_n}$ and $\boldsymbol{\mu}_{q_n}$ are the contributions of speaker s, prosody state q_n and tone t_n, respectively. The normalized pitch contour $\mathbf{y}_n$ is approximated using a zero mean Gaussian distribution $N(\mathbf{y}_n;\mathbf{0},\boldsymbol{\Sigma})$ (where $\boldsymbol{\Sigma}$ is diagonal matrix), or equivalently the observed prosodic feature vector $\mathbf{x}_n$ is modeled by

$$p(\mathbf{x}_n|s,t_n,q_n)=\mathbb{N}\left(\mathbf{x}_n;\boldsymbol{\mu}_s+\boldsymbol{\mu}_{t_n}+\boldsymbol{\mu}_{q_n},\boldsymbol{\Sigma}\right) \tag{5}$$

By this way, the likelihood function of an utterance given an LPM λ is expressed by

$$L(\mathbf{X}|\lambda)=\prod_{n=1}^{N} p(\mathbf{x}_n|s,t_n,q_n)\cdot p(q_1)\prod_{n=2}^{N} p(q_n|q_{n-1}) \tag{6}$$

Moreover, the optimal prosody state sequence $\hat{\mathbf{Q}}$ of an utterance could be automatically labeled using a Viterbi search algorithm (with or without tone tags given) which maximize the likelihood function $L(\mathbf{X}|\lambda)$, i.e.,

$$\hat{\mathbf{Q}}=\arg\max_{\mathbf{Q}} \log\left\{\prod_{n=1}^{N} p(\mathbf{x}_n|s,t_n,q_n)\cdot p(q_1)\prod_{n=2}^{N} p(q_n|q_{n-1})\right\} \tag{7}$$

3. LPM-based Mandarin Accent Identification

Mandarin spoken in Taiwan exhibits several major prosody differences from the Mandarin spoken in Mainland China [10]. Especially, people from Taiwan usually speak slower with a lower voice, and they sound soft and gentle; while Mainlanders have more ups and downs in their intonation, and their voices are higher and faster. These characteristics are likely

attributable, at least in part, to influence from the Southern Fujianese dialect widely spoken throughout Taiwan.

Since there are prosodic differences between Mainlander's and Taiwanese Mandarin, a LPM-based accent identification approach is built to identify these two Mandarin accent types. In the following subsections, the tokenization front-end and the speaker normalization parts of the proposed LPM-based approach and its training procedure are described in detail.

3.1. Tokenization front-end

The operation of the tokenization front-end is shown in Fig. 3. It firstly extracts the raw prosodic contours (log-pitch and log-energy) of an input utterance. The pitch and energy contours are then segmented by an ASR engine. The output is a sequence of voiced and unvoiced segments.

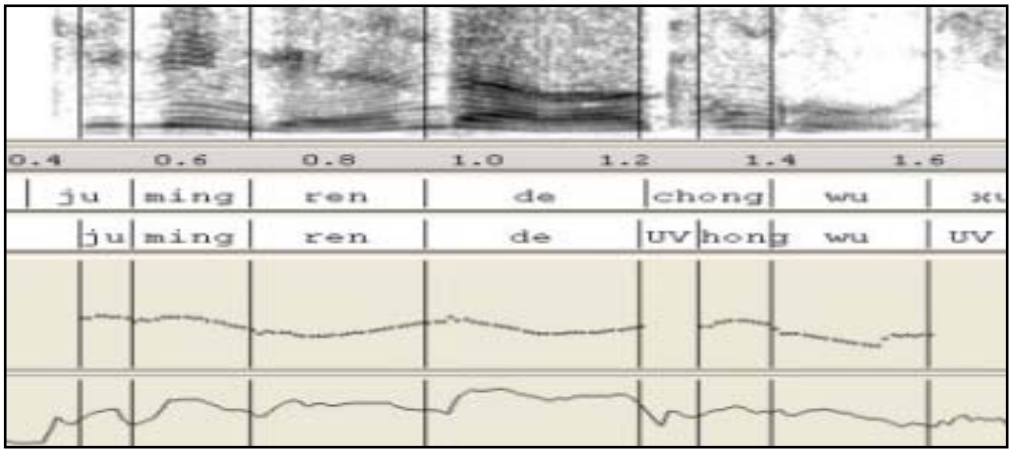


Figure 3. A typical segmentation results of the tokenization front-end (from top to bottom panel: spectrum, syllable and sub-syllable segmentations, log-pitch and log-energy contours).

For each voiced segment, six dimensional prosodic features are extracted including coefficients of 3-order Legendre polynomial function for approximating the log-pitch contour, the log-energy mean and duration of the segment. On the other hand, for each unvoiced segment, only its log-energy mean and duration are utilized.

3.2. LPM training algorithm

To estimate the parameters of the LPM, an unsupervised sequential optimization procedure based on the maximum likelihood criterion is adopted. The training procedure sequentially decodes latent prosody state sequences using Eq. (7) and updates the affecting factors (i.e., tone and prosody state) to optimize the likelihood function in Eq. (6).

In more detail, the sequential optimization training procedure executes the following steps until a convergence has been reached. It is worth noting that each step updates a subset of LPM parameters.

Step 0: Initialization

- Derive the initial affecting factors $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_{t_n}$ of tones by averaging all prosodic feature vector $\mathbf{x}_n$ of a speaker or the whole training data, respectively.
- Cluster and label the prosody state of each segment by vector quantization (VQ) using the residue prosodic feature vector $\mathbf{x}'_n = \mathbf{x}_n - \boldsymbol{\mu}_s - \boldsymbol{\mu}_{t_n}$ and derive the initial prosody state affecting factors $\boldsymbol{\mu}_{q_n}$.
- Derive the initial covariance matrix $\boldsymbol{\Sigma}$.
- Derive the initial prosody state transition probabilities using the statistics of labeled prosody states.

Step 1: Re-Label

- Re-label the prosody state sequence of all utterance using Eq. (7).

Step 2: Re-Estimate

- Update the affecting factors $\boldsymbol{\mu}_s$ of speakers, $\boldsymbol{\mu}_{t_n}$ of tones or $\boldsymbol{\mu}_{q_n}$ of prosody states with all other parameters fixed.
- Update the covariance matrix $\boldsymbol{\Sigma}$ and the prosody state transition probabilities.

Step 3: Iteration

- Repeat step 1 to 2 until the likelihood function Eq. (6) is converged.

4. Experimental Results

4.1. Corpus

To evaluate the proposed LPM approach, two telephone speech corpora were mixed together, one is Mandarin across Taiwan (MAT) [11] released by Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taiwan, and the other is 500-people telephone reading speech corpus (TRSC) [12] released by Chinese Corpus Consortium (CCC), China. There are about 4500 (MAT-2000+MAT-2500) Taiwanese and 500 Mainlander speakers in MAT and TRSC, respectively. The mixed corpus is randomly divided into a training, a development and a test set. The detail of speaker and utterance information is listed in Table. 1. The evaluation is executed utterance by utterance and the average length of an utterance is about 5 seconds.

Table 1. Detail information of the MAT ad TRSC corpora including number of speakers and utterances.

	Training		Development		Test	
	spk	utt	spk	utt	spk	utt
MAT	3936	67633	3742	20192	238	2009
TRSC	409	43340	120	12594	20	2042

4.2. LPM training results

For all following LPM experiments, the number of prosody states was empirically set to 11 (8 for voiced, 3 for unvoiced states) and there are 5 different tones in Mandarin.

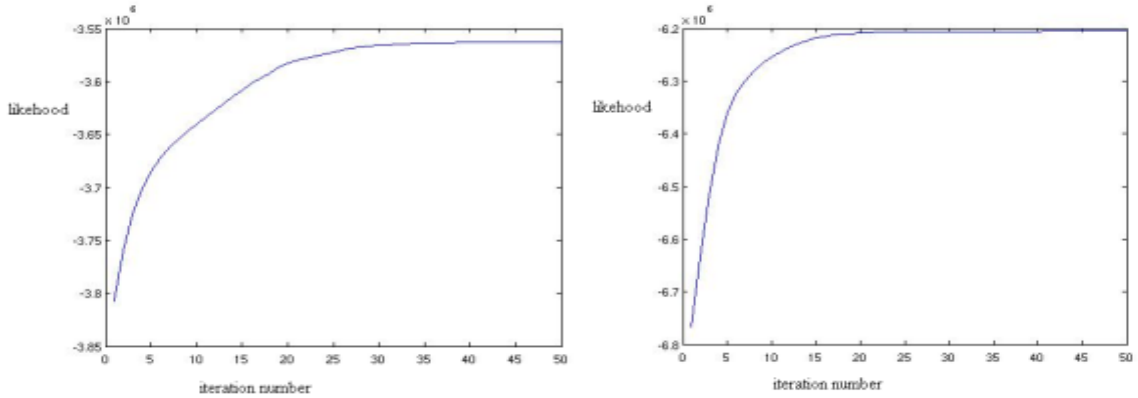


Figure 4. The learning curves of the LPMs training on MAT and TRSC training sets (left: MAT, right: TRSC), respectively.

First of the all, the learning curves of the LPMs were examined. Fig. 4 shows the likelihood functions on the MAT and TRSC training sets, respectively, along with the number of training iterations. It could be found from the figure that LPMs converged quickly, especially for the TRSC set.

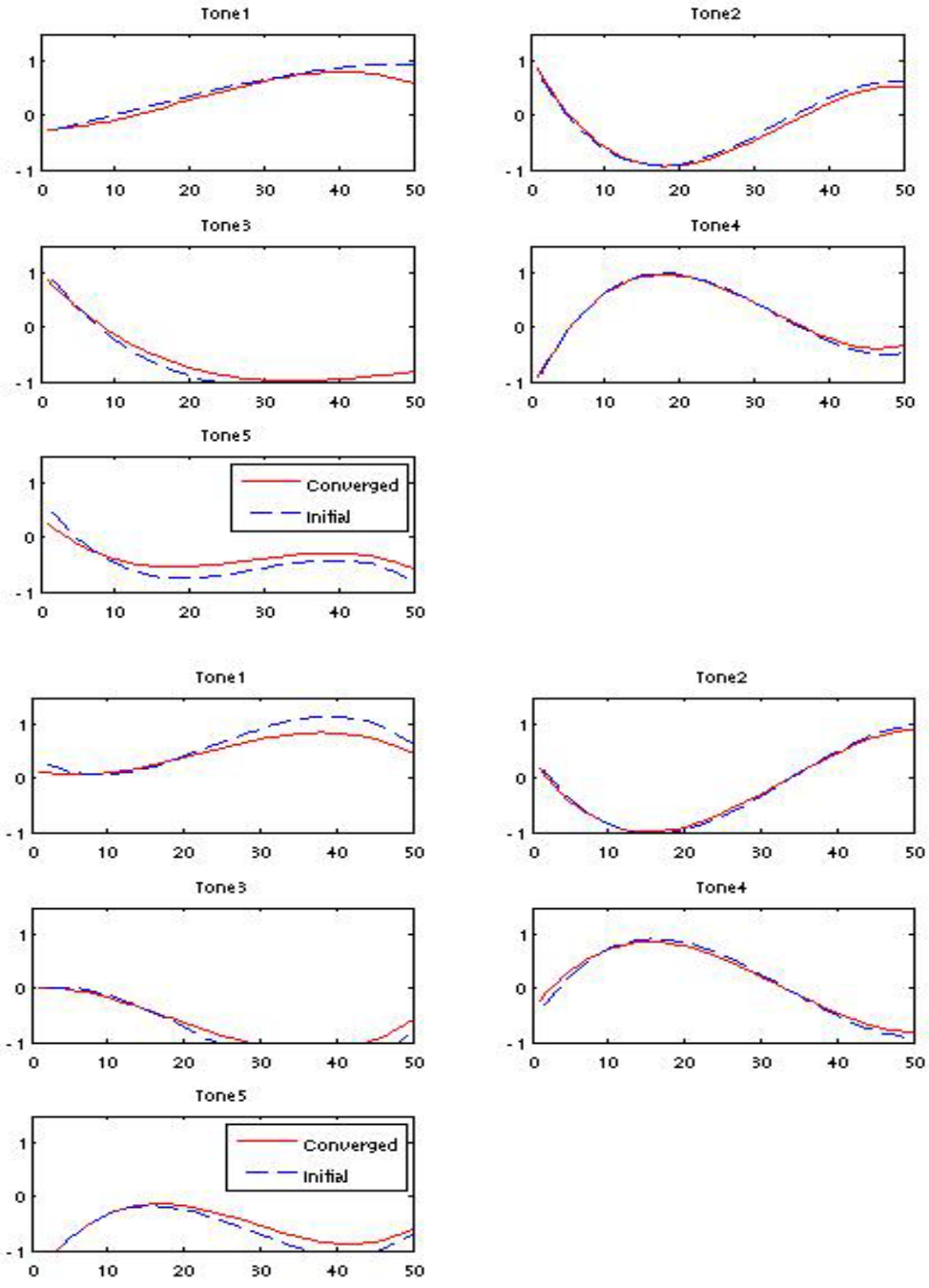


Figure 5. The learned tone affecting patterns on MAT and TRSC corpora (top 5 panels: MAT, bottom 5 panels: TRSC), respectively.

After LPM training was converged, the learned 5 tone affecting patterns of Taiwanese and Mainlanders' Mandarin, respectively, were drawn in Fig. 5. It is found that the major tone differences between Taiwan and Mainland China is the pattern of tone 3 and 5. This is consistent with common linguistic knowledge [10].

These results suggest that LPMs could automatically learn the accent-specific characteristics of Taiwanese and Mainlanders' Mandarin. We therefore expect that LPM-LM-based approach could be successfully used to discriminate these two Mandarin accents.

4.3. Acoustic and Phonotactic baselines

To set up a reference baseline, two popular phonotactic and one acoustic approaches were first tested including (1) PPR-LM, (2) universal phone recognizer (UPR)-LM and (3) shifted delta cepstral (SDC)/Gaussian mixture model (GMM).

For PPR-LM and UPR-LM, 39-dimensional mel-frequency cesptrum coefficient (MFCC) feature vectors were utilized to train the front-end phone recognizers. There are in total 50 phonemes in Mandarin for PPR-LM. But for UPR-LM, the number of phonemes is extended to 63 to reflect the major pronunciation differences (retroflex and nasal-endings sounds) between Mainlander's and Taiwanese Mandarin. All MFCCs were pre-processed by cepstral normalization (CN) to partially compensate the channel and database mismatch. Beside, tri-gram LM backbends were adopted for both PPR-LM and UPR-LM. Moreover, the parameters of SDC were empirically set to 7-3-3-7 and the number of mixtures in GMMs was 512.

Table 2. Experimental results of the individual acoustic, phonotactic and prosodic approaches and their fusion on a mixed TRSC and MAT database.

Approach	Error (%)	System Fusion	Error (%)
(1): PPR-LM	24.88	(5): (1)+(2)	21.84
(2): UPR-LM	23.79	(6): (1)+(3)	22.53
(3): SDC-GMM	29.11	(7): (1)+(2)+(3)	**20.68**
(4): LPM-LM	31.34	(8): (7)+(4)	**16.18**

Table 2 shows the performances of the individual systems and their fusion results. The fusion was done using a softmax-output multi-layer perceptual (MLP) and trained with the development sets. From Table 2, it is found that (1) PPRLM and UPRLM worked better than SDC/GMM and (2) the best performance, 20.68% error rate, was achieved by the fusion of the PPR-LM, UPR-LM and SDC/GMM systems.

4.4. Prosodic approach

The proposed LPM-LM approach was then evaluated. In training phase, the correct tone tags were given but in testing phase MLP-based tone recognizers are adopted to provide estimated tone tags online [7].

Table 2 shows the performances of the proposed LPM-LM and the fusion of LPM-LM with the acoustic and phonotactic baseline. The fusion was also done using the same softmax-output MLP and trained with the development sets. Different from acoustic feature, the prosodic feature extracts another characteristic (example: tone). From Table 2, it is found that LPM-LM worked compatible with the SDC/GMM but is worse than the acoustic and phonotactic baseline. It was caused by just using prosodic feature rather than strong acoustic feature. However, the fusion of LPM-LM and the acoustic and phonotactic baseline could further reduce the error rate from 20.68% to 16.18%. This result may suggest the complementary of those methods.

5. Conclusions

In this paper, a LPM-LM-based approach is proposed to identify two Mandarin accent types spoken by native speakers in Mainland China and Taiwan. Experimental results on a mixed TRSC and MAT database showed that fusion of the proposed LPM-LM and a SDC/GMM+PPR-LM+UPR-LM baseline system could further reduced the average accent identification error rate from 20.7% to 16.2%. Therefore, the proposed LPM method is a promising approach.

6. Acknowledgement

This work was supported by the National Science Council, Taiwan, under the project with contract NSC 96-2221-E-027-100-MY2 and is a partial result of Project 8353C41220 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

References

[1] Language Recognition Evaluation, National Institute of Standards and Technology, http://www.itl.nist.gov/iad/mig/tests/lre/.

[2] Jean-Luc Rouas, "Automatic Prosodic Variations Modeling for Language and Dialect Discrimination," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 1904-1911, Aug. 2007.

[3] Bin Ma, Donglai Zhu, and Rong Tong, "Chinese Dialect Identification Using Tone Features Based on Pitch Flux," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Toulouse, France, May 2006, pp. I-I.

[4] Chi-Yueh Lin and Hsiao-Chuan Wang, "Language Identification Using Pitch Contour Information in the Ergodic Markov Model," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Toulouse, France, May 2006, pp. I-I.

[5] Obuchi, Y. and Sato, N, "Language Identification Using Phonetic and Prosodic HMMs with Feature Normalization," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, Philadelphia, Mar. 2005, pp. 569-572.

[6] Najim Dehak, Pierre Dumouchel, and Patrick Kenny, "Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, no. 17, pp. 2095-2103, Sept. 2007.

[7] Chen-Yu Chiang, Xiao-Dong Wang, Yuan-Fu Liao, Yih-Ru Wang, Sin-Horng Chen, and Keikichi Hirose , "Latent Prosody Model of Continuous Mandarin Speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Hawaii, Apr. 2007, pp. IV-625-IV-628.

[8] Chiu-yu Tseng, Shao-huang Pin, Yehlin Lee, Hsin-min Wang, and Yong-cheng Chen, "Fluent speech prosody: Framework and modeling," *Speech Comminication*, vol. 46:3-4, pp. 284-309, Mar. 2005.

[9] Sin-Horng Chen, Wen-Hsing Lai, and Yih-Ru Wang, " A statistics-based pitch contour model for Mandarin speech," *Journal of the Acoustical Society of America*, 117 (2), pp. 908-925, Feb. 2005.

[10] Chin-Chin Tseng, "Prosodic Properties of Intonation in Two Major Varieties of Mandarin Chinese: Mainland China vs. Taiwan," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing, China, Mar. 2004, pp. 28-31.

[11] Hsiao-Chuan Wang, Frank Seide, Chiu-Yu Tseng, Lin-Shan Lee, "MAT-2000 - Design, Collection, and Validation of a Mandarin 2000-Speaker Telephone Speech Database", in *ICSLP 2000*, Beijing, China, Oct. 2000, pp. 460-463.

[12] 500-People TRSC (Telephone Read Speech, Corpus), Chinese Corpus Consortium, China, http://www.d-ear.com/CCC/corpora/2003-TRSC.pdf, 2003.

高解析度之國語類音素單元端點自動標示

Sample-based Phone-like Unit Automatic Labeling in Mandarin Speech

林宥余　You-Yu Lin
國立交通大學電信工程研究所
Institute of Communication Engineering, National Chiao Tung University
rossi0927.cm97g@g2.nctu.edu.tw
王逸如　Yih-Ru Wang
國立交通大學電信工程研究所
Institute of Communication Engineering, National Chiao Tung University
yrwang@mail.nctu.edu.tw

摘要

在本論文中提出一種以取樣點爲單位(sample-based)的高時間解析度之音素端點自動標示與切割的方法，有別於傳統分析語音信號以音框爲單位(frame-based)或是音段爲單位(segment-based)的研究。本文中，我們提出了一些以取樣點爲單位的聲學參數；由實驗結果顯示，這些聲學參數在不同發音特徵之音素轉換間有明顯的變化率，有利於音素切割位置之標記。我們利用這些發音特徵變化的聲學參數特性，建立一個高時間解析度的自動音素端點標示與切割系統。由TCC-300國語語料庫進行自動端點標示之實驗結果顯示，本論文所提出的方法比傳統以音框爲單位之切割方法，亦即HMM之切割方法，更能有效切出精準的短停頓、摩擦音、塞擦音等之音素端點位置。

Abstract

This paper presents a sample-based phone boundary detection algorithm which can improve the accuracy of phone boundary labeling in speech signal. In the conventional phone labeling method adopted the frame-based approach, some acoustic features, like MFCCs, are used. And, the statistical approaches are employed to find the phone boundary based on these frame-based features. The HMM-based forced alignment method is most frequently used method. The main drawback of the frame-based approach lies in incapability of modeling rapid changes in speech signal; moreover, the time resolution of this approach is too coarse for some applications. To overcome this problem, a sample-wise phone boundary detection framework is proposed in this study. First, some sample-wise acoustic features are proposed which can properly model the variation of speech signal. The simple-based spectral KL distance is first employed for boundary candidates pre-selection in order to reduce the complexity of sample-based methods. Then, a supervised neural network is trained for phone boundary detection. Finally, the effectiveness of the proposed framework has been validated on automatic labeling of TCC-300 speech corpus.

關鍵詞：音素端點切割，帶通信號波封，sample-based 頻譜 KL 距離，監督式類神經網路

Keywords: phone boundary segmentation, sub-band signal envelope, sample-based spectral KL distance, supervised neural network

一、緒論

正確音素切割位置在語音辨認的研究中可以提升辨識模型的可靠度與統計上一致性進而提升辨識率，也扮演著語音合成方面合成聲音品質提升的重要因素之一。在全球有人工切割位置的語料庫不多，最著名的是 TIMIT 語料庫，但是一個大型的連續語音資料庫，使用人工標記切割位置的方式，不僅非常耗時且人工切割的標記位置也伴隨著一個缺點，就是以人工做標記的動作時，會因爲主觀上認定切割位置不同而使得標記的位置缺乏一致性，因此一個能夠自動標記且具有精確切割位置的語料庫是非常重要的。

在語音信號處理中，自動音素之切割是一個非常重要的問題，儘管在過去有非常多自動音素切割的研究[1]，一個具有高精準度的自動音素切割演算法，仍是一個可待持續研究的課題。在過去一些自動音素切割與偵測的研究中，主要可分爲 Model-based 及 Metric-based 或是上述兩種方法結合。

在 Model-based 方法中，最常被使用的就是以概似法則訓練的隱藏式馬可夫模型(maximum likelihood-trained Hidden Markov Model, ML-trained HMM)做自動語音切割，其效能可在正負 20 ms 之內佔有 90%的比率(inclusion rate)，而傳統 HMM 是以整段語句所得到最大相似度函數(maximum likelihood, ML)爲訓練準則，故其自動切割之位置並非爲最佳之音節或音素邊界位置。近年來有學者提出一些方法，其中以最小邊界錯誤(minimum boundary error, MBE)爲訓練準則之 HMM[2]，就使用自動給定之已知端點間誤差最小化作爲 HMM 模型之訓練準則，在 TIMIT 語料庫中，MBE-HMM 自動切割之邊界與人工切割邊界誤差範圍 10 ms 之內的比率高達 79.75%，與傳統 ML-trained HMM 模型其百分比 71.23%相比，提昇許多；然而其自動切割位置只有 7.89%的邊界在人工切割位置誤差 20 ms 之外。此外，也可使用其它圖形識別的方法如支撐向量機(support vector machine, SVM)[3]、類神經網路(neural network, NN)[4]，來對 HMM 之自動切割位置再作進一步地修正，以獲得更好的結果。

而在 Metric-based 方法中，我們知道語音信號在一個音素中穩定的信號，其聲學參數變化的速率就是決定一個音素邊界的重要線索，回顧一些文獻如 Rabiner[5]使用頻譜轉換量測(spectral transition measure)的音素邊界偵測方法，應用在 TIMIT 語料庫其效能可達到在誤差 20ms 的容忍範圍內，只有 23.1%的音素端點位置沒偵測出來 (missed detection rate, MD)、22.0% 誤報率(false alarm rate, FA)。Kotropoulos[6]結合 Kullback-Leibler(KL)距離及貝式資訊法則(Bayesian Information Criterion, BIC)所提出的 DISTBIC 演算法來偵測語音信號之音素邊界，其效能在 NTIMIT 語料庫亦可達到 25.7% MD 與 23.3% FA 的結果。

在先前的音素切割方法中，無論 model-based 或 metric-based 的方法中，常用的語

音信號參數多與信號頻譜相關；且一般假設語音信號在短時間內爲穩定的特性，故使用 frame-based 的聲學參數，例如梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)。然而，在做頻譜分析時會造成時間與頻譜(time-spectrum)上之不確定性(uncertain)，所以頻譜參數越精確就會犧牲時間精確度；但在 frame-based 架構中必須要讓頻譜解析度越精細，以提昇辨認音素能力，而發音器官變化很快的音素如爆破音，其音長可能小於一個音框，使得 frame-based 方法之切割位置與實際正確音素邊界位置之間產生誤差，因此對於自動語音切割之研究提昇時間解析度，必可降低大量因音框之時間解析度所造成的誤差。而語言學家就曾經提出一些用來區別發音特徵的參數，一般稱之爲 Articulation Parameter (AP)。其方法可用低解析度的頻帶，來區分像發音方式或發音位置以及偵測一些 landmark 如 voice on-set，而不是用來辨認像音素的精細分類。由以上敘述，在自動語音切割的應用，我們可以思考爲了使得自動端點標示的時間精確度能夠提昇，降低頻譜精確度的可行性。故在本文中，我們提出 sample-based 音素端點偵測方法的架構，並與 frame-based HMM 切割位置做比較。

在本文中其它章節概要如下：在第二節中，我們首先說明 sample-based 音素端點偵測方法的整體架構；第三節對於本論文中所提出之一些 sample-based 聲學參數的特性做進一步地說明；第四節則是介紹利用上述 sample-based 聲學參數並使用多層感知器(multi-layer perception, MLP)類神經網路架構的 sample-based 音素端點偵測方法；第五節爲實驗結果探討，並於第六節提出簡單的結論。

二、系統架構

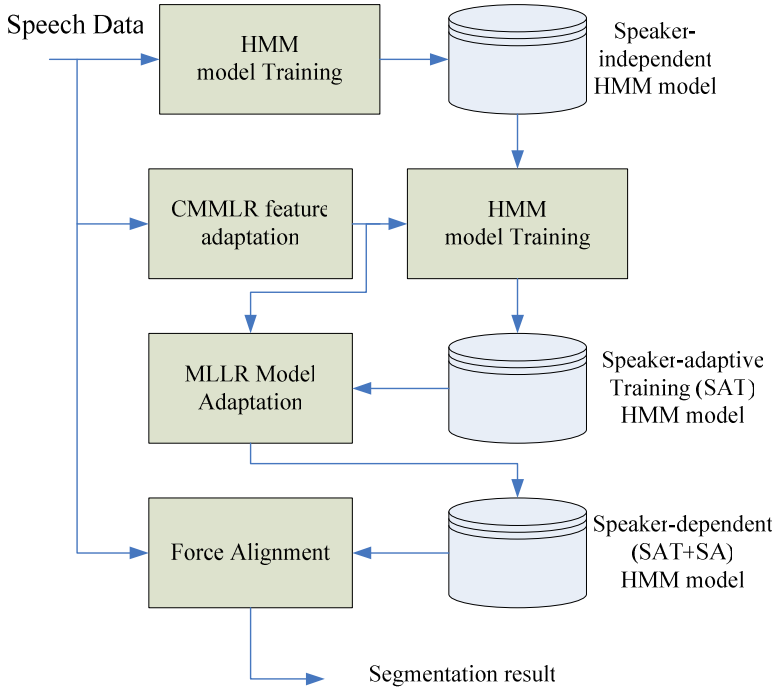
一般傳統切割的方法，主要分成兩個部份，首先利用統計模式爲基礎的方法，如 HMM-based forced alignment 當作初始切割位置，再藉由一些方法如 SVM 等，以進一步修正初始切割位置(refinement)。本研究是對 TCC-300 語料庫做切割，先使用 HMM-based forced alignment 得到初始切割位置；接著，利用 sample-based 聲學參數進一步調整該初始位置；並以 KL distance 挑選其候選端點，訓練一個 MLP 音素端點偵測器以得到最佳之切割位置。由於 TCC-300 語料庫是由不同的語者所組成，所以在取得 HMM-based forced alignment 初始切割位置時，我們使用了語者調適的技術調將 HMM 模型調適成更適合該語者之模型。接下來我們進一步介紹語者調適的流程以及 MLP 音素端點偵測器。

（一）、使用 SAT 及 SA 技術之 HMM phone-like unit alignment 流程

我們將使用下列流程做TCC-300語音資料庫HMM模型類音素層級(phone-like level)之起始切割位置，就是將一個音節區分爲聲母、介音、韻母及韻尾鼻音等部分，其方塊圖如下：

在 HMM phone model training 後，我們再使用做 speaker adaptation training(SAT)；SAT 就是使用 constraint MLLR(CMLLR)對不同語者做語音參數的轉換；使用經語者轉換(CMLLR)後之語音參數再重新訓練新的 HMM 模型將可獲得較佳之 speaker-dependent HMM 模型。做完 SAT 後，我們再做 HMM 做 model adaptation，使用 MLLR 技術來調

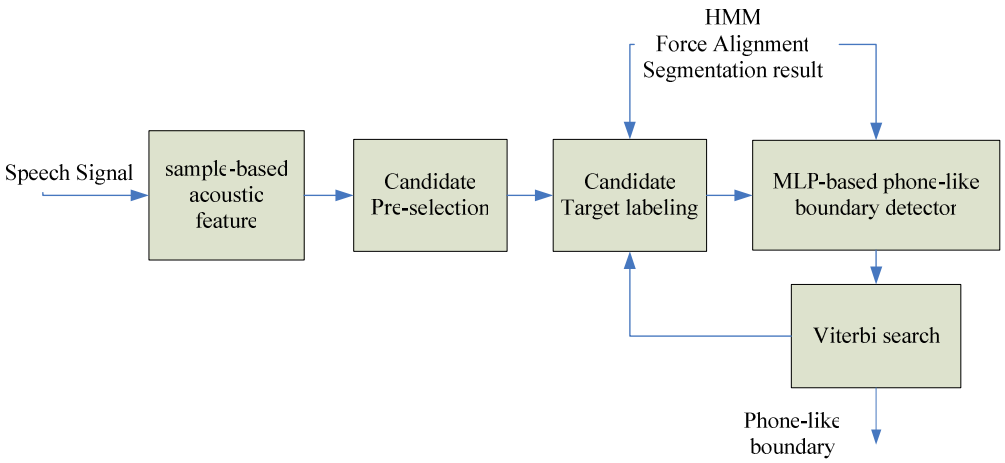
適 HMM 模型，它和 SAT 會有加成性的效果。如此就可以獲得較佳的 HMM 模型來做強制對齊(force alignment)，作爲 TCC-300 語料庫之音素的起始切割位置。



圖一、使用 SAT 及 SA 技術之 HMM　alignment 流程

（二）、MLP 音素端點偵測器訓練流程

MLP 類神經網路被廣泛地運用在各個領域當中作爲資料分類的架構，同時因爲其自我調適的能力、非線性的運算、具有學習能力等特性，故本研究使用此架構訓練一個監督式(supervised)MLP 音素端點偵測器。訓練音素端點偵測器流程，其方塊圖如圖二。



圖二、使用 MLP 架構之類音素端點偵測流程

在圖二中，因經由 HMM alignment 獲得之初始切割位置仍不夠準確，故我們利用 sample-based 的聲學參數所提供之資訊來得到較好的切割位置以作爲訓練 MLP 音素端點偵測器時之答案。且爲減少計算量，由預選擇(pre-selection)即簡單設定一個臨界值的方法來挑選較爲可能之候選端點(candidate)位置。接著將候選端點依目標函數(target function)分類後，訓練 MLP 音素端點偵測器直至收斂，最後使用 Viterbi search 演算法在候選端點中得到該語句最佳之切割位置。

三、sample-based 聲學參數的特性

首先，本研究結合語言學家所提出的 AP，利用數個頻段來區分不同發音特徵之方法，應用於切割語音信號可提高時間解析度由音框進一步精準至取樣點，並在此提出一些 sample-based 的 AP 以用於描述不同語音屬性變化時的 AP 特性，來調整音素切割位置之標記。在此節中將介紹本論文所提出之 sample-based 聲學參數及其在音素端點偵測上之特性。

（一）、Sample-based 聲學參數

我們提出一些 sample-based 的 AP 如帶通信號波封(sub-band signal envelope)、參數上升率(rate of rise, ROR)、頻譜熵(spectral entropy) 、sample-based spectral KL distance 及 spectral flatness，並觀察它們在不同語音屬性，如爆破音、鼻音、靜音等特性。以下，我們進一步介紹本研究所使用的語音特徵參數：

1、帶通信號波封[7]

在語言學家所提出的 AP 中，有許多帶通濾波器，它們各自能用來區別不同的發音方式或發音位置，常見的頻段(filter bank)[7]有以下：

0.0 – 0.4 KHz 0.8 – 1.5 KHz 1.2 – 2.0 KHz

2.0 – 3.5 KHz 3.5 – 5.0 KHz 5.0 – 8.0 KHz

例如在摩擦音、塞擦音中，在頻譜中之高頻段成份能量極強，低頻段成分能量較弱，鼻音韻尾則是在低頻段的成份能量極強。這些頻段中能量在有明顯變化的時候，可視爲是語音信號開始改變的地方。但語言學家所使用的 AP 爲帶通信號波封，而非現今語音辨認器中常用的能量。故我們將這六個頻段之語音信號取出它的波封來當作本研究中所使用的聲學參數。

我們在製作一個波封檢測器(envelope detector)時，爲了保持波封變化快的時候能正確地找到信號的波封，我們使用希爾伯特變換(Hilbert transform)後再經低通濾波器，求取輸入信號的波封，一個信號 $x[n]$ 的希爾伯特變換 $H(x[n])$ 的希爾伯特變換，如下式：

$$H(x[n]) = x[n] \otimes h[n] \quad \text{and} \quad h[n] = \begin{cases} 0, & n \text{ is even} \\ 1/n\pi, & n \text{ is odd} \end{cases} \tag{1}$$

2、上升率[7]

語言學家所稱之上升率，就是在 frame-based 的語音特徵參數中所用的 delta-term：

$$ROR_x[n] = \left(\sum_{i=-w}^{w} i \cdot x[n+i] \right) / \left(\sum_{i=-w}^{w} i^2 \right) \tag{2}$$

其中 $x[n+i]$為輸入參數資料，w 為求上升率所使用的音框寬度。本研究使用波封的上升率、頻譜熵之上升率、各頻段信號波封的上升率等當作語音信號的聲學參數，來描述各 sample-based 聲學參數的變化率。

3、頻譜熵 [9-10]

頻譜熵可用來描述信號在頻譜上的集中程度，若信號越集中在某一個頻段則頻譜熵越小。在此，本研究使用先前所述之 6 個頻段，則頻譜熵 H_s 可以定義如下式表示：

$$H_S = -\sum_i E_i[n]\log\left(E_i[n]\right) \tag{3}$$

$$E_i[n] = \frac{e_i}{\sum_{j=1}^{6} e_j} \tag{4}$$

其中 $E_i[n]$為第 i 個頻段之第 n 點正規化之後的波封。由頻譜熵對應到語音信號上，可以發現短停頓類似於雜訊，在各個頻段都會出現，所以頻譜熵值較高；而韻母在頻譜上的能量則較集中於低頻至中頻的部分，其頻譜熵值相對較低。

4、Sample-based spectral KL distance

將頻譜視爲一個機率分佈，因此可用 KL distance 來描述頻譜上的相似程度。在語音信號中計算兩點不同時間(m 與 n)的 spectral KL distance，$d_x(m,n)$，可以由下式表示：

$$d_x(m,n) = \sum_{i=1}^{6}\left(E_i[n] - E_i[m]\right)\log\left(\frac{E_i[n]}{E_i[m]}\right) \tag{5}$$

以上所敘述的參數頻譜熵、頻譜熵的上升率、sample-based KL distance 來觀察一段語音信號其語音特徵的變化，這些語音特徵證實可以分辨不同語音屬性的邊界。

5、Spectral flatness[11]

使用正規化後之帶通信號波封計算的 flatness，F，表示如下式：

$$F = \frac{\left(\prod_{i=1}^{6}\frac{E_i[n]}{s_i}\right)^{1/6}}{\frac{1}{6}\left(\sum_{i=1}^{6}\frac{E_i[n]}{s_i}\right)} \tag{6}$$

其中 S_i 爲第 i 個頻段靜音信號正規化後波封的平均。若信號爲靜音(silence) 或是短停頓(short pause)，則 F 將會趨近於 1。若 spectral flatness 與波封等參數經過設定適當的臨界值(thresholds)，對於標記靜音及短停頓的切割位置時是一個有效的參數。

（二）、Sample-based 聲學參數之語音特徵

在此我們將觀察 sample-based 聲學參數語音特徵之特性與類音素端點間之關係，並以實例證實先前我們所提出之 sample-based 聲學參數具有正確偵測類音素端點間端點的能力與特性。

以下使用 TCC-300 麥克風語料庫來做爲觀察 sample-based 聲學參數之語音特徵的

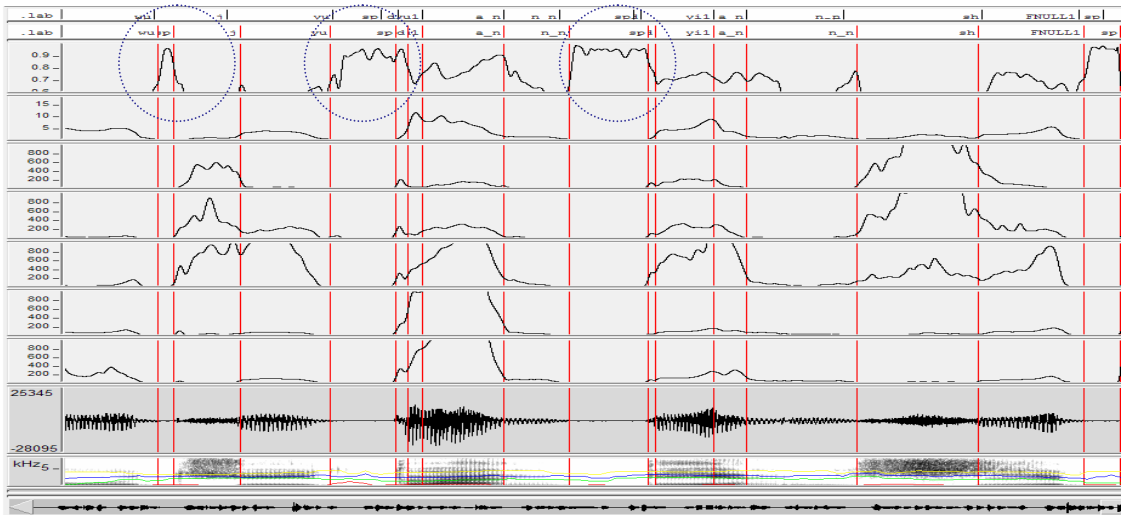
語料，在觀察中我們利用先前由 SAT 及 SA 技術之 HMM phone-like unit alignment 所獲得之切割位置作爲比較對象。

首先，利用 SAT(speaker adaptation transform, feature MLLR)及 SA(speaker adaptation, MLLR)後的語者調適 HMM 模型來得到 TCC-300 的類音素單元之切割位置，接著我們利用此新切割位置以語音屬性的不同做分類，如表一。由新切割位置當做參考位置利用 sample-based 的聲學參數特性觀察是否可用來調整音素端點的位置。由於先前語者調適 HMM 之切割位置，已近乎準確，但是仍有更進一步修正的空間，故我們提出以 sample-based 音素端點偵測的方式以期達到更爲精確的切割位置。

表 1、國語語音發音方法的分類表。

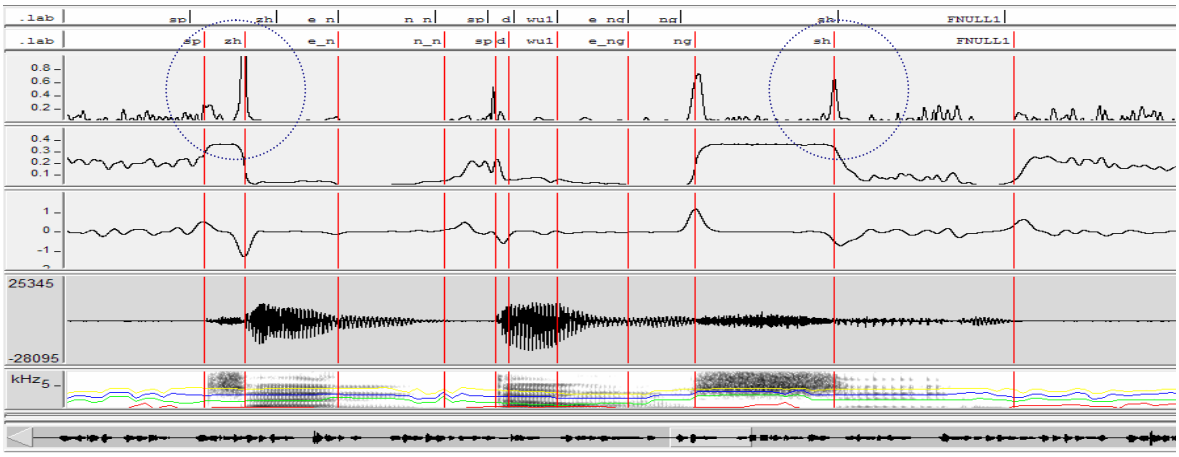
爆破音 Stop	b	p	d	t	g	k
鼻音 Nasal	m	n	(n_n)	(ng)		
摩擦音 Fricative	f	s	x	h	sh	
塞擦音 Affricate	q	j	c	z	zh	ch
流音 Liquid	l	r				
韻母音 Vowel	others					

先前觀察 HMM 自動切割位置的標記時，發現短停頓常無法標記出來，而使得塞擦音與爆破音等音素平均音長過長的現象。如圖一，在這裡我們使用 spectral flatness、波封以及各頻段之信號波封來判斷是否爲短停頓的狀態。在短停頓與爆破音及塞擦音的交界處，短停頓在各個頻段之信號波封與其它有語音信號的地方相比幾乎很低，且 flatness 趨近於 1，波封可與 flatness 產生互補的效果來標記短停頓的端點。



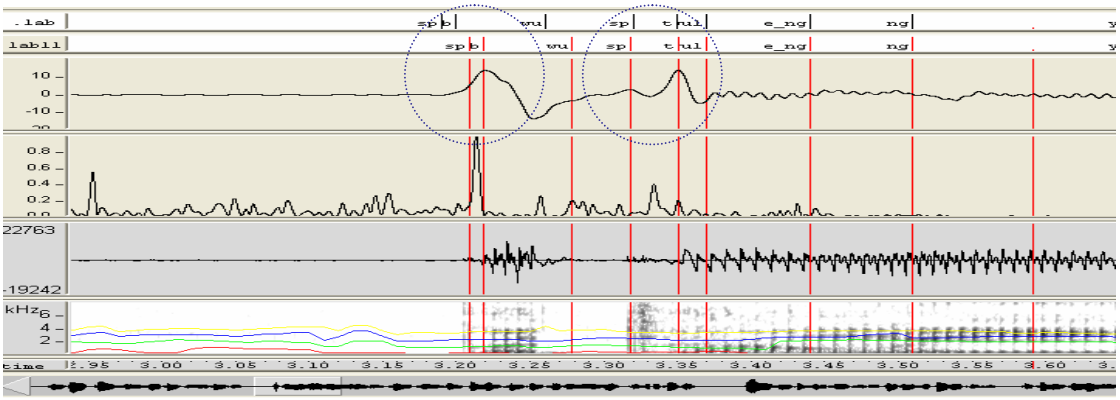
圖三、國語語句觀察短停頓切音位置之例子，由上至下的圖形分別表示 Spectral flatness、波形之波封、第 6 個至第 2 個頻段的信號波封、波形、頻譜。(最上方兩列標音位置分別表示是原語者調適 HMM 切割位置及修正後之切割位置)

接下來我們觀察摩擦音、塞擦音等聲母，它們在頻譜中與相鄰韻母與短停頓有極大的頻譜差異。在此我們使用 spectral KL distance、頻譜熵以及頻譜熵的上升率來偵測音素的端點。如圖四所示之圓圈圈選處中，我們可以看到與摩擦音及塞擦音相鄰頻譜之差異非常大，而 spectral KL distance 在摩擦音、塞擦音等聲母接續至韻母或是韻母轉換至摩擦音、塞擦音之情形有相對其他部分有較高的峰值。且摩擦音、塞擦音相鄰韻母的端點，頻譜熵值上升與下降速度很快，分別在頻譜熵的上升率中造成極大、極小的峰值。頻譜熵的上升率之峰值位置與我們所期望的正確端點位置差距不遠，我們可以了解頻譜熵、 KL distance 等已知在 frame-based 偵測信號變化量是非常有用之聲學參數，同樣在 sample-based 的效果一樣明顯，且標記之切割位置更精準。



圖四、國語語句觀察摩擦音、塞擦音切音位置之例子，由上至下的圖形分別表示 sample-based KL distance、頻譜熵、頻譜熵上升率、波形、頻譜。（最上方兩列標音位置分別表示是原語者調適 HMM 切割位置及修正後之切割位置）

爆破音切割位置的修正時，由波形與頻譜觀察中，我們可以發現通常在爆破音開始的時候會有短停頓出現，波封接著會有急遽上升的現象，故我們使用波封之上升率來描述其現象。如圖 6 所示，在爆破音結束的地方，也是音素轉換的端點。



圖五、國語語句觀察爆破音切音位置之例子，由上至下的圖形分別表示波封上升率、sample-based KL distance、波形、頻譜。（最上方兩列標音位置分別表示是原語者調適 HMM 切割位置及修正後之切割位置）

鼻音部分我們可發現信號在頻譜上多集中在0.0 – 0.4 KHz 與0.8 – 1.5 KHz 的低頻頻段，與相鄰的音素皆有頻譜上的差異，在此我們也使用 spectral KL distance 來判斷。而韻母端點的偵測，是利用相鄰聲母及短停頓之端點位置，當做韻母的邊界切割位置。

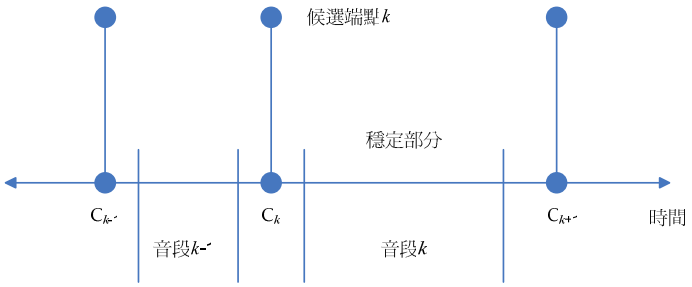
四、使用 MLP 類神經網路架構的 Sample-based 音素端點偵測方法

在前一節的觀察中已證實本論文中所提出之 sample-based 聲學參數有精確偵測類音素端點的能力。在這一節我們將使用這些參數來製作一個監督式(supervised)的 MLP 類神經網路模型來作爲音素端點之偵測器。

我們從語音信號中抽取 sample-based 聲學參數之後，爲了減少在端點偵測器中過於龐大的資料計算量，經由預選擇即簡單設定一個臨界值方法來挑選較爲可能之候選端點位置；而當語音信號在頻譜中的變化量大時，spectral KL distance 是一種很好的測量方式，故若 spectral KL distance 滿足下式：

$$d_x(n-1,n) < d_x(n,n+1),\ d_x(n,n+1) > d_x(n+1,n+2) \text{ and } d_x(n,n+1) > Th_d \tag{7}$$

則代表爲挑選出來的候選端點，最後我們得到這一連串候選端點的序列，$\{c_j; j=1,...,N_c\}$。經過預選擇步驟後，候選端點會將語音信號分割成很多音段(segment)，我們也可由這些語音信號的音段求取一些 segment-based 之聲學參數來協助端點偵測，如圖六所示。



圖六、利用候選端點將語音信號分割成音段的示意圖

在此，我們使用音段中正規化後的各個頻段之信號波封平均值來評斷相鄰 2 個音段 $[c_{k-1},c_k]$、$[c_k,c_{k+1}]$之語音特性。其中 $ES_i(k)$爲在第 k 個音段($[c_k,c_{k+1}]$)中正規化後的各個頻段之信號波封平均值，可定義成下式：

$$ES_i(k) = \left(\sum_{n=c_{k-1}+1}^{c_k-1} E_i[n] \right) / \left(c_k - c_{k-1} - 2 \right) \tag{8}$$

接著，我們對於每個候選端點建立一個 30 維的參數向量(feature vector)，對於第 k 個候選端點，c_k，其參數向量包括以下聲學參數，

(1) 目前候選端點之參數：

$$\{fb_i[k],\ ror\_fb_i[k],\ d_k(n,n+1),\ F(k),\ env(k),\ ror\_env(k), H_s(k),\ ror\_H_s(k)\}; i = 1,...,6 \tag{9}$$

(2) 目前候選端點前($ESp_i(k)$)、後($ESn_i(k)$)音段之參數：

$$\{ESp_i(k)=\left(\sum_{n=c_{k-1}+1}^{c_k-1}E_i[n]\right)/\left(c_k-c_{k-1}-2\right),\ ESn_i(k)=\left(\sum_{n=c_k+1}^{c_{k+1}-1}E_i[n]\right)/\left(c_{k+1}-c_k-2\right)\};i=1,...,6 \tag{10}$$

最後使用 2 個指標指出此候選端點是否爲此候選端點序列之第一個或最後一個端點。

由先前所述的方法，我們已利用 sample-based 聲學參數的調整得到較佳之切割位置以訓練 MLP 端點偵測器。然而最重要的問題是要如何決定 MLP 音素端點偵測器之目標函數。由於所使用的參數不但能描述波形變化，也能使用帶通信號波封來辨別語音之發音特性。因此，我們定義 9 大類的目標函數，分別表示候選端點出現在短停頓(short pause, IS)、聲母(consonant, IC)、韻母(vowel, IV)、韻尾鼻音(nasal endings, IN) 4 種分類與彼此分類的轉換點，短停頓變化至聲母或是韻母(SC)、聲母至韻母(CV)、韻母接韻尾鼻音(VN)、韻母或韻尾鼻音變化至短停頓(VS)和可略過短停頓轉換點(CP)，而目標函數之間的轉移機率由目標函數所產生的 likelihood 作正規化後計算。

在我們的系統中，應用於 MLP 音素端點偵測之訓練演算法的流程。我們利用程式調整過後的 HMM 切割位置，再經過預選擇挑選出來的所有候選端點，進行分類標記來訓練 MLP 音素端點偵測器。

訓練演算法的過程如下：

(1) 經過分類標記的所有候選端點，當做初始目標函數；
(2) 利用給定的目標函數來訓練 MLP-based 音素端點偵測器直至收斂；
(3) 由 MLP-based 偵測器輸出目標函數的 likelihood 來計算候選端點之轉移機率，並依照其轉移機率使用 Viterbi search 得到最佳路徑，再重新進行候選端點的分類標記並將其標記結果當作 MLP 端點偵測器新的目標函數；
(4) 重複(2)與(3)的步驟，直至收斂。

五、Sample-based 音素端點偵測方法實驗結果

本章節主要是將我們所提出的 sample-based 音素端點偵測方法，運用 MLP 類神經網路的架構訓練一個音素端點偵測器，觀察並分析其切割位置之結果。同時以 frame-based HMM 架構之切割位置來比較其結果，觀察本研究所提出的方法切割位置之精準度是否有進一步地提升。

（一）、語音資料庫簡介

本文中的自動語音切割實驗所使用的 TCC-300 麥克風語音資料庫是由國立交通大學、國立成功大學、國立台灣大學所共同錄製，中華民國計算語言學學會所發行，此語料庫屬於麥克風朗讀語音。其中台灣大學語料庫主要包含詞以及短句，文字經過設計，考慮了音節及其相連出現機率，由 100 人錄製而成；成功大學及交通大學主要包含長文語料，文章由中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百個字，

再切割成 3-4 段，每段至多 231 字，由 200 人朗讀錄製，每人所朗讀之文章皆不相同。語音的取樣頻率爲 16kHz，取樣位元數爲 16 位元。

（二）、實驗驗環境與實驗架構設定

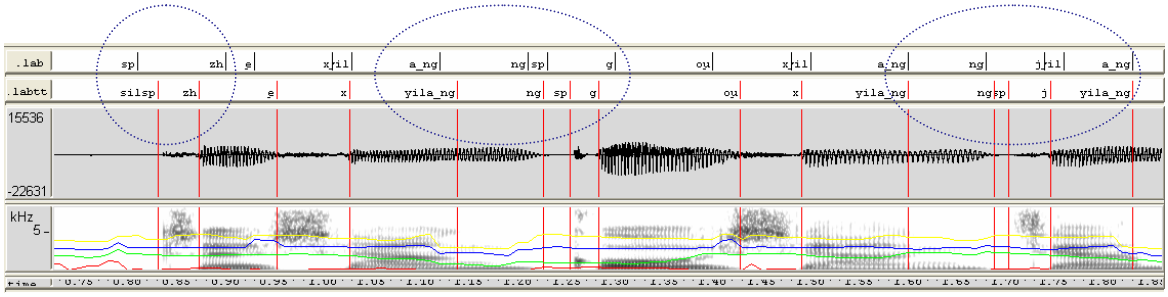
在 TCC-300 語音資料庫之語料選取方面，本論文使用交通大學與成功大學所錄製的長文語料，並隨機選取六分之五的部份當作訓練語料，其它部分爲測試語料。首先使用 SAT 及 SA 技術之 HMM phone alignment 流程，獲得較佳的 HMM 模型後進行強迫對齊之切割結果，作爲 TCC-300 語料庫之類音素起始切割位置。

在本研究使用 NICO Toolkit[12]來訓練我們的 MLP 端點偵測器，並採用 30×50×9 的 MLP 類神經網路架構分爲 3 層，包含一個輸入層、一個隱藏層、一個輸出層，輸入層點數共 30 點，分別輸入 6 個頻段之信號波封及其波封之上升率、頻譜熵及其上升率，波形之波封及其上升率、sample-based spectral KL distance、spectral flatness，與前、後音段正規化 6 個頻段信號波封後的平均值等參數。

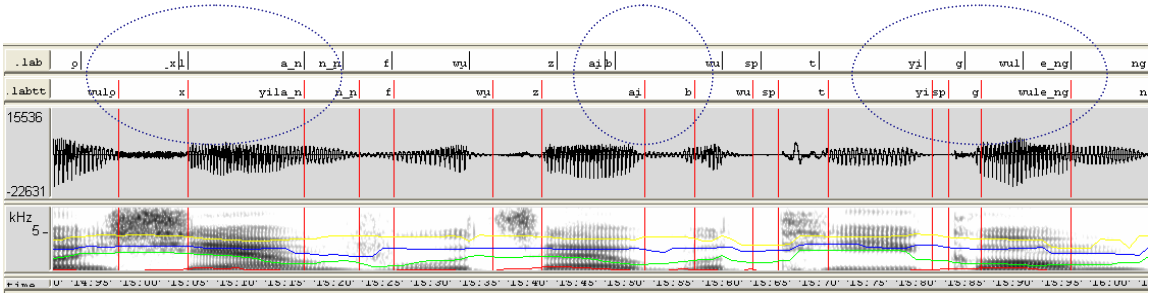
獲得 MLP 端點偵測器後，我們可以使用 HMM 切割位置作爲初始切割位置再以 sample-based MLP 類音素端點偵測器來做更精確之切割。我們將 sample-based MLP 類音素端點偵測器偵測之端點限制於 HMM 初始切割之正負 100 ms 範圍內，亦即起始切割位置不須十分精確，我們都可以找到正確端點。因此我們利用一個加上限制範圍的 Viterbi search 方法來獲得新的切割位置。

（三）、MLP 音素端點偵測器實驗結果比較與分析

在此實驗中我們觀察 Viterbi search 限制範圍是 100 ms 之音素端點偵測結果，並列舉數個圖形比較音素端點偵測結果與 HMM 語者調適模型強迫對齊的切割位置。可以由下列圖八、圖九之中看到 HMM 強迫對齊切割位置之結果，對於音素的切割位置經常有誤差存在，尤其是聲母之前的短停頓之切割效果不好或是沒有切出來，造成聲母長度普遍變長之現象，同時其聲母與韻母之間的切割位置亦不甚理想。我們同樣可由圓圈圈選處之音素端點位置觀察到，無論是音節與音節之間的短停頓或是聲母與韻母之間的端點位置都非常準確。圖八所示之橢圓形圈選處，我們亦可發現在韻母轉變至鼻音韻尾的情形，其音素端點位置之準確度仍能保持良好的水準，由上述結果皆可證實其 sample-based 的聲學參數具有偵測發音特徵變化之效能。



圖八、國語語句音素端點偵測之例子，由上至下的圖形分別表示原語者調適 HMM 切割位置及音素端點偵測之切割位置、波形、頻譜



圖九、國語語句音素端點偵測之例子，由上至下的圖形分別表示原語者調適 HMM 切割位置及音素端點偵測之切割位置、波形、頻譜

在此我們統計了 HMM 模型強迫對齊切割各發音方法之平均音長，並與我們提出的方法作比較，且本研究之音素端點偵測器將介音與韻母合併視爲單一韻母偵測，故韻母之平均音長不作比較。而由各發音方法之平均音長，觀察發現 HMM 語者調適模型之結果(表二)較音素端點偵測器之平均音長(表三)平均結果皆多出 10-20ms 以上的範圍，其原因在於 HMM 音素端點之切割位置皆有誤差，而 sample-based 音素端點偵測器皆能準確地將短停頓的位置標記出來使得聲母之平均音長下降，特別是爆破音與流音之平均音長下降 20-30ms 以上，明顯地較 HMM 切割位置之平均音長更加符合合理的範圍。

表二、HMM 語者調適模型切割位置各發音方法平均音長

單位： 10ms 發音方法	各發音方法平均音長
爆破音　Stop	4.96
鼻音　Nasal	5.95
摩擦音　Fricative	11.13
塞擦音　Affricate	8.92
流音　Liquid	6.23

表三、MLP 自動端點標示各發音方法平均音長

單位： 10ms 發音方法	各發音方法平均音長
爆破音　Stop	2.62
鼻音　Nasal	4.46
摩擦音　Fricative	8.75
塞擦音　Affricate	7.13
流音　Liquid	2.70

六、結論

本篇論文在 TCC-300 語音資料庫無正確音素人工標示資訊下，第一階段使用 SAT 及 SA 技術之 HMM phone alignment 流程，獲得較佳的 HMM 之切割位置資訊，作爲

TCC-300 語料庫之音素起始切割位置；在第二階段我們提出 MLP-based 音素端點偵測器的架構並加入數個 sample-based 的聲學參數對語料庫做自動化類音素單元之端點標示工作。實驗結果顯示，由於 HMM 切割位置的不準確會造成聲母過長或是無法正確切割出短停頓的情形均有大幅改善，證實這些以往使用於 frame-based 之聲學參數在 sample-based 的應用上也確實有顯著的效果。語音信號對於發音方式的不同會有不同的特性，而語音屬性應該是與語言無關的，因此在未來我們即可利用此性質對國內經常使用的聲調語言類型如閩南語、客語等方言來進行跨語言的自動端點標示的工作，並將此架構應用於有人工切音位置之 TIMIT 語料庫以評估此方法之效能。

七、參考文獻

[1] Toledano, D.T.; Gomez, L.A.H.; Grande, L.V., "Automatic phonetic segmentation," *Speech and Audio Processing, IEEE Transactions on* , vol.11, no.6, pp. 617-625, Nov. 2003.

[2] J. -W Kuo and H.-M Wang, "Minimum Boundary Error Training for Automatic Phonetic Segmentation," *The Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)*, September 2006.

[3] J.-W. Kuo, H.-Y. Lo, and H.-M. Wang, "Improved HMM/SVM methods for automatic phoneme segmentation," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2057-2060.

[4] K.-S. Lee, "MLP-based phone boundary refining for a TTS database," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 981–989, 2006.

[5] Sorin Dusan and Lawrence Rabiner, "On the Relation between Maximum Spectral Transition Positions and Phone Boundaries, ", in *Proc. Interspeech* 2006, pp. 17–21.

[6] Almpanidis, G., Kotti, M., Kotropoulos, and C., "Robust Detection of Phone Boundaries Using Model Selection Criteria With Few Observations," *IEEE Transactions on Audio, Speech, and Language Processing,* vol.17, no.2, pp.287-298, Feb. 2009.

[7] Sharlene A. Liu, " Landmark detection for distinctive feature-based speech recognition, ", J. Acoust. Soc. Am. **100** (5), November 1996, pp. 3417-3430.

[8] Hasegawa-Johnson, etc. "Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop," *Acoustics, Speech, and Signal Processing, 2005. ICASSP 2005*. vol.1, no., pp. 213-216, March 18-23, 2005

[9] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR,", in *Proc. ICASSP* 2004, pp. 193–196.

[10] Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee, "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments", *Proc. ICSLP* 1998.

[11] J. D. Markel and A. H. Gray, "A spectral-flatness measure for studying the autocorrelation method of linear prediction speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 207-217, June 1974.

[12] Nico Tool Kit : Available: http://nico.nikkostrom.com/

基於離散倒頻譜之頻譜包絡估計架構及其於語音轉換之應用

A Discrete-cepstrum Based Spectrum-envelope Estimation Scheme and Its Application to Voice Transformation

古鴻炎　　蔡松峰
Hung-Yan Gu　and　Song-Fong Tsai

國立台灣科技大學 資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
e-mail: guhy@mail.ntust.edu.tw

摘要

基於前人以離散倒頻譜來逼近頻譜包絡之觀念、及其穩定化係數值之求解方法，本論文進一步研究實際實施時所面臨的兩個問題，其一是"頻譜峰點挑選"的問題，其二是"頻率軸尺度轉換"的問題，對這兩個問題我們提出了不錯的解決方法，然後用以建構一個頻譜包絡的估計架構(scheme)，測試實驗顯示該架構所估計出的頻譜包絡，確實比原始方法所估計出的準確不少。接著，我們應用所提出的頻譜包絡估計之架構，去製做出一個語音轉換系統，經由頻譜包絡估計、頻譜包絡伸縮、基頻移動、和信號重新合成等處理步驟，可把輸入語音信號的音色轉換成不同性別和年齡的其它音色。由聽測實驗的結果顯示，我們的語音轉換系統，的確可有效地達成音色轉換的功能。

關鍵詞: 頻譜包絡, 離散倒頻譜, 語音轉換, 音色轉換

一、前言

這裡"頻譜包絡"指的是頻譜振幅包絡(magnitude-spectrum envelope)，關於一個語音音框的頻譜包絡的估計，先前研究者已提出了一些方法，例如基於線性預測編碼(linear prediction coding, LPC)之方法[1, 2]，以全極(all pole)模型之頻率響應曲線來逼近語音的頻譜包絡，不過 LPC 頻響(頻率響應)曲線，在一個共振峰頻率的附近會比理想的頻譜包絡曲線低，而在頻譜變化較快速的頻率區段，則會比理想的頻譜包絡曲線高很多，如圖 1 裡一個/i/音音框的 LPC 頻響曲線所示，所以 LPC 頻響曲線和理想的頻譜包絡曲線之間會存在著不能忽略的誤差，這樣的誤差在一些應用裡(如語音轉換)，將會造成語音品質的衰退。

除了 LPC 之外，過去也有幾個以倒頻譜(cepstrum)爲基礎的頻譜包絡估計方法被提出，最簡單的一個是倒頻譜平滑法[1]，此法只保留倒頻譜係數的前幾個，而把後面的係數全部砍除(即令爲 0 值)，再作離散傅利葉轉換(discrete Fourier transform , DFT) ，就可得到平滑的頻譜曲線，如圖 1 裡下方的那一條平滑曲線，很明顯地這樣的一條頻譜曲線並不是頻譜包絡，因爲它走在原始 DFT 頻譜的波峰與波谷之間，而不是沿著波峰行走。因此，Imai 和 Abe 兩人提出一個以倒頻譜爲基礎再作改進的方法[3, 4] ，稱爲 true envelope 估計法，然而此法的計算量很大而缺乏效率。另外，Galas 和 Rodet 兩人提出以離散倒頻譜(discrete cepstrum)來估計頻譜包絡的觀念[5]，後來 Cappé 和 Moulines 兩人則提出穩定化(regularization)的技術[6]，以解決使用離散倒頻譜來逼近頻譜包絡時所遇到的困難。我們覺得基於離散倒頻譜之估計法是一個不錯的方法，因此就著手研究此

方法，並且設法解決實際使用時所遇到的問題。

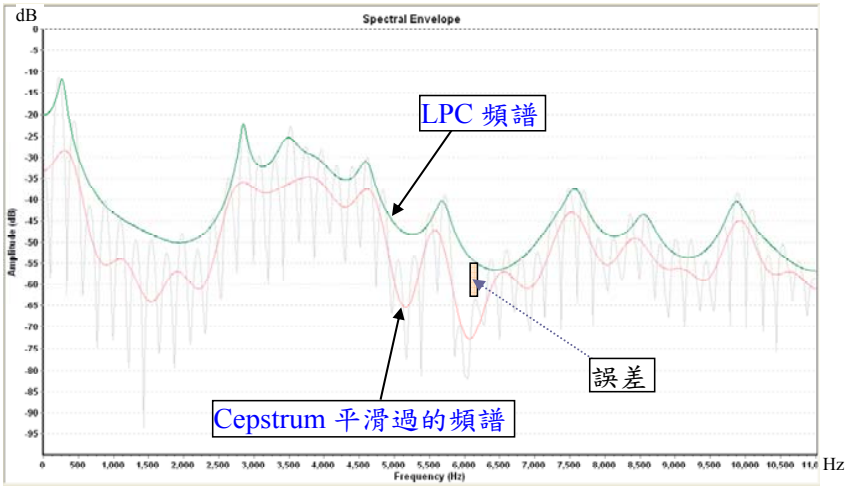


圖 1　/i/音音框之 LPC 頻譜包絡 和 倒頻譜平滑後之頻譜

本論文以離散倒頻譜之估計法爲基礎，研究、提出一個頻譜包絡估計的架構(scheme)，架構如圖 2 所示之處理流程，一個輸入的 20ms 之語音音框，首先進行基頻

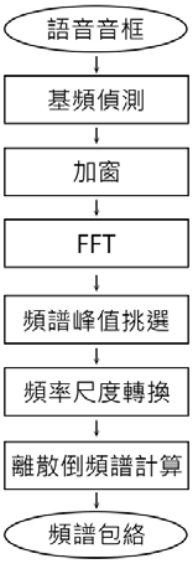


圖 2　頻譜包絡估計之架構

的偵測，以求出該音框的基頻值及判斷是否爲有聲(voiced)，求出的基頻值將會在”頻譜峰值挑選”之方塊內使用，在此我們使用了一種自相關(autocorrelation)函數搭配 AMDF (absolute magnitude difference function)的基頻偵測方法[1, 7, 8]。接著，將該音框信號乘上漢寧(Hanning)窗[1]，並於信號序列後面補上零，使序列長度成爲 1024 點，然後對該序列作 FFT (fast Fourier transform)計算而得到頻域上的頻譜振幅。之後，對頻譜振幅曲線作峰點(spectral peak)的挑選，並將選中的振幅峰點的頻率值作頻率尺度之轉換。使用挑選出的振幅峰點之振幅值和相對應的頻率值，就可帶入頻譜包絡的逼近準則(critera)[6]，去解出離散倒頻譜係數的數值，而離散倒頻譜係數則可用以算出所逼近的頻譜包絡。

在圖 2 的流程裡，我們主要作探討且提出新作法的是，”**頻譜峰點挑選**”和”**頻率尺度轉換**”兩方塊。雖然，核心的”**離散倒頻譜計算**”方塊，我們只是直接參照前人的成果

[6]，但是，如果沒有挑選到正確的頻譜峰點，或者沒有使用正確的頻率尺度，則"**離散倒頻譜計算**"所逼近出的頻譜包絡曲線，仍然會出現不能忽略的誤差，而不能算是理想的頻譜包絡曲線。由於過去的文獻，並未記載"**頻譜峰點挑選"**和"**頻率尺度轉換"**的確實作法，因此我們便著手研究這兩個問題，詳細情形在第三節和第四節分別說明。此外，我們也把圖 2 的頻譜包絡估計架構，應用於作語音轉換(voice transformation)，例如把成年女性的原始發音，經由轉換處理而得到女小孩的聲音、或成年男生的聲音，詳細情形在第五節裡說明。

二、 基於離散倒頻譜之頻譜包絡估計

2.1 離散倒頻譜

離散倒頻譜之觀念是由 Galas 和 Rodet 所提出[5]，他們採取以頻域上的最小平方準則(least-squares criterion)來求取倒頻譜係數，這和原本的實數倒頻譜(real cepstrum)係數的求取方式不同。原本的求取方式是把對數頻譜振幅, $\log|X(k)|$, k=0,1, …, N-1, 去作反離散傅利葉轉換(IDFT)而得到，令所得的倒頻譜係數爲 $c_0, c_1, \ldots, c_{N-1}$，之後，再將這些倒頻譜係數作 DFT，就可還原求得對數振幅頻譜，其公式[8]爲

$$\log\left|X(k)\right| = \sum_{n=0}^{N-1} c_n e^{-j\frac{2\pi}{N}kn}, \quad 0 \le k \le N-1 \quad (1)$$

由於對數頻譜振幅 $\log|X(k)|$是偶對稱的，即 $\log|X(k)| = \log|X(N-k)|$，所以由對數頻譜振幅求出的倒頻譜係數也是偶對稱的，即 $c_k = c_{N-k}$，依據這個偶對稱的特性，公式(1)可被推導成爲

$$\log\left|X(k)\right| = c_0 + 2\sum_{n=1}^{N/2-1} c_n \cos(\frac{2\pi}{N}kn) \; + \; c_{n/2}\cos(\pi k), \quad 0 \le k \le N-1 \quad (2)$$

這是因爲在轉換核心的虛部(imaginary part) 奇函數 sin(·)和偶對稱的倒頻譜係數序列會加總成 0 值。

若公式(2)裡只保留前面少數幾個(例如 p+1 個)倒頻譜係數，則它可用以計算出一個平滑過的振幅頻譜，計算公式爲

$$\log S(f) = c_0 + 2\sum_{n=1}^{p} c_n \cos(2\pi fn) \quad (3)$$

但是，如果想要以公式(3)來逼近頻譜包絡，則公式(3)裡的倒頻譜係數 c_k，就不是使用 IDFT 來求取了，而是要先定義一些欲被滿足的頻譜包絡限制，然後在儘量滿足這些頻譜包絡限制的條件下，去求解出最佳的倒頻譜係數 c_k 的數值，如此求出的倒頻譜係數就稱爲離散倒頻譜係數。前述所謂的頻譜包絡限制，其實是從原始的 DFT 頻譜$|X(k)|$上，找出 L 組具有代表性的頻譜振幅峰值及其頻率(a_k, f_k), k=1, 2 …, L。由於 L 通常比倒頻譜階數 p 大許多，所以需要使用一個加權式最小平方準則(weigthed least-squares criterion)，來最小化這 L 個頻率點上(f_k, k=1, 2 …, L)，$S(f_k)$和 a_k之間的誤差，也就是最小化

$$\varepsilon = \sum_{k=1}^{L} w_k \cdot \left|\log a_k - \log S(f_k)\right|^2 \quad (4)$$

其中 w_k 是一個權重值與頻率 f_k 有關，對於不同的頻率值給予不同的加權，可藉以求得較好的頻譜包絡。若把公式(4)以矩陣形式來表示，則可改寫成下式[6]

$$\varepsilon = \left|a - Mc\right|^2 \cdot W = (a - Mc)^T W (a - Mc) \quad (5)$$

其中 $a = \left[\log(a_1), ,\cdots \ \log(a_L)\right]^T$ ； $c = [c_0, \cdots; \ c_p]^T$ 是一個維度 p+1 的向量，代表未知的倒頻譜係數；

$W = \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_L \end{bmatrix}$ 爲對角矩陣，對角原素爲加權值；

$$M = \begin{bmatrix} 1 & 2\cos(2\pi f_1) & 2\cos(2\pi f_1 2) & \cdots & 2\cos(2\pi f_1 p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2\cos(2\pi f_L) & 2\cos(2\pi f_L 2) & \cdots & 2\cos(2\pi f_L p) \end{bmatrix}$$

若要得到一組最佳的倒頻譜係數 c，也就是要最小化公式(4)的誤差，這可由公式(5)推導出如下之線性解

$$c = (M^T W M)^{-1} M^T W a \quad (6)$$

如此，透過矩陣逆轉換與矩陣相乘，即可得到一組最佳的離散倒頻譜係數。

2.2 離散倒頻譜之穩定化(regularization)

由前一小節的說明可知，離散倒頻譜的原理是，在頻域上以最小平方準則去求解倒頻譜係數，這樣的求解方法在實務上會遭遇到一個問題，而使它變成不實用，那就是 ill-conditioning 問題。由於矩陣 $M^T W M$ 通常是條件很差的矩陣，這將會導致包絡曲線有非常大的誤差，這意味著只要數據(a_k, f_k) 有些微的改變(例如四捨五入)，則計算出的倒頻譜係數就會有劇烈的變化，而包絡曲線也會因爲 f_k 的改變而產生過大的起伏。

以圖 3 爲例，虛線曲線所表示的是，使用 40 階離散倒頻譜係數、和非線性頻率尺度所求得的頻譜包絡，雖然包絡曲線有正確的經過前 7 個振幅峰點，不過峰點與峰點之間的曲線變化得非常劇烈。在許多頻譜包絡的應用裡，相鄰峰點之間的振幅值必須是可以計算的，如果有可能出現如圖 3 所示的情況，那麼此曲線在實務上將無法作爲頻譜包絡曲線。

先前研究者也發現，在以下三種情況下，上述問題發生的機會將會提高，分別是：(a)當有很寬的頻率軸區間沒有可逼近的振幅點時；(b)當兩個相鄰的頻率點 f_k, f_{k+1} 很靠近，並且此兩點的振幅差異很大；(c)當離散倒頻譜係數的個數 p 很接近欲逼近的頻率點數 L 時，通常在基頻較大時發生。因此，Cappé 和 Moulines 提出一個穩定化技術[6]，用以排除包絡曲線的不合理起伏，他們在求解一組離散倒頻譜係數時，除了依據最小化平方誤差準則，還將包絡曲線的平滑程度納入考慮，如此修改過的準則就變成如下公式：

$$\varepsilon = \sum_{k=1}^{L} w_k \cdot \left|\log a_k - \log S(f_k)\right|^2 \ + \ \lambda \cdot R\left(S(f)\right) \quad (7)$$

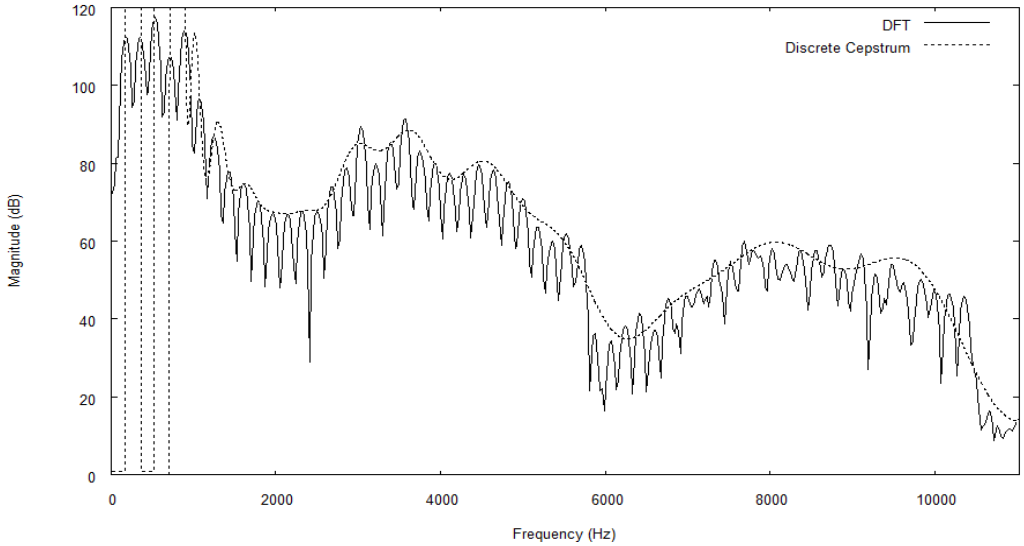


圖 3　使用 40 階離散倒頻譜係數、非線性頻率尺度求得之基本頻譜包絡

其中 $R(S(f))$是一個補償函數，用以量測包絡曲線的平滑程度，若包絡曲線越平滑則其值越小，反之越大；λ 是正規化參數，用以控制平滑程度和最小平方誤差準則之間的相對權重，λ 值越大則逼近出的頻譜包絡越平滑。一個典型的補償函數如下式[6]：

$$R\left(S(f)\right)=\int_0^{\pi}\left[\frac{d}{df}S(f)\right]^2 df \tag{8}$$

當把公式(3)式代入公式(8)中，可推導得

$$R\left(S(f)\right)=c^T U c\ ,\quad U=8\pi^2\begin{bmatrix}0 & & 0\\ & 1^2 & \\ & \ddots & \\ 0 & & p^2\end{bmatrix} \tag{9}$$

如此從公式(7)推導出的最佳解如下式[6]

$$c=(M^T W M+\lambda U)^{-1}M^T W a \tag{10}$$

其中正規化參數λ的較佳值在 0.0001 附近，如此矩陣 ill-conditioning 的問題就可獲得解決，並且依然能逼近出良好的頻譜包絡曲線，圖 4 裡的虛線曲線就是使用穩定化離散倒頻譜估計方法，所逼近出的頻譜包絡，很明顯地在低頻部分過度起伏的情況已經獲得改善，至於圖 4 裡的實線 DFT 曲線則與圖 3 裡的相同。

三、 頻譜峰值挑選

一般來說，頻譜包絡可看成是 DFT 振幅頻譜上連結各峰點(spectral peak)的曲線，因此離散倒頻譜的估計，採取以最小化所選出的峰點振幅值 a_k 與包絡曲線 $S(f)$之間的平方誤差作爲準則。由此可推知，峰點的挑選是一個相當重要的前置步驟，如果採取的是簡單的峰點挑選方法，例如選出全部的頻譜峰點，這將會導致不良的頻譜包絡曲線被逼近出來，而使用此種頻譜包絡進行語音編碼或音高調整，也將會降低語音品質和造成音色的不一致。

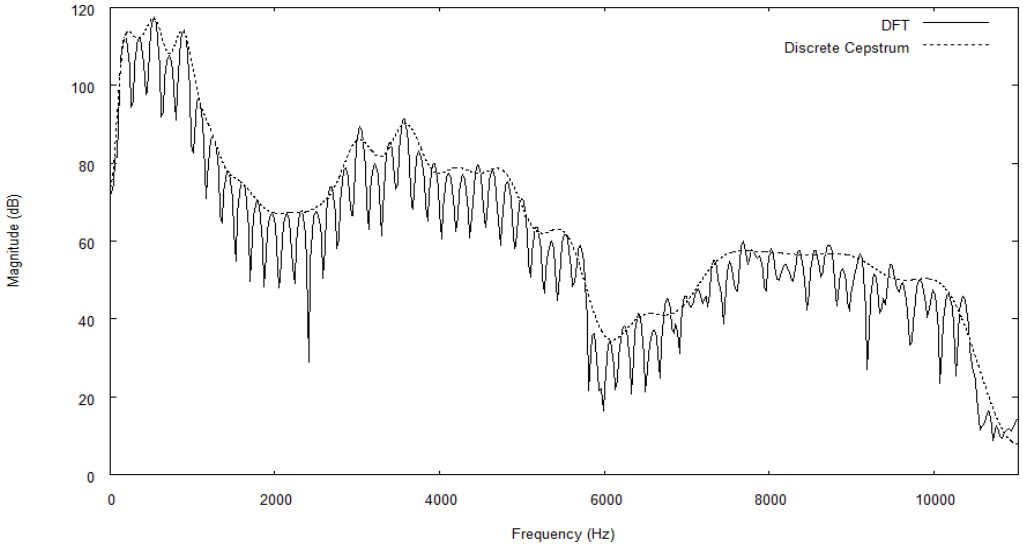


圖 4　使用 40 階離散倒頻譜係數、非線性頻率尺度求得之穩定化頻譜包絡

因此，我們採用 HNM (harm onic-plus-noise model) 之觀念[9, 10] ，當輸入的音框作基週偵測後判斷爲有聲時，就依據 HNM 偵測出的最大有聲頻率(maximum voiced frequency, MVF)，把該音框的 DFT 頻譜切割成低頻的諧波(harmonic)部分、和高頻的雜音(noise)部分。然後，對於諧波部分，依據偵測出的基頻值 F_0，在頻率範圍 [0 .5×F_0, 1.5×F_0] 內尋找出峰點的振幅值 a_1、及其對應的頻率值 f_1；接著在頻率範圍 [f_1+0.5×F_0, f_1+1.5×F_0] 尋找出另一峰點的振幅值 a_2、及其對應的頻率值 f_2；如此繼續找出其它峰點。如果在尋找的頻率範圍內沒有峰點，則會把尋找的範圍往後移動 0.5×F_0，再嘗試尋找峰點。此外，我們也設定了峰點振幅的門檻，以排除振福較小的峰點。

對於有聲音框的雜音部分，由於頻譜已無明顯的諧波結構，如圖 4 裡 5800Hz 之後的 DFT 頻譜曲線，相鄰峰點之間的頻率間隔變得隨機而非固定值，且峰點的振幅高度也變得隨機地起伏，因此我們認爲高頻雜音的頻譜，不能夠再使用與低頻部分相同的峰點挑選方法。在此，我們先使用 30 階的實數倒頻譜係數去計算出一個平滑過的頻譜曲線，然後依據此平滑的頻譜曲線，把 MVF 之後所有 DFT 頻譜振幅大於平滑頻譜振幅的峰點，都算是要選出的峰點。至於無聲的音框，我們就直接把 MVF 設爲 0，然後採取上述雜音頻譜峰點的相同挑選方法。圖 5 裡顯示了一個以我們的頻譜峰點挑選方法，所挑選出的頻譜峰點結果，選出的頻譜峰點以符號"+" 表示。

四、 離散倒頻譜階數與頻率軸尺度

4.1 離散倒頻譜之階數

由第二節的說明可知，離散倒頻譜係數的個數 p 必須先固定，然後才能去解 p 個聯立方程式來求得離散倒頻譜係數。至於 p 值要設爲多少? 若使用太小的 p 值(如 $p < 10$)，則包絡曲線的起伏次數會較少，而無法準確地逼近大多數的頻譜包絡形狀。然而隨著 p 值的增加，解聯立方程式的計算量也會增加，不過爲了準確地逼近大多數的頻譜包絡形狀，以避免音質下降及保持音色的一致，我們認爲加大 p 值是必要的。那麼 p 值應設爲多少? Shiga 和 King 曾提到[11]，若要得到精確的頻譜包絡，則較高階數(如 48~64)的倒頻譜係數是需要的。

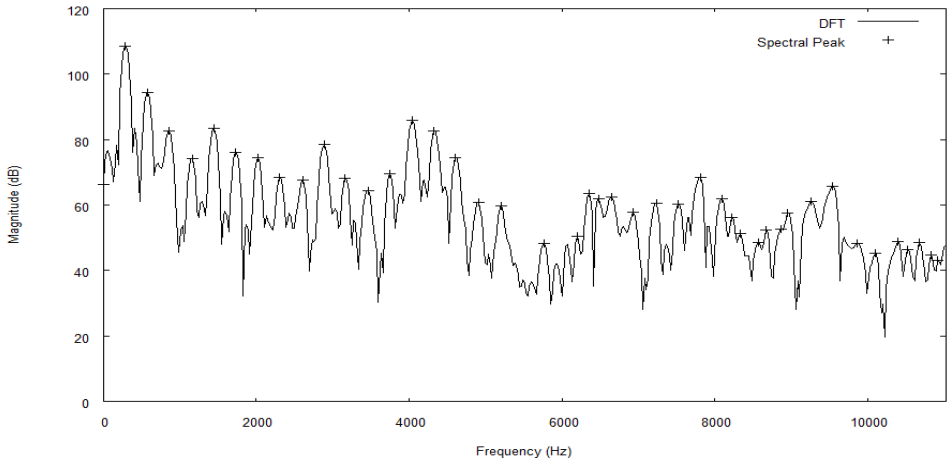


圖 5　一個頻譜峰點挑選的例子

在此，我們以實驗的方式來探討離散倒頻譜階數和頻譜包絡之逼近誤差的關係，實驗裡我們使用的誤差量測公式如下:

$$Es = \frac{1}{Nr}\sum_{t=0}^{Nr\text{-}1}\left[\frac{1}{L}\sum_{k=1}^{L}\left|20\log_{10} a_k^t - 20\log_{10} S(t,f_k)\right|\right] \qquad (11)$$

其中 Nr 表示語音音框的總數，a_k^t 表示第 t 個音框裡的第 k 個頻譜峰點的振幅，$S(t,f_k)$ 表示第 t 個音框以離散倒頻譜所逼近出的頻譜包絡。實驗後計算出的的逼近誤差如圖 6 所示，橫軸表示離散倒頻譜的階數，縱軸則是量得的逼近誤差 Es 的數值，觀察圖 6 可發現， 隨著階數的增加 Es 值會明顯地下降，直到階數高於 30 時，Es 值下降的幅度才趨於緩和，因此我們決定把 p 值設定爲 40，以確保逼近出的頻譜包絡的準確性。

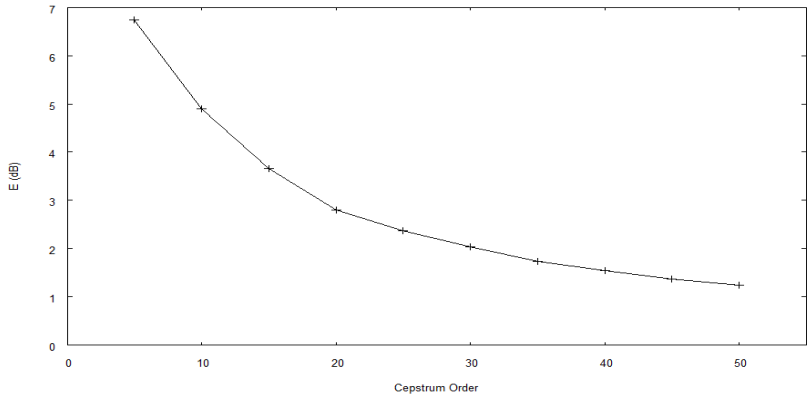
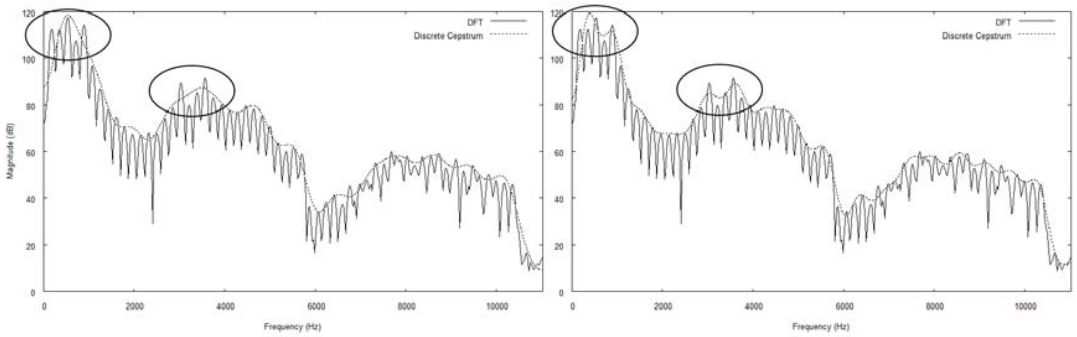


圖 6　不同階數之離散倒頻譜的頻譜包絡逼近的誤差

4.2 頻率軸尺度

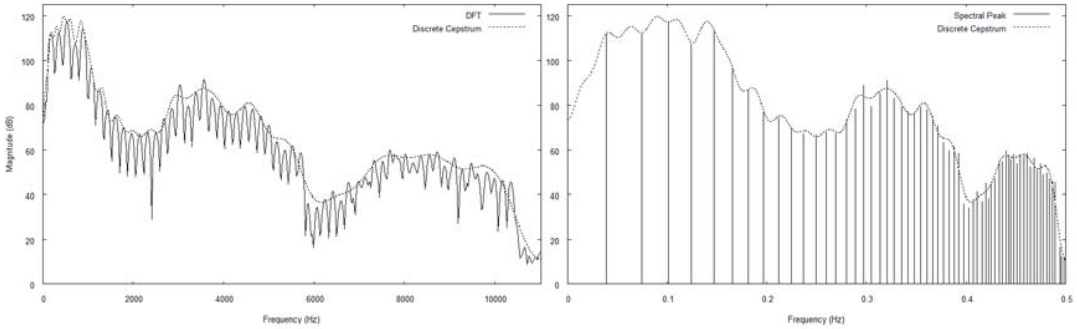
雖然高階數的離散倒頻譜已經可以逼近出不錯的頻譜包絡，不過當我們觀察某些語音信號的頻譜時，發現高階數的離散倒頻譜所逼近出的頻譜包絡上，仍然會出現不小的、不能忽略的逼近誤差。例如圖 7(a)裡的頻譜包絡曲線，它是使用 30 階的離散倒頻譜所逼近出的，圖中圈起來的部分顯示發生較大誤差的地方，如果我們將階數提高到 40，則會得到如圖 7(b)裡的頻譜包絡曲線，雖然 1,000Hz 與 3,000Hz 附近的誤差，已經獲得了一些改善，不過還是不夠理想。



(a)使用 30 階之離散倒頻譜　　(b)使用 40 階之離散倒頻譜

圖 7 以線性頻率尺度之離散倒頻譜所逼進之頻譜包絡

圖 7 裡的情況通常是發生在基頻較低的語音音框裡，由於相鄰峰點的頻率值非常接近，而發生頻譜包絡快速變化的情況，這種快速變化的頻譜包絡無法由低階數的離散倒頻譜去作準確的逼近，尤其是在低頻的區段。要解決這種問題，一個普遍被採取的觀念是，使用非線性的頻率軸來擴大低頻區段在整個頻率軸所佔的比率，而常見的非線性頻率尺度如梅爾尺度(Mel Scale)或巴克尺度(Back Scale)。實際上的作法是，在挑選出頻譜峰點之後，先將各峰點對應的頻率 f_k 作頻率尺度的轉換 $\hat{f}_k = warp(f_k)$，才去求解離散倒頻譜係數，之後當要計算所逼近的頻譜包絡時，也需作頻率軸尺度的轉換。圖 8(a)裡的頻譜包絡曲線就是使用梅爾尺度所逼近出的，相較於圖 7(b)的線性尺度，低頻的峰點都有確實地被包絡曲線通過，但是 3,000Hz 附近的峰點仍然存在誤差。

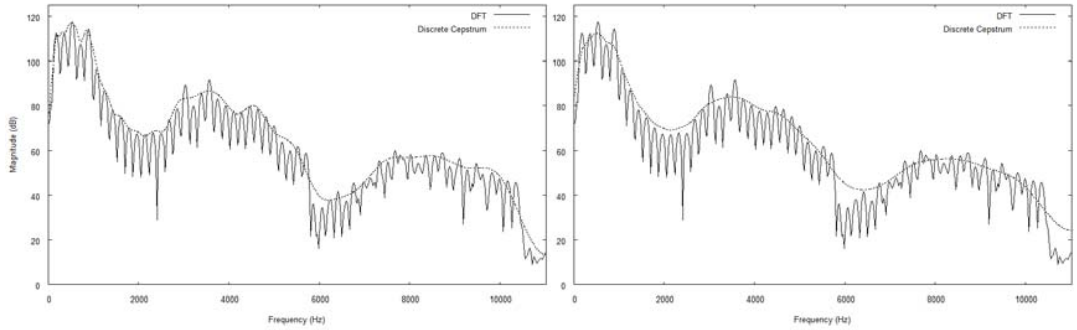


(a)橫軸爲線性頻率之包絡曲線　　(b)橫軸爲梅爾頻率之包絡曲線

圖 8　以梅爾頻率尺度和 40 階之離散倒頻譜所逼進的頻譜包絡

此外，在圖 8(a)的低頻部分，也可看到頻譜包絡曲線過度起伏的情況，原因是梅爾頻率尺度會把低頻部分的相鄰峰點的間隔擴大而引起過度起伏，這種過度起伏可從圖 8(b)裡觀察到。如果改成使用巴克頻率尺度，則將使包絡曲線過度起伏的情況更加嚴重，因爲巴克尺度會讓低頻部分所佔的比率更爲變大，並且相鄰的高頻峰點間的間隔也會變得更窄，而會使高頻部分的包絡曲線變得更爲平滑。

第二節裡曾提到包絡曲線過度起伏的解決方法，亦即穩定化之技術。到目前爲止，我們都只把正規化參數 λ 設爲 0.0002，爲了改善上述的過度起伏情形，我們嘗試加大正規化參數 λ，結果得到如圖 9 所示的頻譜包絡曲線，圖 9(a)裡設定 $\lambda = 0.001$，而在圖 9(b)裡設定 $\lambda = 0.01$。雖然隨著正規化參數 λ 的增加，低頻部分的頻譜包絡曲線會漸趨平滑，不過頻譜包絡的起伏程度也受到了限制，而導致整體 L 個峰點的逼近誤差也隨著增加。



(a) $\lambda = 0.001$ (b) $\lambda = 0.01$。

圖 9　在梅爾頻率尺度和 40 階倒頻譜之條件下加大穩定化參數 λ

因此，我們嘗試調整頻率軸尺度，也就是調整低頻部分在整個頻率軸所佔的比率，目的是讓離散倒頻譜階數大於 30 時，頻譜包絡曲線不會在低頻部分出現過度起伏的情況，並且在 2,000Hz ~ 6,000Hz 的頻率軸區段可以減小峰點振幅的逼近誤差。經過多次的實驗觀察，我們歸納出一種如下列公式所示之頻率尺度，

$$warp(f) = \log(1 + \frac{f}{1{,}750}) \tag{12}$$

圖 4 裡的頻譜包絡曲線就是使用此種頻率尺度來求解離散倒頻譜而逼近出的，可以發現小於 1,000Hz 之頻率部分，頻譜包絡曲線沒有過度起伏之情形，並且可準確地通過各個峰點，此外在 2,000Hz ~ 6,000Hz 的頻率部分，頻譜包絡曲線比起梅爾尺度所求出者，有著更佳的起伏能力。

爲了比較兩種頻率尺度，即我們提出的公式(12)之頻率尺度和梅爾頻率尺度，對於以離散倒頻譜逼近頻譜包絡所產生的誤差，在此我們就分別在不同的頻率範圍做逼近誤差的量測實驗，四種頻率範圍分別是 (a) 0 ~ 2,000Hz，(b) 0 ~ 4,000Hz，(c) 0 ~ 6,000Hz，(d) 0 ~ 11,025Hz，而誤差量測的方式仍然如公式(11)。實驗後我們得到如圖 10 所顯示的結果，也就是說在 0 ~ 2,000Hz 之頻率範圍，兩種頻率尺度有著類似的逼近誤差，但是在其它三種頻率範圍作量測時，梅爾尺度的頻譜包絡逼近誤差，明顯地會高於我們提出的頻率尺度，並且頻率範圍愈大時，我們所提的頻率尺度的逼近誤差會和梅爾尺度的逼近誤差愈來愈拉開。

五、語音轉換之應用

5.1 語音轉換系統

這裡所說的語音轉換(voice transform ation)，指的是作頻譜包絡曲線的伸展或收縮(相當於作頻率軸的 scaling)，及基本頻率的移動(frequency shifting)，以把輸入語音信號的音色改變成另一個人的音色，例如把成年女生的語音轉變成成年男生的語音、或小孩的語音。過去，作語音轉換常被使用的是 phase vocoder (PV) 之技術[12, 13]，但是基本的 PV 技術，並不能讓**頻譜包絡伸縮**和**基頻移動**兩者作獨立的控制。在本論文裡，我們決定採取電腦音樂之加法式合成法(additive synthesis) [12]及 HNM 的觀念[9]，來讓頻譜包絡伸縮和基頻移動兩者可以被分別地控制，而實作上則需要應用前面說明的頻譜包絡之估計方法，來求得輸入語音各音框的頻譜包絡曲線，然後才據以作頻譜包絡的伸縮。我們所製做的語音轉換系統，其程式介面如圖 11 所示，pitch contour 區塊顯示從原始語音所分析出的基週軌跡，waveform 區塊上下分別顯示原始語音及轉換過語音的波形。

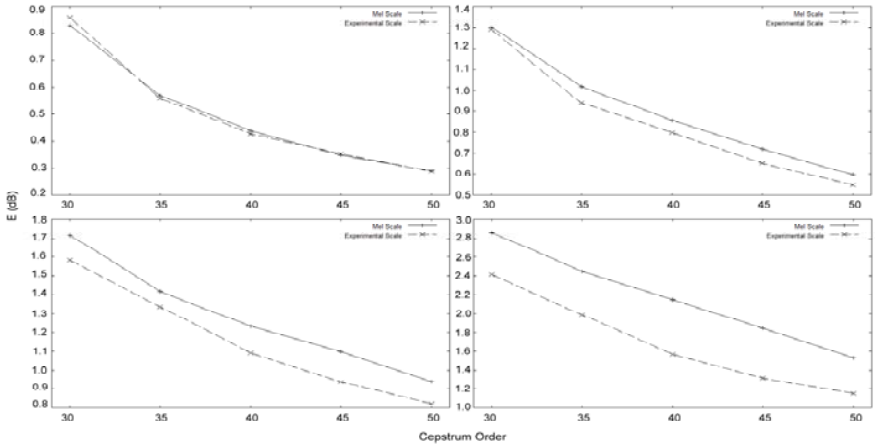


圖 10　在四種頻率範圍比較本論文尺度、梅爾尺度的頻譜包絡逼近之誤差: (左上) 0~2,000Hz，(右上) 0~4,000Hz，(左下) 0~6,000Hz，(右下) 0~11,025Hz

至於此系統的處理流程則如圖 12 所示，輸入的語音信號，先切割成長 20ms 重疊 10ms 的音框序列，然後對各個音框作如圖 2 所示的處理步驟，接著依據公式(10)去求取 40 個離散倒頻譜係數，再依據公式(3)就可計算出頻譜包絡曲線；至於"頻譜包絡伸縮"、"基頻移動"、"信號重新合成"等三個方塊的細節，將於下面各子節分別作說明。關於此系統的處理速度是，在 Intel T5600 1.83GHz CPU 的筆記本電腦上，處理 1 秒鐘的語音，平均需花 0.75 秒的時間。

5.2 頻譜包絡伸縮

不同年齡與性別的語者所發出的語音信號，在聲學上的主要差異是，語音頻譜上之共振峰頻率(formant frequency)值的高低會有明顯的差別。一般來說男生由於聲道(vocal track)較女生的長，所以男生語音的共振峰頻率值會比女生的低。因此，若要把輸入的語音信號轉換成不同性別與年齡的語音信號，則共振峰峰頻率的調整是必需的，不過，要求取出正確的共振峰頻率值、及直接修改它，並不是容易的事，因此我們採取對頻譜包絡作伸展或收縮的處理，以達到共振峰頻率的移動。

對頻譜包絡作伸、縮處理一個例子如圖 13 所示，圖 13(a)畫的是原始的頻譜包絡，令表示此包絡的函數是 $vs(f)$，如果我們對頻譜包絡作收縮，例如令收縮過的包絡的表示函數爲 $vc(f)$，且令 $vc(f) = vs(\frac{10}{7}f)$，則收縮的包絡將如圖 13(b)所示，如此就可把全部的共振峰頻率都調低；相反地如果我們對頻譜包絡作伸展，例如令伸展過的包絡的表示函數爲 $ve(f)$，且令 $ve(f) = vs(\frac{7}{10}f)$，則伸展的包絡將如圖 13(c)所示，如此就可把全部的共振峰頻率都調高。

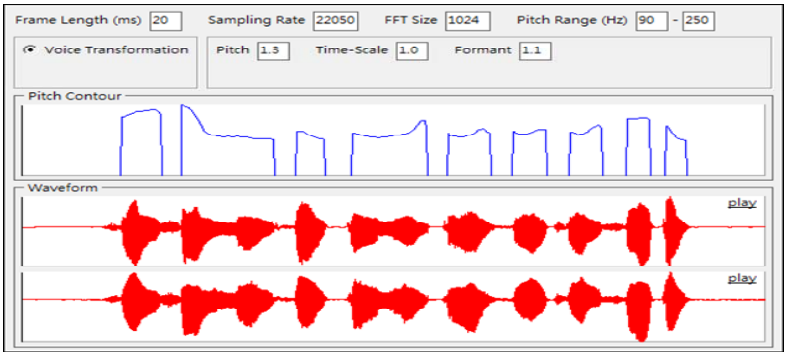


圖 11　語音轉換程式之介面

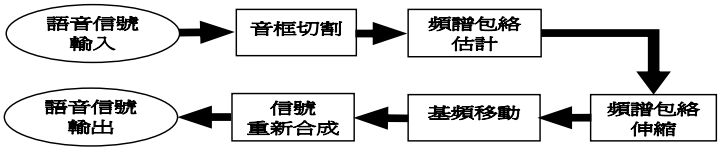
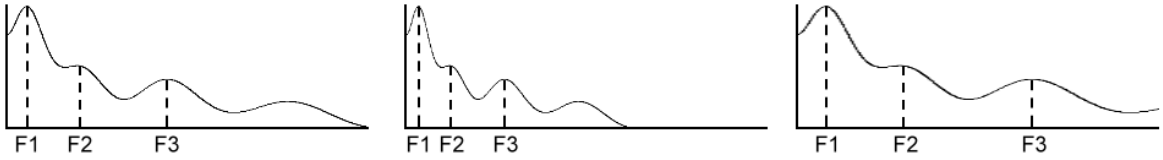


圖 12 語音轉換處理之主流程



(a)原始之頻譜包絡　　(b)收縮之頻譜包絡　　(c)伸展之頻譜包絡

圖 13 頻譜包絡之收縮與伸展

5.3 基頻移動

調整音高(pitch)的高低可讓聲音轉爲尖銳或低沉，但只調整男性語音的音高，並不能轉換出具有女性或孩童音色的語音，因此，若要把語音轉換成不同性別與年齡的語音，則共振峰頻率的調整(也就是頻譜包絡伸縮的調整)是必需先作的，然後才來作音高的調整。

當作完 5.2 小節的頻譜包絡伸縮之後，就可使用該頻譜包絡 $vc(f)$ (或者 $ve(f)$)來設定新的諧波參數，假設目前音框的原始基頻是 180Hz，而現在要把基頻調高到 250Hz, 則在此基頻的各個倍頻上的諧波，它們的振幅高度可根據 $vc(f)$來求得，也就是 $vc(250)$, $vc(500)$, $vc(750)$, …, 這相當於在新的基頻及其倍頻上對頻譜包絡取樣，用以取代原先的諧波頻率和振幅，結果會得到如圖 14 所示的新諧波結構。在此一個諧波的參數只有頻率和振幅，我們暫時不使用相位之資訊。

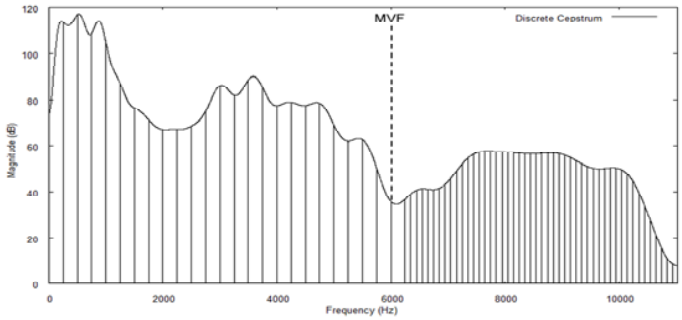


圖 14 基頻爲 250Hz 之諧波結構

5.4 信號重新合成

由於我們是根據語音音框在頻域上分析出的頻譜包絡，來製做語音轉換的功能，所以相對地也必須採取以頻域參數建構的信號模型，來作語音信號的重新合成。本論文採取的信號模型是諧波加雜音模型(harmonic-plus-noise m odel, HNM) [9, 10] ，它是由 Y. Stylianou 所提出，除了考慮語音信號裡低頻部分的諧波特性之外，還考慮了高頻部分的雜音特性，所以可以更確切地掌握語音信號的特性。HNM 模型提供了一個最大有聲頻率(maximum voiced frequency , M VF)的偵測方法，找出 MVF 之後，就可把一個語音音框的頻譜分割成低頻的諧波部分、和高頻的雜音部分，如圖 14 所示。

令第 i 個語音音框由 5.3 子節所求出的諧波參數是 f_k^i , a_k^i , k=1, 2, …, L^i , f_k^i 與 a_k^i 分

別表示第 k 個諧波的頻率與振幅；再令第 i+1 個語音音框由 5.3 子節所求出的諧波參數是 f_k^{i+1}, a_k^{i+1}, k=1, 2, …, L^{i+1}。如此，當要合成第 i 和第 i+1 音框之間時刻 t 的諧和(harmonic)信號之樣本 $h(t)$時，我們先以如下公式作線性內差，

$$f_k(t) = f_k^i + \frac{f_k^{i+1} - f_k^i}{N} t, \quad k = 1,2,...,L$$
$$a_k(t) = a_k^i + \frac{a_k^{i+1} - a_k^i}{N} t, \quad k = 1,2,...,L \tag{13}$$

以求取時刻 t 時各諧波的頻率與振幅，其中 N 表示相鄰音框之間的位移樣本數(frame shift)，L 是 L^i 和 L^{i+1} 的較大者，因此當 L^i 小於 L^{i+1} 時，就要把 a_k^i, k= L^i +1,…, L^{i+1} 設爲零值。然後，以如下公式計算 $h(t)$，

$$h(t) = \sum_{k=1}^{L} a_k(t) \cdot \cos(\phi_k(t)), \quad 0 \le t < N$$
$$\phi_k(t) = \phi_k(t-1) + 2\pi \cdot f_k(t) / 22,050 \tag{14}$$

其中 $\phi_k(t)$ 表示第 k 個諧波累積到時刻 t 時的相位量，關於初始值 ϕ_k(-1) ，我們令其等於前一音框裡的 $\phi_k(N\text{-}1)$ 以便維持相位的連續性，而當音框編號 i 爲 0 時就以亂數來設定，此外 22,050 是取樣率。

關於雜音(noise)信號的合成，我們採取 HNM 文獻上提到的一個作法，就是把雜音當作是 MVF 之後頻率間隔固定爲 100Hz、但振幅會隨時間改變之一些弦波的加總。令 mvf 爲第 i 和第 i+1 音框的 MVF 的較大者，則依 mvf 可決定頻率 index 之下限 KL=mvf / 100, 而其上限明顯地是 KU=22,050 / 100，如此，對於第 i 和第 i+1 音框之間時刻 t 的雜音信號樣本 $g(t)$，我們以如下公式來計算，

$$g(t) = \sum_{k=KL}^{KU} b_k(t) \cdot \cos(\psi_k(t)), \quad 0 \le t < N$$
$$\psi_k(t) = \psi_k(t-1) + 2\pi \cdot k \cdot 100 / 22,050 \tag{15}$$

其中 $b_k(t)$ 表示時刻 t 時第 k 個弦波的振幅，其值也是以類似公式(13)之線性內差來求得，$\psi_k(t)$ 表示第 k 個弦波累積到時刻 t 時的相位量，其初始值也是以亂數來設定。最後，將 $h(t)$與 $g(t)$相加，即可得到時刻 t 的合成信號樣本。

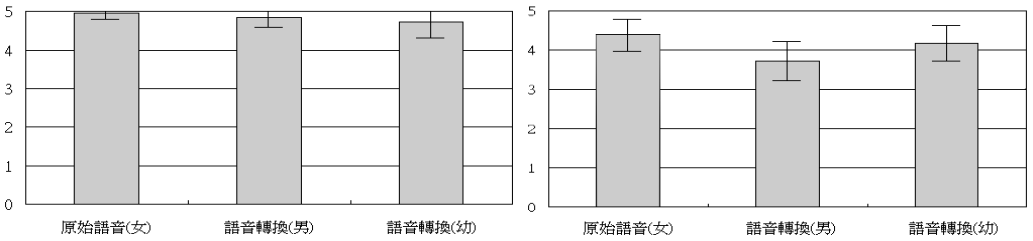
5.5 聽測實驗

爲了評估我們的語音轉換系統的效能，接著就進行主觀的聽測評估實驗。聽測的語料包含了成年女性發音的 3 句原始語句，和成年男性發音的 2 句原始語句。依據女性原始語句，我們系統進行了兩種轉換處理，第一種是設定頻譜包絡收縮成 80%、且基頻移動到原基頻的 60%，以轉換出男性的語音；第二種是設定頻譜包絡伸展成 130%、且基頻移動到原基頻的 140%，以轉換出孩童的語音。另外，依據男性原始語句，我們系統也進行了兩種轉換處理，第一種是設定頻譜包絡伸展成 120%、且基頻移動到原基頻的 210%，以轉換出女性的語音；第二種是設定頻譜包絡伸展成 130%、且基頻移動到原基頻的 150%，以轉換出孩童的語音。如此，對於女性和男性原始語句，各有 2 組轉換出的語句，可供作聽測評估，這些語句(原始的和轉換出的)，可從網頁 http://guhy.csie.ntust.edu.tw/dcc/vt.html 去下載和試聽。在此評估的項目有兩項，分別是音色辨識度、和語音品質，音色辨識度在於評估轉換出來的語音音色與目標音色的接近

程度，而語音品質在於評估轉換出來的語音信號聽起來是否清楚且無失真。

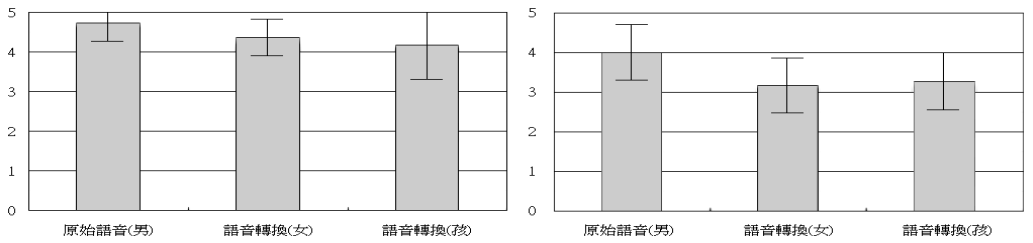
我們邀請了 13 位受測者來進行聽測評估，首先是對女性原始語句及轉換出的語句作聽測，實驗時一次讓一位受測者聆聽 3 組語句(女性原始、男性轉換、孩童轉換)，然後請他對這 3 組語句各給一個評分，評分的範圍由最高 5 分到最低 1 分，5 分代表非常相似(or 好) 而 1 分代表非常不相似(or 差)，可以打至小數點下一位。關於音色辨識度的評估，以女性語句爲例是詢問「聽測音檔的音色和女性語音音色的相似度爲何?」，至於男性語句、孩童語句可依此類推。關於語音品質的評估，詢問的方式則是「聽測音檔的語音品質是非常好、普通、或非常差?」。聽測實驗後，我們將 13 位受測者的評分作平均，所得到的平均值如圖 15 所示，在音色辨識度方面，轉換過的語句皆有相當高的辨識度，並且很接近原始語音的辨識度，如圖 15(a) 所示；在語音品質方面，轉換過的孩童語音跟原始音檔的語音品質比較接近，但下降一些，而轉換過的男性語音之語音品質則約比原始音檔的低 0.5 分。

另外，我們也請這 13 位受測者來對男性原始語句及轉換出的語句作聽測評估，實驗時一次讓一位受測者聆聽 3 組語句(男性原始、女性轉換、孩童轉換)，然後請他對這 3 組語句各給一個評分，評分方式如前段所述。實驗後將 13 位受測者的評分作平均，得到的平均分數如圖 16 所示，在音色辨識度方面，轉換出的孩童音色的分數比另二者的低一些，原因應是頻譜包絡伸展比例 130%不夠高，而使得音色感覺像國中生男孩而非孩童；在語音品質方面，轉換出的女性和孩童語句的分數都比原始男性語句的低，而且圖 16(b)三者的分數都分別比圖 15(b)三者的低，這似乎暗示，男性原始音作語音轉換所得的語音品質會比較差。



(a)音色辨識度　　(b)語音品質

圖 15　使用女性原始語句之聽測結果



(a)音色辨識度　　(b)語音品質

圖 16　使用男性原始語句之聽測結果

六、結論

雖然以離散倒頻譜來逼近頻譜包絡的觀念，多年以前就被提出了，但是實際實施時會面臨到三個問題，其一是求解穩定的頻譜包絡(無劇烈起伏)所對應的倒頻譜係數，這個問題已由前人解決，本論文則研究了另外二個問題，即”頻譜峰點挑選”和”頻率軸尺度轉換”的問題。關於頻譜峰點之挑選，我們應用 HNM 的觀念，先把頻譜分割成低頻

諧波和高頻雜音兩個部分，然後在低頻諧波部分依據所求出的基頻值去偵測諧波頂點，而在高頻雜音部分，則依據一般 cepstrum 平滑過的頻譜曲線去找出高過曲線的頻譜峰點。此外，關於頻率軸尺度的轉換，文獻上雖然提到 mel 尺度和 bark 尺度，但是我們從觀察逼近出的頻譜包絡曲線，發現 mel 尺度和 bark 尺度所得到的頻譜包絡仍然是不夠理想的，在中頻帶(3KHz~6KHz)常常會出現錯誤的頻譜包絡，因此我們在一番嘗試後，設計了一種頻率軸的尺度轉換公式，測試實驗顯示我們的轉換公式可明顯地降低頻譜包絡和頻譜峰點之間的逼近誤差。使用上述的解決方法，我們建構了一個頻譜包絡的估計架構。

此外，我們應用所提出的頻譜包絡估計之架構，去製做出一個語音轉換系統，該系統經由頻譜包絡估計、頻譜包絡伸縮、基頻移動、和信號重新合成等處理步驟，可把輸入語音信號的音色轉換成不同性別和年齡的其它音色。在信號重新合成步驟裡，我們採用了 HNM 信號模型，來分別合成諧波信號和雜訊信號，再作相加。爲了評估此系統的效能，我們進行了聽測實驗，由 13 位受測者的平均評分來看，我們系統的確可以有效地達成音色轉換之功能。根據這樣的音色轉換之表現，未來我們將會把本論文研究的頻譜包絡估計之架構，應用於特定語者之間的音色轉換的研究。

參考文獻

[1] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, Piscataway, NJ, 2000.

[2] D. Schwarz and X. Rodet, "Spectral e nvelope estim ation and representation for sound analysis-synthesis", *Int. Computer Music Conference*, Beijing, China, pp. 351-354, Oct. 1999.

[3] S. Im ai and Y . Abe, "Spectral enve lope extraction by im proved cepstral m ethod", Electron. and Commun. in Japan, Vol. 62-A, No. 4, pp. 10–17, 1979. (in Japanese)

[4] A. Robel and X. Rodet, "Ef ficient spectra l en velope es timation and its app lication to pitch shifting and envelope preservation", *Int. Conference on Digital Audio Effects*, Madrid, Spain, pp. 1-6, September 2005.

[5] T. Galas and X. Rodet, "An im proved cepstral method for deconvolut ion of source filter systems with discre te s pectra: Applica tion to m usical sou nd signals ", *Int. Computer Music Conference* (ICMC), Glasgow, Scotland, pp. 82-44, 1990.

[6] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation", *IEEE Signal Processing Letters*, Vol. 3, No. 4, pp. 100–102, 1996.

[7] 古鴻炎、張小芬、吳俊欣，"仿趙氏音高尺度之基週軌跡正規化方法及其應用"，第十六屆自然語言與語音處理研討會(ROCLING XVI)，台北，第325-334 頁, 2004。

[8] 王小川，語音訊號處理(修訂二版)，全華圖書公司，台北，2009。

[9] Y. S tylianou, "Modeling speech bas ed on harmonic p lus n oise m odels", in *Nonlinear Speech Modeling and Applications*, eds. G . Cholle t *et al*., S pringer-Verlag, Berlin, pp. 244-260, 2005.

[10] Y. Stylianou, *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification,* Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

[11] Y. Shiga and S. King, "Estim ating deta iled spectral envelope s using articulatory clustering", *Int. Conference on Spoken Language Processing* (ICSLP2004), Jeju, Korea, October 2004.

[12] F. R. Moore, *Elements of Computer Music*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[13] M. Dolson, "The phase vocoder: A tutorial", *Computer Music Journal*, Vol. 10, No. 4, pp. 14–27, 1986.

電腦輔助句子重組試題編製[1]

黃志斌　　劉昭麟　　郭韋狄　　孫瑛澤　　賴敏華
國立政治大學 資訊科學系
{g9614, chaolin, s9441, s9538, g9523}@cs.nccu.edu.tw

摘要

句子重組試題要求受測者把所給的一組詞彙組合成特定詞序的正確語句，是語文測驗常見的試題類型，可以檢驗受測者的句型和文法知識。我們試圖利用文法分析工具，協助出題教師編製句子重組試題。編製句子重組試題的第一個工作是適當地切割語句，以控制試題難度；一個好的剖析器就足以提供很好的服務。切割所得的詞彙集合，常常可以組合成與教師所欲學生回答的正確語句之外的合法語句。透過限制試題詞彙集的相對位置，我們可以限制學生的答案。我們的研究經驗顯示要自動協助出題教師預示所有可能的合法詞序卻是一件艱難的工作，而且這一研究問題與語法學有密切關係。本文詳述我們利用史丹佛剖析器，建構英文句子重組試題編製環境的經驗與實際系統。

關鍵詞：電腦輔助語文教學、文法學習輔助、句子重組練習、連結樹句法

1. 簡介

語文是人們溝通的基礎，爲了避免詞序的使用錯誤以及恰當的使用語言，本文將探討如何利用軟體技術建構出句子重組試題編製環境，以供學生練習語文的正確詞序之用。

句子重組試題是將句子打亂詞序後讓學生重新組合詞序的測驗試題，不過在重新組合詞序的時候有可能會遭遇到下面的問題。如「今天我被狗咬了。」與「I put a book on a table.」這兩句打亂詞序產生句子重組試題之後，學生除了排出原本的句子之外，也可以排出「今天狗被我咬了。」與「I put a table on a book.」在文法上合法但語意上卻有微妙差異的同字詞不同詞序組合。

目前在中文句子重組出題方面，香港宣道會葉紹蔭紀念小學提供的中文科網上學習網頁[1]裡有中文句子重組的練習題；在英文句子重組出題方面，台灣成德國小提供的英語檢測練習網頁[2]裡有句子重組的練習題。我們刻意地排出非這兩個系統預設的答案卻又是合理的句子來測試，結果發現此兩系統均無法辨別其正確性。另外，王昱鈞等學者[3]也曾提出過句子重組系統，該系統以語境(context)的方式提供了句子重組之線索來引導學生答出系統預設的答案，不過仍無法避免學生答出預設答案之外的合理句子，同時也無法辨別這些句子的正確性。考慮到建置所有同字詞不同詞序的答案之成本，本論文設法尋求其它方法，讓不同的學生都只能排出相同的答案。

建置句子重組試題編製環境第一個要面臨到的問題就是如何才能讓不同的學生都只能排出某些特定的答案。爲了稱呼上的方便，原始句子內的詞序如果經過調動，我們就

[1] 本篇論文爲符合論文集頁數限制而刪減爲 14 頁，限於篇幅不能在本文中全面交代相關細節。我們另外準備了一份較詳細的 25 頁版本之論文，公開在 http://www.cs.nccu.edu.tw/~chaolin/papers/rocling_huang.pdf

稱詞序調動後的句子爲***亂序句***(deranged sentences)。若亂序句仍然符合文法規範的話，這樣的亂序句就稱爲***重組句***(scrambled sentences)。Becker 等學者[4]是尋找重組句的先進，他們爲了解決在德語上使用「Tree Adjoining Grammars」(TAG)[5]分析重組句時的不足，將 TAG 擴展成「Multi-Component TAG」與「Free Order TAG」來分析與預測可能的重組句。我們打算將某些字固定在特定的詞序上，藉此來過濾掉其它合法的同字詞不同詞序之答案(也就是重組句)，而這些被固定在特定詞序上的字我們就稱爲***錨字***(anchors)。當一個句子在固定某些字的詞序之後，如果不同的學生只能排出一樣的答案的話，那麼我們的目的就達成了。

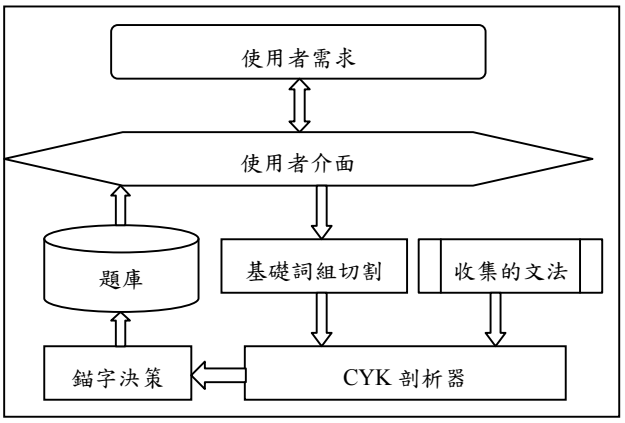


圖 1、編製環境架構圖

重組試題編製環境的第二個問題就是如何利用電腦輔助篩選出可能的重組句？我們將一句 n 個字的句子透過「基礎詞組」的幫助把這個句子切割成 m 個區塊($n \geq m$)，並藉由這 m 個區塊排列出 $m!$(m 階乘)個句子。然後，經由「史丹佛剖析器」(Stanford Parser)[6]，從中收集用來剖析(parse)的文法，接著將這些文法提供給利用動態規劃剖析演算法「Cocke-Younger-Kasami Algorithm」(簡稱 CYK 演算法)[7][8]所組建的剖析器，來對這 $m!$個句子篩選出可能的重組句，這些等待 CYK 剖析器篩選的句子我們就稱爲***候選句***。篩選的方式則是利用 CYK 剖析器，依據所給定的文法規則，來檢驗所輸入的候選句是否合乎文法；若候選句可以被完整剖析，就通過篩選，反之則被篩選掉。

圖 1 爲本論文之編製環境架構圖。當教師想要建置句子重組題目時，可以透過我們所提供的使用者介面輸入想要做爲題目的句子，這個被輸入的句子經過基礎詞組切割的流程後所排列產生出的句子，會和預先收集好的文法一起送入 CYK 剖析器篩選出重組句，接著透過錨字的決策固定住某些字，最後再存入題庫，供教師編輯或學生練習。

我們會在第二節說明基礎詞組的切割並呈現相關數據的分析，然後在第三節簡述我們收集到的文法和相關實驗的結果，接著在第四節討論錨字決策，再來我們在第五節展示編製環境提供給教師編輯與學生練習介面，最後第六節則是簡短的結語。

2. 切割基礎詞組與篩選重組句

我們在第一小節介紹基礎詞組切割的想法，然後在第二小節提出尋找基礎詞組與合併詞組的方式，在第三小節簡單介紹如何利用機率分數給亂序句排序，第四小節則是篩選門檻值與合併詞組的比較實驗。

2.1 基礎詞組與重組句的關聯

表 1 是假設教師以(1a)的句子輸入使用者介面做爲原始句所產生出的一個重組句範例，(1a)、(1b)以及(1c)這三句都是重組句。同時仔細觀察這三句，都擁有基礎詞組「good students」。我們企圖透過重組句中的基礎詞組，發展出找重組句時較好的方式。

最直覺的做法就是將一串 *n* 個字的英文句打亂詞序之後重新排列的 *n*!個句子透過 CYK 剖析器來篩選可能的重組句。但打亂詞序之後重新排列產生「10!」個句子，就表示剖析器就必須處理「10!」(3628800)個候選句，其總花費的時間或許會令出題教師感到不耐。

表 1、簡單的重組句範例

1a	I have never seen such good students.
1b	Such good students I have never seen.
1c	Never have I seen such good students.

爲解決這樣子的問題，我們觀察到了基礎詞組的一般性概念，即句子中最小的形容詞片語(adjective phrase，ADJP)、名詞片語(noun phrase，NP)以及介系詞片語(prepositional phrase，PP)的結構。在一個句子中可能只有一個基礎詞組，也可能有很多個基礎詞組，在表 1 中(1a)裡的基礎詞組只有「good students」。而在(1a)打亂詞序後，除了可以排成原本的句子(1a)之外，也可以排成(1b)或是排成(1c)包含「good students」的重組句。但在文法的規範下，重組句不會出現「Students good such seen never have I.」這種不包含「good students」的句子。也就是說，不同的重組句之間詞序的調動並不會把基礎詞組作進一步的切割。

透過這樣的想法，在表 1 的(1a)在排列成亂序句時，若將「good students」綁成一個區塊，CYK 剖析器處理的句子就可從原本的「7!」(5040)句減爲「6!」(720)句。藉由這樣子的方式就能省下不少時間。

2.2 詞組分析與合併

我們在 2.2.1 節中探討如何分析句子中的基礎詞組，2.2.2 節則討論當句子經過基礎詞組切割程序後的區塊數目仍然相當多時，使用合併詞組來處理的程序。

2.2.1 詞組分析

我們利用史丹佛剖析器分析試題編輯者輸入的句子，在取得機率分數最高的結構樹後，再根據該結構樹的內部結構來擷取基礎詞組。

圖 2 是用史丹佛剖析器分析表 1 中的(1a)所得到的機率最高之結構樹結構(詞性標記的部分請參照[9] Penn Treebank Tags)。爲了能夠精確地分析基礎詞組，我們設計了圖 3 的尋找基礎詞組的演算法，用來從結構樹中分析基礎詞組，並搭配圖 2 的結構樹來說明演算法的程序。首先根據圖 3 演算法的第 4 列在圖 2 的結構樹中尋找「ADJP」、「NP」、「PP」的蹤影，而在圖 2 我們找到了第四層的「PP」(假設節點「ROOT」爲第零層)，以及第二層和第五層的「NP」。但我們要找的「ADJP」、「NP」、或「PP」其子樹必須爲單字層級，如同圖 3 中第 5 列所述；然而第四層的「PP」其左子樹雖爲單字層級，右子樹卻爲「NP」的片語層級，也因此第四層的「PP」並不是我們想尋找的基礎

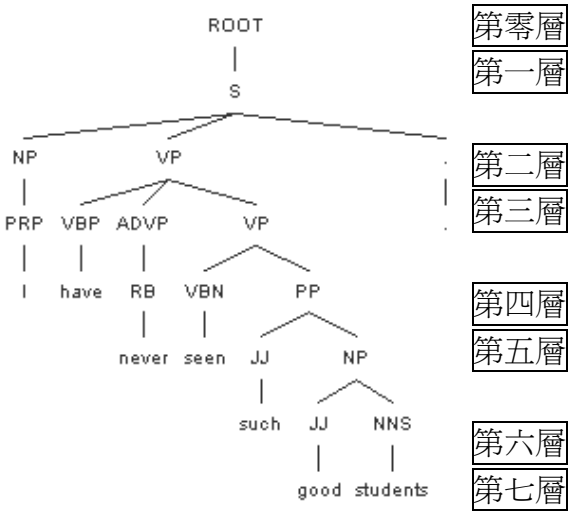


圖 2、史丹佛剖析器產生的結構樹範例

```
01  輸入：英文句 E={w1, ..., wn}及其結構樹 T。
02  輸出：英文句 E 中的所有基礎詞組。
03  程序：
04  在 T 中尋找所有的 ADJP、NP 以及 PP。
05  如果某個 ADJP、NP 或 PP 在 T 中的子節點為 E 中的單字層級，
06      那麼就將這個 ADJP、NP 或 PP 視為基礎詞組。
```

圖 3、尋找基礎詞組的演算法

詞組，第二層和第五層的「NP」才符合我們要找的基礎詞組。所以我們就將「I」和「good students」視爲基礎詞組，並把基礎詞組分別綁在一起產生亂序句。透過這種方式，我們就可以將表 1 中的(1a)切成「I」、「have」、「never」、「seen」、「such」以及「good students」這六個區塊來產生「6!」個候選句。

但表 1 中的(1a)是一個較簡單的句子，假若輸入一個 10 個字的句子，而且這個句子經過基礎詞組切割的程序之後產生了 8 個區塊，就代表我們還是得產生「8!」(40320)個候選句傳給 CYK 剖析器處理。剖析器要處理的句子越多，所花費的時間越長，同時剖析句長爲 n 的時間複雜度(computational complexity)爲$\theta(n^3)$。因此我們設計了兩個簡單的處理方法，試圖減少時間的花費。

2.2.2 合併詞組

若是句子經過基礎詞組切割後的區塊數仍很多的話，則用合併詞組來處理。我們從結構樹中最深的葉節點(leaf node)開始合併，直至亂序句降到理想的範圍內。

本論文合併詞組的方式是取由下而上(bottom up)的方式來實作。我們首先尋找結構樹中深度(depth)最深的詞組，接著再將該詞組向上拉高一層來合併，如果區塊還很多，就重複「尋找」和「合併」的動作，直至區塊數 m 降到理想的範圍內。我們用圖 2 的結構樹和圖 3 的演算法找到的基礎詞組，以及圖 4 合併詞組的演算法來說明這樣的方式。

```
01  輸入：英文句 E={w1, ..., wn}與其結構樹 T，英文句 E 的基礎詞組集合
        K={k1, ..., ky}，以及英文句 E 經過基礎詞組分析後切割出的
        區塊數 m，y≤m≤n。
02  輸出：集合 K
03  宣告：
04  令 K 中最深的詞組為kh。
05  令 near 為與kh同子樹且互為兄弟節點的相鄰詞組或英文字，|near|≤2。
06  隨機函數 R：隨機取出集合中的某個元素，每個元素被取出的機會相等。
07  參數：u，u 為理想的範圍。
08  程序：
09  當 m≥u 時，則重複以下步驟：
10      如果集合 K 中最深的詞組不只一個，
11          則透過 R 從這些同是最深的詞組集合中選一個設為kh。
12      在 K 中移去kh。
13      將kh與此kh的 near 合併成新詞組，
14          但若是此kh的|near|等於 2，
15              則透過 R 從此kh的 near 集合中選一個與kh合併成新詞組。
16      將合併後的新詞組加入 K 中。
17      區塊數 m 減 1。
```

圖 4、合併詞組的演算法

首先從圖 3 輸出的基礎詞組中尋找最深的詞組，在圖 2 的結構樹中即是(NP(JJ good)(NNS students))，接著將此詞組與其 *near* 合併來達到縮減區塊數 m，以求將 m 降到合理的範圍內。在圖 4 中，這個合理的範圍被參數化爲第 7 列的 u；而 *near* 則被描述在第 5 列，指的是與最深的詞組k_h同子樹且互爲兄弟節點的鄰接詞組或英文字，同時因爲與k_h鄰接的位置只有k_h的左相鄰與右相鄰兩個位置，所以第 5 列也明列出$|near|\leq 2$。我們透過第 9 列到第 17 列的迴圈來重複從詞組集合 K 中尋找最深的詞組k_h，以及將最深詞組k_h和此k_h的 *near* 合併，最後，當區塊數 m 降到理想目標，則將集合 K 做爲輸出。

在這個迴圈裡面，第 13 列會將最深詞組k_h及同子樹且互爲兄弟節點的鄰接詞組或英文字(即此k_h的 *near*)來合併，以在圖 2 中最深的詞組(NP(JJ good)(NNS students))來說，找到的 *near* 就是(JJ such)，於是就合併成(PP(JJ such) (NP(JJ good)(NNS students)))，並在第 16 列加入詞組集合 K 中，做爲下一次尋找最深詞組的候選人。這樣就可以把「such good students」綁在一起，如此一來變成「I」、「have」、「never」、「seen」以及「such good students」五個區塊，產生「5!」的候選句。

另外，有兩個問題的處理原則是必須額外交代的。第一個問題是當合併區塊時，若同時有多個基礎詞組在結構樹中都是最深的，那應該挑選何者來合併。第二個問題則是我們設定的區塊數並不是每一個結構樹都能找到剛好的切分點來符合我們的要求。

第一個問題可能沒有標準的答案，我們是透過一個隨機函數 R 來從多個最深的詞組中挑選一個。這個 R 被描述在圖 4 中的第 6 列，每個詞組被取出的機會均等。隨機的機制則在圖 4 中的第 10 到 11 列中說明。

而第二個問題目前也沒有標準答案，也是由隨機函數 R 來決定哪兩個相鄰的區塊要做合併的動作。以三個連續區塊 A、B 和 C 來說，我們先用解決第一個問題使用的隨機機制挑選其中一個最深的區塊，如果挑到 A 或 C 就沒有合併的困擾，因爲此時兩者的 *near* 均爲區塊 B(即$|near|$等於 1)，無論挑到哪個都只能和 B 合併。但若挑到區塊 B 的話，便有 A 和 C($|near|$等於 2)，這是又引出了隨機的機制，進入第 15 列用隨機函數 R 來選擇要合併的是哪個區塊。

我們在此試圖用((green)(and)(pink))這個小結構來說明合併成兩個區塊時，上述兩個問題的實際情況。首先「green」、「and」和「pink」都是最深的詞組，那麼我們就隨機選擇一個來合併；若是隨機選到「green」或「pink」都只能和「and」合併產生「『green and』、『pink』」或「『green』、『and pink』」；而若是隨機選到「and」的話則可再隨機選擇與「green」或「pink」合併，產生「『green and』、『pink』」或「『green』、『and pink』」。

2.3 利用機率分數篩選重組句

當史丹佛剖析器處於「probabilistic context-free grammar」(簡稱 PCFG；也稱爲 stochastic context-free grammar，SCFG)[7][8]的模組下，當輸入一個句子後，就可以操作參數來產生出許多該句的可能結構樹結構，同時每個結構樹都帶有一個機率分數，該機率分數代表著句子被剖析爲對應的結構樹之機率有多少。

我們利用結構樹機率分數越高，合法句的機會也越高的想像，將切割後的區塊所產生的亂序句輸入史丹佛剖析器，擷取每句亂序句中最高的機率分數，同時依此機率分數將亂序句排名。相關的實驗我們將在 2.4.1 節中提出。

2.4 篩選門檻值與合併詞組的實驗

我們在 2.4 節中進行比較實驗。2.4.1 節探討機率分數排名的門檻值，2.4.2 節分析合併詞組後產生的亂序句是否能涵蓋重組句，2.4.3 節則是篩選門檻值與合併詞組的綜合實驗。

2.4.1 篩選門檻值的比較實驗

此節是用基礎詞組切割的區塊產生出所有的亂序句之後，利用史丹佛剖析器所給的機率分數，把亂序句排名次然後再用門檻取出特定數量的亂序句做爲候選句，藉此降低 CYK 剖析器花費在篩選的時間。

表 2、實驗資料區塊切分

	4 區塊	5 區塊	6 區塊	7 區塊	8 區塊
7 字	7 句	9 句	13 句	7 句	0 句
8 字	1 句	9 句	7 句	11 句	7 句
9 字	0 句	1 句	10 句	11 句	7 句

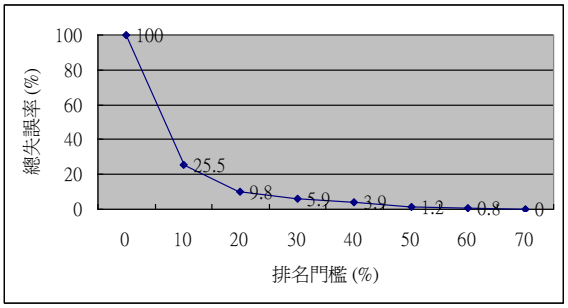


圖 5、排名門檻總失誤率的實驗數據

我們進行了一個排名門檻的實驗。實驗對象是 100 個從網路上隨機擷取的句子，當中包含了 36 個 7 字句、35 個 8 字句、以及 29 個 9 字句，並對這些句子用基礎詞組的方式切割，形成 4、5、6、7、8 五種區塊類別，如表 2 所示。表 2 裡第二列第四行的「13 句」表示「有 13 句 7 個字的句子被切分成 6 個區塊」，其他依此類推。我們也用人工的方式盡可能地尋找這 100 句的重組句，並對重組句組用史丹佛剖析器的機率分數做排名百分比後，進行門檻值的分析。

圖 5 即爲排名門檻的實驗數據，橫軸代表著我們取句子產生的亂序句中「前 10%」、…、「前 70%」來篩選的門檻值，縱軸則是以「255 句重組句」爲分母，「未在門檻內的重組句數」爲分子的總失誤率。這 100 個重組句組一共有 255 句，因此當門檻值爲「0%」，在圖 5 中會對應到有 255 個重組句不在我們所取的門檻內；而當我們的門檻值設爲「70%」的時候，則可以將 255 句全部囊括進來。

雖然取「70%」爲門檻值是個穩定作法，但 7 區塊句子有 5040 個亂數句，取「70%」的句數(3528 句)對於 CYK 剖析器來說是一個不小的篩選數量。或許可以嘗試用較低的門檻值來篩選。若以總失誤率來看的話，將門檻值設爲「20%」失誤率有 9.8%，設在「30%」、「40%」的話失誤率則有 5.9%、3.9%。

雖然將門檻值由「70%」改爲較低的門檻值會有失誤率的風險存在，但卻可以爭取到計算時間上的效益。也就是說門檻值的選定並沒有絕對的標準，而是可以有彈性變化的。

2.4.2 合併詞組的比較實驗

我們利用表 2 中第四列第六行的那 7 組重組句組(參見附錄二)做爲實驗語料，分析合併詞組後產生的亂序句是否能涵蓋重組句。

我們將 7 組重組句組裡的 7 句原始句使用圖 4 提到的演算法從 8 個區塊分別合併成 7、6、5 和 4 個區塊的句子，接著再以這些句子各自產生亂序句。實驗目的在分析這些

表 3、合併詞組實驗

	合併區塊數			
	7	6	5	4
1.1	T	T	T	T
1.2	T	T	T	F
2.1	T	T	T	T
2.2	T	F	F	F
2.3	T	T	T	T
2.4	T	F	F	F
3.1	T	T	T	T
3.2	T	T	T	T
3.3	T	F	F	F
4.1	T	T	T	T
4.2	T	T	T	F
5.1	T	T	T	T
5.2	T	T	T	T
6.1	T	T	T	T
6.2	T	T	T	F
6.3	T	T	T	T
7.1	T	T	T	T
7.2	T	T	F	F
7.3	T	F	F	F

表 4、綜合實驗數據

	合併區塊數							
	7		6		5		4	
	排名	排名百分比	排名	排名百分比	排名	排名百分比	排名	排名百分比
1.1	468	9.29	69	9.58	16	13.33	2	8.33
1.2	208	4.13	37	5.14	8	6.67	F	F
2.1	19	0.38	2	0.28	1	0.83	1	4.17
2.2	7	0.14	F	F	F	F	F	F
2.3	486	9.64	78	10.83	15	12.5	6	25
2.4	487	9.66	F	F	F	F	F	F
3.1	2	0.04	2	0.28	2	1.67	2	8.33
3.2	3	0.06	3	0.42	3	2.5	3	12.5
3.3	965	19.15	F	F	F	F	F	F
4.1	3	0.06	1	0.14	1	0.83	1	4.17
4.2	12	0.24	10	1.39	2	1.67	F	F
5.1	3	0.06	2	0.28	1	0.83	1	4.17
5.2	64	1.27	40	5.56	16	13.33	6	25
6.1	312	6.19	79	10.97	37	30.83	6	25
6.2	163	3.23	37	5.14	23	19.17	F	F
6.3	1272	25.24	248	34.45	78	65	15	62.5
7.1	97	1.92	27	3.75	6	5	1	4.17
7.2	98	1.94	28	3.89	F	F	F	F
7.3	605	12.00	F	F	F	F	F	F

各自產生的亂序句是否能涵蓋重組句，如果包含指定的重組句的話，則標記「T」；反之則標記「F」。實驗數據如表 3，最上面的「7」、「6」、「5」和「4」即爲將原始句合併成 7、6、5 和 4 個區塊數來產生亂序句。

從表 3 中可以看到，當區塊數目越合併越少的時候，所產生的亂序句能夠涵蓋的重組句也就越來越少。因此當我們在圖 4 中考慮將區塊數 m 降到合理的範圍 u 時，也和 2.4.1 節的門檻值一樣，是可以有彈性變化的。

2.4.3 篩選門檻值與合併詞組的綜合實驗

我們把表 3 中 7、6、5 和 4 個區塊各自產生的亂序句，再用史丹佛剖析器產生這些亂序句的機率排名，並使用 2.4.1 節討論的門檻值來分析是否能涵蓋重組句。

表 4 即爲篩選門檻值與合併詞組的綜合實驗數據。表 4 中「合併區塊數」下方的「7」、「6」、「5」和「4」與表 3 的意義相同；「排名」表示該重組句在所屬的亂序句中的機率分數排名；「排名百分比」則是該重組句在所屬的機率分數排名中的百分比。由於表 4 是承接表 3 所進行的實驗，因此表 3 裡「F」代表沒有被涵蓋在亂序句中的重組句，在表 4 裡也不會有機率分數，所以也用「F」表示無法排名。

如果機率分數的門檻值設在「30%」與「40%」的話，這 19 個重組句在「7」、「6」、「5」和「4」個區塊數所對應的總失誤率(表 4 標記爲「F」的也算失誤)爲 0%、26.3%、36.8%和 47.4%，以及 0%、21.1%、31.6%和 47.4%。這樣子的結果顯示出區塊數與門檻值稍高，總失誤率就能夠稍微降低，但要付出的代價則是 CYK 剖析器從候選句中篩選出可能的重組句之花費時間。

3. 文法分析

由於 CYK 剖析器需要文法來篩選亂序句，但我們缺乏現成的文法規則資料，於是想到了利用語料來獲取文法規則。我們在第一小節說明如何從史丹佛剖析器收集文法，接著在第二小節提報相關實驗的結果。

3.1 文法萃取

我們隨機從網路中收集七千多句國中英文課本的相關例句做爲語料，並將這些語料傳入史丹佛剖析器做結構樹的分析。接著從分析出來的結構樹中統計出現的文法規則，再將這些規則做爲 CYK 剖析器的文法輸入。

當我們在史丹佛剖析器中輸入表 1 中的(1a)時，就可以獲得圖 2 這個最高機率的結構樹。這個結構樹除了葉節點之外的內部節點分支其實都是一組文法規則，如第二層的內部節點「VP」擁有「VBP」、「ADVP」以及「VP」三個子節點，那麼就表示在這邊使用到了「VP → VBP ADVP VP」這樣子的文法規則來進行剖析。

我們對這七千多句的每一句的結構樹都做這樣子的分析，並做出一個統計資料，如圖 6。圖 6 的橫軸是我們統計相同文法規則出現的次數之後取其對數的值，縱軸則是出現一樣次數的不同文法規則之條數。

在這個數據中總共有 985 條文法規則，出現次數最多的前四名從第一名開始依序是 4450 次的「NP → PRP」、4329 次的「S → NP VP」、3079 次的「NP → DT NN」以及 2415 次的「PP → IN NP」。

第一名的文法說明了其相關例句不外乎圍繞在與人稱代名詞相關的句子，如「I love you.」等等之類的句子；第二名的文法則點出了這些例句是以直述句居多；第三、第四名則是告訴了我們簡單的名詞片語以及介系詞片語也是常出現於這些例句，如「an apple」、「in the morning」等等。另外我也可從圖 6 中得知大多數的規則都出現很少次，實際上真正常被拿來做剖析的文法只有 985 條中的一小部份而已。

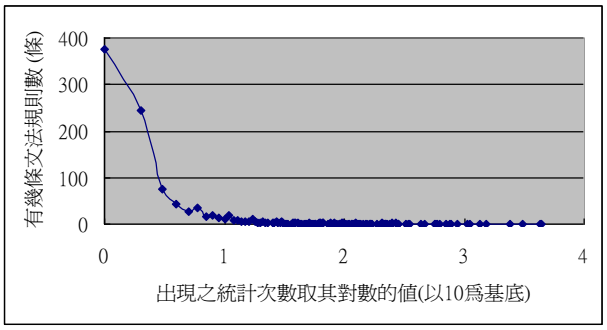


圖 6、文法規則的統計資料

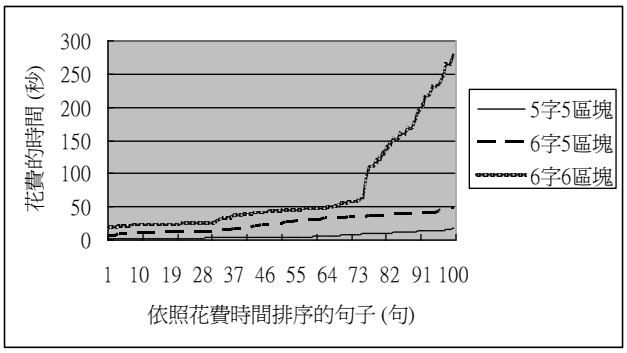


圖 7、所費時間之數據

我們在一部使用 2GB 的 RAM 和 Windows XP 的 Pentium 4 2.68G 機器上做了一個簡單的時間實驗，該實驗是在 CYK 剖析器使用上述的 985 條文法規則下，在網路上隨機找了 5 個字、6 個字的 100 個句子，並將 5 個字切成 5 個區塊、6 個字切成 5 和 6 兩種區塊產生亂序句，再把它們匯入 CYK 剖析器記錄所需時間。實驗目的在探討 CYK 對於同句數不同字數及同字數不同句數的篩選時間。圖 7 是這 200 句資料跑的三組實驗所得的花費時間之內部情況，縱軸是 CYK 篩選產生的亂序句時所花的時間，橫軸則是依照篩選每句產生的亂序句總花費時間由小到大排序。

我們從圖 7 的記錄中可以計算出 CYK 剖析器篩選 5 字 5 區塊、6 字 5 區塊以及 6 字 6 區塊的平均時間分別爲 6.4 秒、24.3 秒及 72.3 秒。雖然同樣都是切 5 區塊，但 6 個字的平均篩選時間是高於 5 個字的，可得知字數越多則所需時間越多。再來看到 6 字 5 區塊、6 字 6 區塊的篩選時間，可得知句數越多則所需時間越多。

值得注意的是，6 字 6 區塊的折線在最後的部份開始急速上升，這個奇怪的部份我們試著用以下的例子來說明。假設文法集中有很多規則，當中有四條規則爲「A → a c」、「B → a c」、「C → a c」、「D → a c」，同時我們有 3 個字的原始句「a b c」。如果將「a b c」切成 2 個區塊來產生候選句，無論什麼情形都不會引入上述的四條規則來剖析。然而當我們將「a b c」切成「a」「b」「c」這 3 個區塊來產生候選句時，就會出現「bac」和「acb」的組合，此時就會引入上述四條規則，所花費的時間也就可能以倍數成長了。我們推測圖 7 中 6 字 6 區塊的折線在最後急速上升的部份和上面例子的 3 字 3 區塊情況類似，是由文法集本身的規則所產生的時間倍數成長，如在我們所萃取的文法集當中的「NP VP」就有「SQ → NP VP」、「NP → NP VP」與「S → NP VP」這三條規則。

由以上我們對時間實驗數據的討論可知，句子字數越多、切分的區塊越多，那麼所花時間也會越長，同時與文法集的規則數也有關。

3.2 文法集分析

我們將 985 條文法規則分別取不同的範圍進行實驗，藉此比較不同範圍的文法集對 CYK 剖析器判斷重組句時的影響。

我們可以試著去想像，當文法的數目相當龐大複雜的時候，即使是亂組的亂序句，都有可能找到剛剛好的文法來剖析(這樣的問題在計算語言學裡面稱爲「overgeneration」[10])，使得 CYK 剖析器誤判爲重組句；又或者，當文法數目相當稀少的時候，即使是教師透過使用者介面輸入的原始句，都有可能找不到合適的文法來剖析，該原始句也會被 CYK 剖析器誤判成不合法。前述第一種情況，將導致 CYK 剖析器在過濾可能的重組句時太過寬鬆，使得我們必須產生較多的錨字(錨字的選擇問題將會在第四節仔細說明)，最後影響到的就是句子重組試題的效果。試想，如果一個句子重組試題有 *n* 個字，而我們固定住了 *n*-1 個字，這樣子的句子重組試題完全沒有意義。而第二種情況，則將導致 CYK 剖析器在過濾可能的重組句時太過嚴苛，使得最後甚至沒有任何亂序句能通過篩選。

我們對於這些通過篩選的亂序句中是否含有重組句，以及在面對真實生活中的句子時文法是否能夠篩選出可能的重組句感到興趣。於是從生活週遭的對話，以及網路文章中隨機挑選出來九個句子做爲測試語料，同時也用人工的方式找出每一個句子可能的重組句(參見附錄一)。實驗數據如表 5。表 5 中的「n」、「m」代表這個句子是 *n* 個字的句

表 5、測試通過篩選的亂序句中是否含有重組句的實驗數據

	n	m	前 150 條			前 300 條			前 450 條			全部 985 條		
			通過句數		包含在內	通過句數		包含在內	通過句數		包含在內	通過句數		包含在內
1.1	8	6	32	4%	T	78	11%	T	96	13%	T	240	33%	T
1.2					T			T			T			T
2.1	6	6	120	17%	T	120	17%	T	120	17%	T	120	17%	T
2.2					F			F			F			F
2.3					T			T			T			T
2.4					T			T			T			T
3.1	6	5	48	40%	T	48	40%	T	48	40%	T	72	60%	T
3.2					T			T			T			T
3.3					T			T			T			T
3.4					T			T			T			T
4.1	6	4	6	25%	T	6	25%	T	6	25%	T	6	25%	T
4.2					F			F			F			F
4.3					F			F			F			F
5.1	7	6	120	17%	T	120	17%	T	120	17%	T	120	17%	T
5.2					F			F			F			F
5.3					F			F			F			F
6.1	7	5	0	0%	F	0	0%	F	0	0%	F	0	0%	F
6.2					F			F			F			F
7.1	7	6	192	27%	T	192	27%	T	240	33%	T	240	33%	T
7.2					F			F			F			F
8.1	8	5	12	10%	T	18	15%	T	24	20%	T	24	20%	T
8.2					F			F			F			F
9.1	6	6	33	5%	T	120	17%	T	120	17%	T	120	17%	T
9.2					F			F			F			F

子且以基礎詞組方式切割成 m 個區塊來產生亂序句。左邊最上面的「1.1」與「1.2」代表除了原始句「1.1」之外，我們還用人工的方式找到了另外一句重組句「1.2」；同理，「2.1」、「2.2」、「2.3」以及「2.4」依此類推。而「前 150 條」…「全部 985 條」代表的是文法規則集中出現次數最高的前多少條文法規則。「通過句數」代表該限制的文法規則集下原始句產生的 $m!$ 個亂序句中到底有多少個能通過 CYK 剖析器的篩選；其下的左欄位是通過句數的實際句數，右欄位則是實際句數在 $m!$ 個亂序句當中所佔的比例。「包含在內」代表這些通過 CYK 剖析器篩選的亂序句中，是否包含了這一個重組句；若是包含在內，則標記「T」，反之則標記爲「F」。透過表 5 中的數據，我們做出了以下的推論與探討。

在這九組中，有某些句子即使放鬆了限制，通過 CYK 剖析器篩選的亂序句也不會增加。如第 8 組的重組句組，該句組的文法規則限制由「前 150 條」放鬆到「前 450 條」時，通過篩選的亂序句的確是增加的；但是再放鬆到「全部 985 條」時，數字卻維持在「前 450 條」時的 24 句。此現象驗證了我們在討論圖 6 時所論述的「常被拿來做剖析的文法，實際上只有 985 條中的一小部份而已」。

接著我們在表 5 中看到了一個特例，即第 6 句組。這句組相當奇特，不僅僅是 6.2 的重組句，連做爲原始句的 6.1 都沒辦法通過篩選。原始資料顯示，6.1 並非是一個簡單的人稱代名詞(如：「I」、「you」……等等)做爲主詞的句子。而國中程度的英文句大部份都是以簡單的名詞片語或人稱代名詞爲主詞的句子較多，這點也可以從被 CYK 篩選過後的原始資料中發現，以人稱代名詞做爲主詞開頭的句子，即使是不合法的亂序句，就算將文法規則限制在「前 150 條」，這些不合法的句子也可以通過篩選。所以當我們將超過國中英語課本難度之外的句子輸入時，就會被 CYK 剖析器擋下來而無法通過篩選。

4. 錨字決策

錨字決策即是在最少錨字的限制下，讓通過 CYK 剖析器篩選的亂序句中只有某些句子能夠符合該錨字限制的決策，並藉此滿足教師只要學生排出特定答案的要求。本節提出一個簡單的演算法來說明處理一道試題只有一個唯一答案的情形。

從表 1 的三個英文句來說，如果我們限制第二個字只能填入「have」的話，那麼就還要再限制其它詞序的字才能把這三句過濾到只剩下一句，因爲(1a)和(1c)會共同來競爭「第二個字只能填入『have』」這個條件；但是如果我們一開始就決定第三個字只能填入「never」的話，那麼能夠被排出來的句子就只剩(1a)這一句，同時我們的目的也就達成了。

由於教師從使用者介面輸入句子時，是希望獲得該句的句子重組試題，所以我們將教師輸入的原始句做爲目標句來產生亂序句。圖 8 是我們用來決定錨字的演算法。從CYK剖析器得到通過篩選的亂序句之後，就可以從這些亂序句中去統計原始句裡的字在原來位置出現的句數；爲了往後說明上的方便，我們稱這樣子得到的句數爲***對位數***。圖 8 中第 9 列就選取了對位數最少的非錨字(即不是錨字的字)固定詞序做爲錨字，並在集合 S 中留下其它和原始句在相同位置上有錨字的亂序句(圖 8 中第 13 列)，同時在集合 S 中刪去其他的亂序句。如果$|S|$大於 1 的話，那麼就重複執行第 8 列到第 13 列的步驟，直到這些亂序句被過濾後只剩下原始句。

如果教師從使用者介面輸入表 1 中的(1a)，同時假設通過 CYK 剖析器篩選的亂序句剛好也只有(1a)、(1b)和(1c)這三句的話，那麼此時圖 8 中$|S|$即等於 3，依照演算法的步

```
01  輸入：經過 CYK 剖析器篩選後的亂序句集合 S={s1, …, se}，
        e 為亂序句之數目。
02  輸出：集合 A
03  宣告：
04  隨機函數 R：隨機取出集合中的某個元素，每個元素被取出的機會相等。
05  輸出的錨字集合 A = {}
06  程序：
07  當|S|大於 1 時，則重複以下步驟：
08      對所有亂序句中的每一句計算原始句裡每個字的對位數。
09      選擇原始句中擁有最小對位數的非錨字作為錨字。
10          但若是擁有最小對位數的非錨字不只一個，
11              則用隨機函數 R 從這些擁有最小對位數的非錨字所成的
                集合中挑選一個做為錨字。
12      將錨字加入集合 A。
13      在 S 中留下擁有錨字的亂序句，刪除其他的亂序句。
```

圖 8、決定錨字的演算法

驟計算，得到原始句中「I」的對位數爲 1，「have」的對位數爲 2，「never」的對位數爲 1，「seen」的對位數爲 2，「such」的對位數爲 2，「good」的對位數爲 2，「students」的對位數爲2。然後依據程序在擁有最小對位數的非錨字中隨機選取一個字(假設選到「I」)，並把原始句第一個字「I」設爲錨字。而當「I」被取爲錨字之後，亂序句就只剩下原始句(1a)，此時$|S|$等於 1，程序就結束了。所以我們得到了錨字「I」的輸出。

此外，當我們尋找擁有最小對位數的非錨字時，如果擁有最小對位數的非錨字不只一個的話，那麼演算法將會在這些同是擁有最小對位數的非錨字中隨機選取一個字做爲錨字。這個隨機的機制是透過一個被描述在圖 8 中第 4 列的隨機函數 R 來完成，相關的程序則在圖 8 中第 10 到 11 列。實際的例子則可以看到上述我們假設選到「I」做爲錨字時，其實「never」也是錨字的候選人之一，因爲這兩個字的對位數同時都是最小的 1。

5. 使用者介面

我們分別爲教師、學生製作了不同的使用者介面。教師介面爲試題編輯主選單，選單中可以建立新的題目，或編輯、刪除已有題目。而學生介面爲試題練習主選單，選單中可以選擇已經建立好的題目來練習。

5.1 編輯介面

圖 9 是試題編製環境的實際編輯介面，第一個和第四個白色的文字編輯欄位允許教師填入語境線索以提供給學生做爲重組句子時的參考，王昱鈞等學者[3]也認爲這種語境誘答設計的想法是有其價值的。第二個和第三個欄位則是填入題目的內容以及要測驗的句子。

在圖 9 的最下方有三個功能性按鍵，分別是「自動完成編輯」、「手動完成編輯」、以及「放棄編輯」。當教師把四個欄位都填完後，可透過「自動完成編輯」讓此編製環境透過前面章節介紹的程序自動地選取錨字，使得學生在練習時只能排出測驗的句子；也可透過「手動完成編輯」，讓教師自己從通過 CYK 剖析器篩選的亂序句中挑選滿意的重組句做爲答案，以保持學生答題時的自由度。

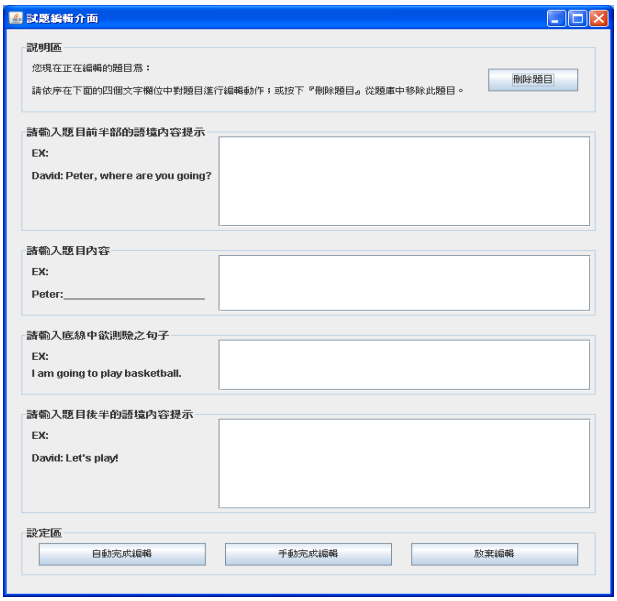


圖 9、編輯介面

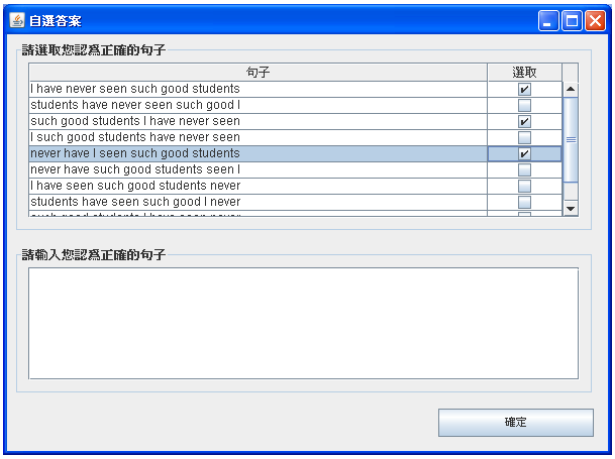


圖 10、挑選答案介面

圖 10 即爲按下「手動完成編輯」後的挑選介面，教師可在介面中選取想要的答案，也可以直接在圖 10 下方白色的文字編輯欄位中輸入答案。圖 9 最右上角的按鍵是「刪除題目」，可以用這個按鍵將編輯介面中的題目從題庫中刪除。圖 9 最右下角是「放棄編輯」，按下此鍵即不做任何更動並返回試題編輯主選單。

5.2 練習介面

當我們以學生的身分登入後，就會進入試題練習主選單。圖 11 的學生練習介面分爲最上面的「說明區」、中間的「作答區」和最下面的「解答區」三個主區塊。說明區是負責學生導覽的工作，藉由上面文字的說明來引導學生操作，同時該區還有一個「開始測驗」按鍵，按下該鍵後介面就會出現題目。

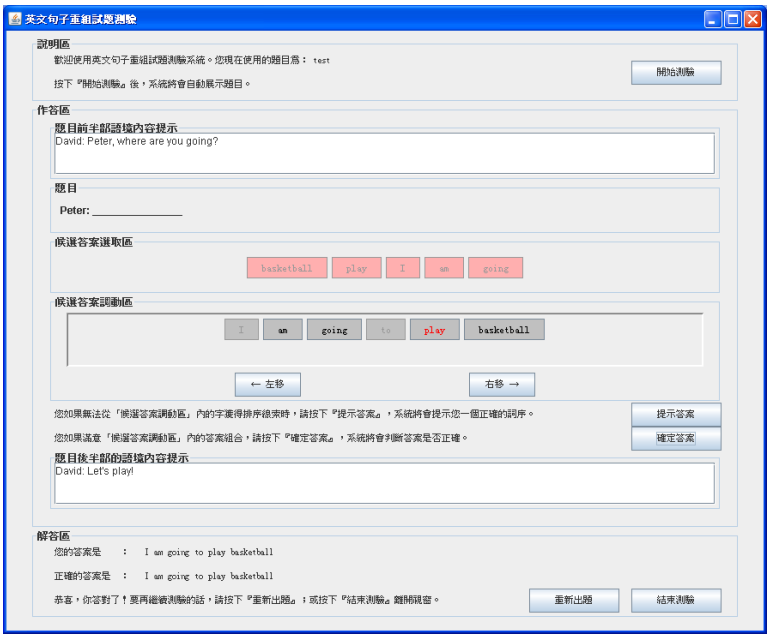


圖 11、練習介面

作答區中則包含了五個子區塊。教師之前在圖 9 的第一個和第四個欄位所填入的語境提示內容，就會出現在圖 11 作答區裡的第一個和第五個子區塊。我們希望透過前後文語境的提示，能夠帶給學生多一些解題的直覺。第二個子區塊則是放置題目的位置。要測驗的句子則在隨機打亂後放置在第三個子區塊中等著學生來挑選；錨字則是以淡灰色的字體顏色出現在第四個子區塊裡，同時無法透過「左移」或「右移」按鍵來移動錨字。

當學生從第三個子區塊中挑選某個字之後，該字會出現在第四個子區塊裡，並可用「左移」或「右移」按鍵來改變位置。我們在作答區中還設置了「提示答案」按鍵，若是學生對於題目不知從何下手，按下此鍵後介面就會在原始句裡面隨機挑選一個正確詞序的字來提示學生。最後的解答區則是比對題目的答案和學生排出的答案是否一致的地方，學生在這裡可以按下「結束測驗」返回試題練習主選單，或是按下「重新出題」繼續練習句子重組試題。

6. 結語

本篇論文報告了我們如何建構一個電腦輔助句子重組試題編製環境來幫助語言學習者掌握使用語言的能力。我們從產生可能的重組句出發，並嘗試用錨字的方式來解決以人工方式建置所有重組句成本過高的問題。透過相關的實驗數據分析，本編製環境的文法規則在處理國中程度的初學者入門句子時可能會稍微好些。

我們在實驗中看到了本編製環境的文法規則處理進階句子時所面臨的問題，因此未來會逐漸試著在國中英文課本相關的語料之外，收集其它較進階、或是現實生活中報紙

雜誌內文章的句子，來改善本編製環境文法規則的情況；又或者可以從剖析器的正確性著手，來探討文法規則剖析時的相關情況。同時未來在時間與精力的許可下，也會將本編製環境提供給教師與學生使用，並以問卷收集相關的數據來評估本編製環境的輔助效果。

致謝

本研究承蒙國科會研究計畫 NSC-97-2221-E-004-007-MY2 的部分補助謹此致謝。我們感謝匿名評審對於本文初稿的各項指正與指導，雖然我們已經在從事相關的部分研究議題，不過限於篇幅因此不能在本文中全面交代相關細節。

參考文獻

[1] 香港宣道會葉紹蔭紀念小學。中文科網上學習網頁。網址：http://www.casyymps.edu.hk/IT/Internet-Learning/chinese/chinese.htm (最後瀏覽日期爲 2009/01/04)。

[2] 台灣成德國小。英語檢測練習網頁。網址：http://w3.ctps.tp.edu.tw/today/teacher/hotexam/index.htm (最後瀏覽日期爲 2009/01/04)。

[3] M.-S. Lu, Y.-C. Wang, J.-H. Lin, C.-L. Liu, Z.-M. Gao, and C.-Y. Chang. Supporting the translation and authoring of test items with techniques of natural language processing, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 12(3), 234-242, 2008.

[4] T. Becker, A. K. Joshi, and O. Rambow. Long distance scrambling and Tree Adjoining Grammars, *Proceedings of the Fifth Conference on European chapter of the Association for Computational Linguistics*, 21-26, 1991.

[5] A. K. Joshi. An Introduction to Tree Adjoining Grammars, *Mathematics of Language*, 1987.

[6] The Stanford Natural Language Processing Group. The Stanford Parser: A statistical parser. http://nlp.stanford.edu/software/lex-parser.shtml (Last visited on 2009/05/10).

[7] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, the MIT Press, 1999.

[8] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2000.

[9] Penn Treebank Tags. http://bulba.sdsu.edu/jeanette/thesis/PennTags.html (Last visited on 2009/08/9).

[10] D. Lin. Principle-based parsing without overgeneration, *Proceedings of the Thirty-First Annual Meeting on Association for Computational Linguistics*, 112-120, 1993.

附錄一

1.1 She can neither sing well nor dance beautifully.
1.2 She can neither dance beautifully nor sing well.
2.1 They are sometimes late for work.
2.2 Sometimes they are late for work.
2.3 They are late for work sometimes.
2.4 They sometimes are late for work.
3.1 Only I saw the cake yesterday.
3.2 I only saw the cake yesterday.
3.3 I saw only the cake yesterday.
3.4 I saw the cake only yesterday
4.1 Solemnly the minister addressed her congregation.
4.2 The minister solemnly addressed her congregation.
4.3 The minister addressed her congregation solemnly.
5.1 I have never seen such good students.
5.2 Such good students I have never seen.
5.3 Never have I seen such good students.
6.1 One of my favorite hobbies is reading.
6.2 Reading is one of my favorite hobbies.
7.1 He goes to the library every Sunday.
7.2 Every Sunday he goes to the library.
8.1 The hotel is next to a movie theater.
8.2 A movie theater is next to the hotel.
9.1 I can agree in neither case.
9.2 In neither case can I agree

附錄二

1.1 Fashion goes hand in hand with compassion for life.
1.2 Compassion for life goes hand in hand with fashion.
2.1 Girls are born with more sensitive hearing than boys.
2.2 Boys are born with more sensitive hearing than girls.
2.3 Born with more sensitive hearing than boys are girls.
2.4 Born with more sensitive hearing than girls are boys.
3.1 Boys and girls should be educated in different ways.
3.2 Girls and boys should be educated in different ways.
3.3 In different ways should boys and girls be educated.
4.1 Neither he nor I want to attend the meeting.
4.2 Neither I nor he wants to attend the meeting.
5.1 He finally passed the exam because he studied hard.
5.2 Because he studied hard he finally passed the exam.
6.1 Here is some good food for you to try.
6.2 Here some good food is for you to try.
6.3 Here some good food for you to try is.
7.1 A brown and white dog is at your doorsteps.
7.2 A white and brown dog is at your doorsteps.
7.3 At your doorsteps is a brown and white dog

主題語言模型於大詞彙連續語音辨識之研究

On the Use of Topic Models for Large-Vocabulary Continuous Speech Recognition

陳冠宇　Kuan-Yu Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
696470203@ntnu.edu.tw

陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@ntnu.edu.tw

摘要

本論文研究使用主題資訊之語言模型(Language Model)。當語言模型用於大詞彙連續語音辨識時，其主要的任務是藉由已解碼歷史詞序列資訊來預測下一個候選詞出現的可能性。傳統的 *N* 連(*N*-gram)語言模型容易受限於模型參數過多的問題，僅能用來擷取短距離的詞彙接連資訊，並不能考慮完整的歷史詞序列之語意資訊。因此，近十幾年來許多研究學者陸續提出各式主題模型(Topic Model)，包括討論文件與詞之關係的機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)和潛藏狄利克里分配(Latent Dirichlet Allocation, LDA)，以及討論詞虛擬文件與詞關係的詞主題模型(Word Topic Model, WTM)。這些模型主要都是透過一組潛藏的主題機率分布來描述文件與詞、或者詞虛擬文件與詞之間的關係，用以擷取出歷史詞序列長距離的潛藏語意資訊。本論文提出一種新的主題模型，稱之爲詞相鄰模型(Word Vicinity Model, WVM)，它直接地基於語言中詞與詞相互關聯資訊以建構一個機率式的潛藏主題空間，並且透過線性模型結合的方式建立歷史詞序列之主題模型來預測下一個候選詞出現的可能性，藉此輔助傳統 *N* 連語言模型。實驗結果顯示本論文所提出的詞相鄰模型不僅相較大部分主題模型具有較低的模型參數量，同時能對於僅使用三連語言模型的基礎大詞彙連續語音辨識系統也有相當程度的語音辨識率提升。

關鍵詞：主題模型、機率式潛藏語意分析、潛藏狄利克里分配、詞主題模型、詞相鄰模型、大詞彙連續語音辨識。

一、緒論

語言是人與人之間最自然且有效率的溝通方式，不需透過其他的手勢或是動作，就可以讓對方了解我們想要表達的意思。正因爲如此，長久以來我們希望能讓機器聽懂人類的語言、直接與人類對話溝通，開啓了語音辨識的研究。在進行語音辨識時，我們以人類發聲的特性以及考量人耳聽覺感知爲基礎，將數位語音訊號轉換成易於電腦處理的聲學特徵向量(Acoustic Feature Vector)序列。接著，利用機率模型對於所收集到的訓練語音聲學特徵向量建立起聲學模型(Acoustic Model)藉此在測試階段比對測試語句聲之學特徵向量序列，判斷語句中所有可能的音素或詞段落。最後，使用語言模型(Language Model)來估測自然語言中每一個詞彙基於不同上下文之所可能出現的機率分布，用以解決聲學模型的混淆、限制辨識的搜尋空間和評估各個候選詞序列在自然語言中的合理性，因而輸出最有可能之候選詞序列。

當語言模型實際運用於語音辨識時，最主要的方式是從已解碼之歷史詞序列擷取短距離的詞彙接連資訊、或是長距離的語意資訊，據此預測下一個候選詞出現的可能性。在傳統統計式語言模型中，*N*連(*N*-gram)語言模型[1]是最爲人所知且廣泛地運用於各種自然語言處理領域。*N*連語言模型嘗試紀錄詞與詞之間同時出現的關係，估測每一個詞在其先前緊鄰 *N*-1 個詞已知的情況下出現的條件機率，並以多項式(Multinomial)分布表示之。但由於詞與詞序列有相當多種排列組合，致使 *N*連語言模型的參數量相當可觀。*N*連語言模型常因訓練語料的不足而限制其 *N* 值的大小(通常 *N* 設爲 2 或 3)，以致於它僅能用以計算短距離詞彙接連機率，而缺乏擷取出語句中(或候選詞與歷史詞序列間)所隱含長距離語意資訊的能力。爲了解決 *N* 連語言模型參數量龐大的問題，前人的研究認爲詞序列中每一個詞都有一個其隸屬的詞類別(Word Class)，隸屬於同一個詞類別的詞可能有具有相同的語法角色或相近的語意資訊，透過詞類別資訊可以將 *N* 連語言模型的參數量降低並保有適當的模型預測能力，因而有所謂的類別 *N* 連模型(Class-based *N*-gram Model)[2]。常見的類別 *N*連模型將每一個詞對應到一個固定的詞類別，但因每一個詞實際上或許並非只有一種語意或是文法角色，所以亦有學者嘗試放寬詞與詞類別的對應，也就是讓一個詞可以隸屬於多個詞類別，爲此提出了聚合式馬可夫模型(Aggregate Markov Model, AMM)[3]。

不論是類別 *N* 連模型或是聚合式馬可夫模型的提出，皆是希望改善 *N* 連語言模型參數量過多的問題。另一方面，近十幾年來許多研究提倡探索在完整的文件或歷史詞序列中所隱含的語意資訊或是語句結構資訊等，以補足 *N*語言連模型的不足[4, 5, 6]。其主要發展可追溯到早期使用潛藏式語意分析(Latent Semantic Analysis, LSA)的研究[7]，潛藏式語意分析利用線性代數的方法，將文件（或歷史詞序列）與詞投影至一個低維度空間，在這個低維空間中試圖描述文件與詞之間的關係，同時也可解決在高維度的情況下參數量過多和訓練語料量不足的問題。後來，更有所謂的機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)[8, 9, 10]、潛藏狄利克里分配(Latent Dirichlet Allocation, LDA)[11, 12, 13, 14]及一些延伸方法陸續被提出。基本上，這些方法希望藉由機率模型的使用在低維度的語意空間中找出文件與詞的相關性。不同於潛藏語

意分析，機率式潛藏語意分析與潛藏狄利克里分配皆是機率式的生成模型，藉著對每一篇文件建立機率模型，直接表示文件與詞之間的關係，並且可以描述同義詞或者一詞多義的現象。新近，亦有所謂的詞主題模型(Word Topic Model, WTM)[15, 16]被提出。詞主題模型根據語言中每一個詞在訓練語料出現的資訊，將訓練語料作重新整理與安排，爲語言中每一個詞收集其對應的詞虛擬文件(Word Pseudo-document)，以訓練每一個詞專屬的機率生成模型，最後用來組成文件或歷史詞序列之機率生成模型以預測下一個候選詞出現的可能性。上述這些模型在本論文將統稱爲主題模型(Topic Models)[17]。

本論文提出一種新的主題模型，稱之爲詞相鄰模型(Word Vicinity Model, WVM)，它直接地基於語言中詞與詞的相互關聯資訊，建構出一個機率式的潛藏主題空間；並且透過線性模型結合的方式建立語音辨識中已解碼歷史詞序列之主題模型，用來預測下一個候選詞出現的可能性，藉此輔助傳統 N 連語言模型。本論文的安排如下：第二節將介紹近年來蓬勃發展的主題模型，包含機率式潛藏語意分析、潛藏狄利克里分配、詞主題模型；第三節將闡述本論文所提出之詞相鄰模型，並說明詞相鄰模型與現有各種主題模型之間的差異；第四節則是實驗結果與分析；第五節是結論。

二、主題語言模型相關研究

（一）機率式潛藏語意分析 (Probabilistic Latent Semantic Analysis, PLSA)

潛藏語意分析(Latent Semantic Analysis, LSA)假設文件集中文件與詞的組合存在若干潛藏語意結構成分[5]，藉由線性代數之奇異質分解(Singular Value Decomposition, SVD)可將高維度的文件向量與詞向量共同投影至一個低維度空間，其中每一維度代表某種語意結構成分；文件與詞之間語意的相似度可藉由它們在這個低維度空間的向量距離或者夾角的計算而得。如此一來，不僅可以簡化文件與詞表示方法的複雜度，也可以去除語料集中文件與詞的組合所含有的部分雜訊資訊。機率式潛藏語意分析(PLSA)是由潛藏語意分析延伸發展而來，不同於潛藏語意分析以線性代數的方法尋找語料集中隱含的主要語意結構成分，機率式潛藏語意分析利用機率模型爲每一篇文件建立生成模型，透過一組共享潛藏主題機率分布來描述每一篇文件生成文件中詞的關係[9]。

當機率式潛藏語意分析運用於語音辨識時，會將每一段歷史詞序列 H 視爲是一篇文件並估測其對應的機率式潛藏語意分析模型，用以計算在給定一段歷史詞序列 H 後下一個候選詞 w_i 出現的可能性，其機率式可表示成：

$$P_{\text{PLSA}}(w_i / H) = \sum_{k=1}^{K} P(w_i / T_k) P(T_k / H) \tag{1}$$

其中 T_k 代表一個潛藏主題，具有某種語意結構成分； $P(w_i / T_k)$ 是給定潛藏主題 T_k 的情況下，候選詞 w_i 出現的機率； $P(T_k / H)$ 是歷史詞序列產生潛藏主題 T_k 的機率。在語音辨識時，我們假設每一個潛藏主題產生候選詞的機率 $P(w_i / T_k)$ 不因詞序列搜尋及拓展過程而變動，可在執行語音辨識前就先以期望值最大化演算法(Expectation-Maximization Algorithm)[18]最大化訓練語料發生的機率而求得。另一方面，因歷史詞序列會隨著語音辨識的搜尋過程一直擴展、變動，所以歷史詞序列 H 產生每一個潛藏主題 T_k 的機率 $P(T_k / H)$ 需要不斷地被重新估算，同樣地也可以使用期望值最大化演算法來最大化歷

史詞序列發生的機率而得。機率式潛藏語意分析所擷取的長距離語意資訊可以彌補傳統 N 連語言模型在此的不足。在一般語音辨識的使用上，會將機率式潛藏語意分析與傳統 N 連語言模型經由線性插補法(Linear Interpolation)作結合，以提供在歷史詞序列 H 已解碼出的情況下每一個候選詞 w_i 發生的機率：

$$\widehat{P}(w_i / H) = (1 - \lambda_{\text{PLSA}}) \cdot P_{N-\text{gram}}(w_i / H) + \lambda_{\text{PLSA}} \cdot P_{PLSA}(w_i / H) \quad (2)$$

其中 $P_{N-\text{gram}}(w_i / H)$ 就是傳統 N 連語言模型的機率分布，我們可以使用一個介於 0 到 1 之間的可調整參數 λ_{PLSA} 來控制 N 連語言模型與機率式潛藏語意分析模型的權重。

機率式潛藏語意分析的提出讓主題空間的概念得以由線性代數描述轉往機率式模型發展，但機率式潛藏語意分析本身仍然存在著許多問題：首先，它假設在給定某一個潛藏主題的前提下，文件與詞的關係是獨立的。由語意的觀點省視，這樣的假設過度強化了詞與整體文件（或者歷史詞序列）之間的獨立性。其次，隨著我們所收集到的訓練語料集中文件數的增加，機率式潛藏語意分析模型所需的參數也會呈線性增加，有可能會讓模型參數過度符合(Overfitting)訓練語料。一個理想的機率生成模型，對於描述未見過的(Unseen)文件中的詞應具備良好的預測能力。但事實上，機率式潛藏語意分析並沒有具備健全的預測能力，其主要原因在於它對於每一套訓練語料都會產生一組獨特的潛藏主題，並非使用一組全域性的參數描述所有語料。因此當用於估測一篇嶄新文件之主題機率模型時，會受到原始訓練語料的強烈限制。另外，在模型參數的估測過程，機率式潛藏語意分析使用期望值最大化演算法來逼近訓練語料的最大相似度。但以期望值最大化演算法來估測模型參數未必能找到全域最佳(Global Maximum)解，所以模型參數的起始值設定就變得格外重要。過去有研究學者也對訓練起始值提出不少研究討論，諸如多重隨機初始(Multiple Random Initialization)、預先使用非監督式分群(Unsupervised Clustering)或是利用傳統潛藏語意分析找出較好的模型起始值皆是常用的方法[7]，不過使用這些方法卻也會成爲模型訓練過程中一種額外的負擔。最後，當將機率式潛藏語意分析被應用於語音辨識時，需要不斷地使用期望值最大化演算法來對每一歷史詞序列估算其產生潛藏主題分布的機率，但這樣的估算過程事實上是相當耗費時間的，特別在潛藏主題數目龐大時，其所需的運算時間複雜度更是驚人。雖然有學者提出使用漸進式的期望值最大化演算法估測歷史詞序列產生潛藏主題分布的機率，但其能節省的運算時間有限，並且其結果相對地顯得較差。

（二）潛藏狄利克里分配 (Latent Dirichlet Allocation, LDA)

爲了改善機率式潛藏語意分析對於未見過的文件之預測能力以及模型參數量會隨著訓練語料中文件數量的增加而呈現線性成長的缺點，有學者提出了潛藏狄利克里分配[11]。潛藏狄利克里分配的模型詮釋方式與機率式潛藏語意分析不同，並且它可僅以兩組參數 α 與 β 來代表訓練語料的潛藏語意資訊，茲簡述如下。首先，假設訓練語料集 D 中共有 M 篇文件，而每一篇文件 d 中有 N_d 個詞，我們先由一組狄利克里分配 α 的參數求得每一篇文件 d 產生所有潛藏主題的機率向量 θ_d，而文件中每一個詞在每一個潛藏主題 $T_{d,n}$ 下產生的機率分布則由 β 生成。潛藏狄利克里分配對參數的估算是最大化整個訓練語料 D 的邊際機率：

$$P_{\mathrm{LDA}}(\mathrm{D}\mid\alpha,\beta)=\prod_{d=1}^{M}\int P(\theta_d/\alpha)\left(\prod_{n=1}^{N_d}\sum_{T_{d,n}}P(T_{d,n}/\theta_d)P(w_{d,n}/T_{d,n},\beta)\right)d\theta_d \tag{3}$$

對於潛藏狄利克里分配參數的估算，前人提出不少方法諸如變動性貝氏期望值最大化(Variational Bayesian Expectation Maximization)演算法[7, 8]或吉卜森取樣(Gibbs Sampling)[9, 10]等。對於語音辨識中某一已解碼歷史詞序列 H（將 H 視爲一篇文件），潛藏狄利克里分配可以透過變動性貝氏期望值最大化演算法求得最佳參數解或是利用最大事後機率估測法估測其產生所有潛藏主題的機率向量 θ_H，而當 H 長度足以表達語意資訊後，也可以對 α 進行重新估測。另一方面，若使用吉卜森取樣重估歷史詞序列產生所有潛藏主題的機率，則是結合訓練時對訓練語料集中詞的取樣資訊與對歷史詞序列中詞的重新取樣資訊，以期望逼近潛藏主題在歷史詞序列已知情況下的事後機率。事實上，不論是變動性貝氏期望值最大化法或是吉卜森取樣，在進行重估歷史詞序列的主題分布時都是非常耗費時間的。

（三）詞主題模型 (Word Topic Model, WTM)

不論是機率式潛藏語意分析或是潛藏狄利克里分配皆是希望擷取文件或歷史詞序列中隱含的長距離語意資訊，以彌補 N 連語言模型僅考慮短距離詞彙接連規則之不足。詞主題模型則是希望在建立語言模型時不僅考慮詞彙相鄰資訊，並且透過詞彙間潛藏語意資訊的組合，建立起文件或歷史詞序列之長距離語意資訊[16]。

詞主題模型的特色是透過一組共享的潛藏主題機率分布，爲語言中每一個詞 w_j 建立一個主題模型 M_{w_j}。爲達此目的，在模型建立之前，必須從訓練語料中擷取每一個詞出現處其鄰近文字段落內其它詞出現的資訊，並將所有出現處的上下（或左右相鄰）文字段落聚集成每一個詞主題模型對應的訓練文件，稱之爲詞虛擬文件(Word Pseudo-document)。然後，透過一組共享的潛藏主題機率分布，估算每一個詞 w_j 之詞虛擬文件與其它詞 w_i 之共同出現關係；更明確些，即是 w_j 的詞主題模型 M_{w_j} 產生另一詞 w_i 的機率：

$$P_{\mathrm{WTM}}\left(w_i/\mathrm{M}_{w_j}\right)=\sum_{k=1}^{K}P(w_i/\mathrm{T}_k)P\left(T_k/\mathrm{M}_{w_j}\right) \tag{4}$$

其中 $P(w_i/T_k)$ 是給定潛藏主題 T_k 的情況下，詞 w_i 出現的機率；$P\left(T_k/\mathrm{M}_{w_j}\right)$ 是 w_j 的詞主題模型產生主題 T_k 的機率；K 則是潛藏主題總數。

當詞主題模型運用於語音辨識時，就如同機率式潛藏語意分析一般，我們將需要估算在給定了候選詞 w_i 的歷史詞序列 H 後，w_i 出現的機率。在此假設每一個潛藏主題產生候選詞 w_i 的機率 $P(w_i/T_k)$ 不隨語音辨識搜尋過程變動，並且每一個詞主題模型也已經由訓練語料求得最佳參數。因此，對於歷史詞序列 H，我們首先將它視爲由一連串的詞所組成的詞串，接著我們將詞串中每一個詞的詞主題模型利用線性插補法的方式結合，以此做爲歷史詞序列的主題模型[15]。

相較於機率式潛藏語意分析需使用期望值最大化演算法在語音辨識搜尋過程中不斷地估測歷史詞序列產生潛藏主題 T_k 的機率 $P(T_k/H)$，在使用詞主題模型時歷史詞序列

產生潛藏主題的機率 $P(T_k / H)$ 可由歷史詞序列中每一個詞 w_j 的詞主題模型產生主題 T_k 的機率 $P(T_k / \mathrm{M}_{w_j})$線性組合而成，此舉可大大地提升了語音辨識時的搜尋速度。

三、詞相鄰模型 (Word Vicinity Model, WVM)

（一）原理

與詞主題模型作法相似，本論文嘗試透過一組共享的潛藏主題分布，估算訓練語料集中相鄰詞彙間的語意關連性，稱之爲詞相鄰模型。不同於詞主題模型的是，詞相鄰模型直接對訓練語料中任意兩個詞 w_i 與 w_j 的聯合機率 $P(w_i, w_j)$透過一組潛藏主題分布所建構的語意空間作機率分解：

$$P_{\mathrm{WVM}}(w_i, w_j) = \sum_{k=1}^{K} P(w_i/T_k) P(T_k) P(w_j/T_k) \tag{5}$$

觀察式(5)與式(4)，我們可以發現詞相鄰模型包括了每一個潛藏主題的事前機率 $P(T_k)$，以及每一個潛藏主題產生每一個詞的機率分布 $P(w_i/T_k)$；而詞主題模型則是對詞彙間條件機率 $P_{\mathrm{WTM}}(w_i/\mathrm{M}_{w_j})$透過潛藏主題分布所建構的語意空間作機率分解，故有詞主題模型產生潛藏主題分布的機率 $P_{\mathrm{WTM}}(T_k/\mathrm{M}_{w_j})$以及每一個潛藏主題產生每一個詞的機率分布 $P_{\mathrm{WTM}}(w_i/T_k)$。相較之下，詞相鄰模型需要較少的模型參數量，在使用相同的訓練語料下，應會有較佳的模型參數估測表現。

當詞相鄰模型運用於語言模型的使用，諸如用於預測在給定詞 w_j 時另一詞 w_i 發生的可能性（亦即條件機率 $P(w_i \mid w_j)$），我們可以經過適當的機率式轉換，將此條件機率以詞相鄰模型的兩組機率分布 $P(T_k)$與 $P(w_i / T_k)$表示：

$$P_{\mathrm{WVM}}(w_i/w_j) = \frac{P(w_i, w_j)}{P(w_j)} = \frac{\sum_{k=1}^{K} P(w_i/T_k) P(T_k) P(w_j/T_k)}{\sum_{k=1}^{K} P(T_k) P(w_j/T_k)} \tag{6}$$

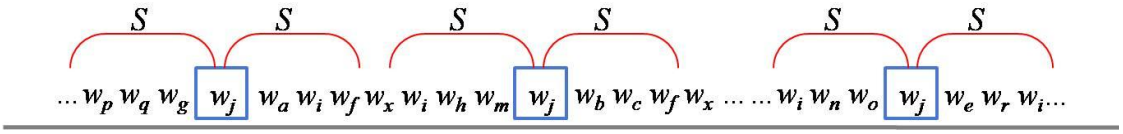
當詞相鄰模型用於語音辨識之語言模型使用時，對於每一個候選詞 w_i 以及其歷史詞序列 $H = w_1, w_2, \ldots, w_{i-1}$ ，欲估算 w_i 在給定 H 下出現的可能性時，可利用 H 中每一個詞產生 w_i 的條件機率（如式(6)）之線性組合來近似：

$$P(w_i \mid H) \approx \gamma_1 \cdot P_{\mathrm{WVM}}(w_i \mid w_1) + \gamma_2 \cdot P_{\mathrm{WVM}}(w_i \mid w_2) + \cdots + \gamma_{i-1} \cdot P_{\mathrm{WVM}}(w_i/w_{i-1}) \tag{7}$$

其中 $\{\gamma_1, \gamma_2, \ldots, \gamma_{i-1}\}$ 爲線性組合係數。再進一步來看，歷史詞序列 H 產生某一主題 T_k 的機率可經由歷史詞序列中所有詞在潛藏語意空間的分布特性而決定：

$$\begin{aligned} \tilde{P}(T_k \mid H) &= \sum_{j=1}^{i-1} \gamma_j \cdot P(T_k \mid w_j) \\ &= \sum_{j=1}^{i-1} \gamma_j \cdot \frac{P(w_j \mid T_k) P(T_k)}{\sum_{k=1}^{K} P(w_j \mid T_k) P(T_k)} \end{aligned} \tag{8}$$

其中， $P(T_k)$與 $P(w_i / T_k)$爲詞相鄰模型所求得之模型機率分布。因此，當我們將詞相鄰模型用於語音辨識之語言模型的使用時，亦可以如同機率式潛藏語意分析及潛藏狄利克里



圖一、詞相鄰模型訓練框

分配般的方式來表示歷史詞序列與候選詞間長距離的語意相關性：

$$P_{\mathrm{WVM}}(w_i / H) = \sum_{k=1}^{K} P(w_i / T_k)\tilde{P}(T_k / H) \qquad (9)$$

值得注意的是，詞相鄰模型對於式(9)其實是透過歷史詞序列中每一個詞與候選詞 w_i 兩兩間在潛藏語意空間上的機率分布關係而計算出，這一點與詞主題模型相似；而機率式潛藏語意分析及潛藏狄利克里分配是將歷史詞序列視爲一整體，計算其與候選詞在潛藏語意空間上的機率分布關係。

在詞相鄰模型的訓練方面，爲了估算語料庫中任意兩個詞 w_i 與 w_j 共同出現的聯合機率 $P(w_i, w_j)$，我們首先須決定一個訓練框 S，用來固定選取每一個詞 w_i 出現時上下（左右相鄰）文段中有哪些的詞彙出現，並估算它們個別與詞 w_i 在訓練語料中會共同出現在此訓練框的次數，以 $n(w_i, w_j)$ 表示，而會有 $n(w_i, w_j) = n(w_j, w_i)$ 的性質。再者，我們假設在討論任意兩兩詞彙間共同出現的關係時，不受到其它詞彙或其它詞彙之間的關係的影響。因此，詞相鄰模型的訓練是以最大化詞典 V 中任意兩個詞 w_i 與 w_j 在訓練語料共同出現在一定範圍上下文段（或訓練框）的聯合機率 $P_{\mathrm{WVM}}(w_i, w_j)$（參見式(5)）之對數機率值總和 L_{WVM} 爲目標：

$$L_{\mathrm{WVM}} = \sum_{w_i, w_j \in \mathrm{V}} n(w_i, w_j) \log P_{\mathrm{WVM}}(w_i, w_j) \qquad (10)$$

我們藉著使用期望值最大化演算法來進行其中詞相鄰模型機率式之估測。

（二）其他主題模型之比較

在此，我們由圖形模型(Graphical Model)表示、模型參數量多寡、以及於語音辨識時之執行效能等幾個觀點分析與比較各種主題模型之間的關係與優劣，如表一所歸納。

1. 圖形模型表示

首先，藉由圖形模型表示機率式潛藏語意分析(PLSA)。如圖二(a)所示，我們可觀察出機率式潛藏語意分析是先考慮每一篇文件生成每一潛藏主題的機率，接著聯合一組潛藏主題分布分別產生每一個詞的機率，使每一篇文件成爲一個具有預測能力的生成模型。另一方面，詞主題模型(WTM)則收集每一個詞在語料庫中出現位置鄰近處的詞合成對應的詞虛擬文件，考慮每一個詞與詞虛擬文件之間的關係，其圖形模型表示如圖二(b)所示。詞主題模型的圖形模型表示與機率式潛藏語意分析之圖形模型表示非常相似，主要差別在於機率式潛藏語意分析以訓練語料庫中每一篇文件爲模型單位，而詞主題模型則爲語言中每一個詞重新整理其在訓練語料出現資訊而有所謂的詞虛擬文件來作爲模型單位。

表一、各主題模型之比較

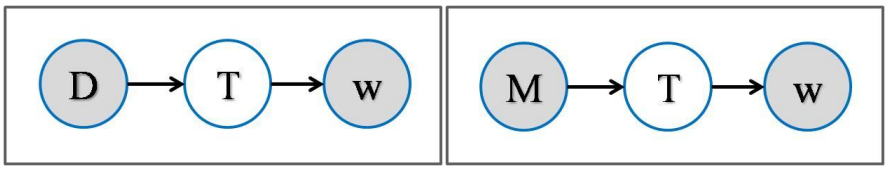
模型	機率式潛藏語意分析	潛藏狄利克里分配	詞主題模型	詞相鄰模型
模型對象	文件與詞	文件與詞	虛擬詞文件與詞	詞與詞
模型參數量	$N \times K + M \times K$	$K + N \times K$	$2 \times N \times K$	$K + N \times K$
用於語音辨識之方式	即時重新估算歷史詞序列之主題分布	即時重新估算歷史詞序列之主題分布	詞主題模型線性結合	機率模型線性結合
速度	中等	慢	快	快

再者，機率式潛藏語意分析在訓練語言模型時，將每一篇文件的語意資訊考慮進模型參數，求取參數的過程中希望最大化每篇文件產生詞的機率，過去研究人員認爲這樣的訓練過程會使得模型參數估測受到訓練語料中文件的限制。如用於未見過的文件其主題分布能符合訓練語料的特質，其模型的預測能力將會有不錯效果；但若用於預測主題偏差較大的未知文件，則可能就無法得到良好的效果。有別於將文件語意資訊直接考慮於模型參數，潛藏狄利克里分配(LDA)的文件主題生成方式僅以兩組參數（α 與 β）描述，如此可以避免潛藏主題機率參數過度符合訓練語料所收集到的文件，讓生成模型對於預測新文件時更有彈性，潛藏狄利克里分配可表示成一個三層架構的之圖形模型，如圖二(c)所示，其中 α 與 β 是潛藏狄利克里分布模型之所擁有的兩組參數。

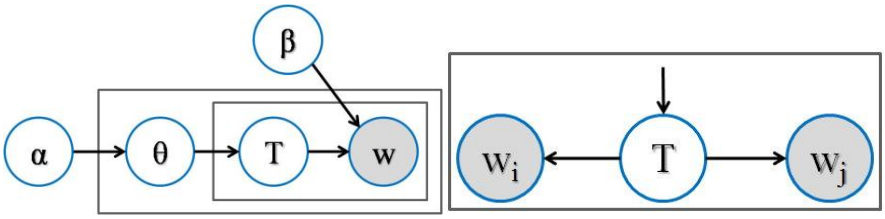
有別於將文件層次語意資訊考慮於模型訓練之中，我們希望模型可以不受文件過度束縛。透過不同層次（詞層次）潛藏主題模型的引入，讓文件生成潛藏主題的過程更具彈性，但卻不至於讓模型演繹過程過度複雜。因此，在本論文我們提出了詞相鄰模型，直接透過估算訓練語料中相鄰詞彙間的語意關連性，建立一組詞彙間共享的主題語意空間，描繪詞與詞之間的相互關係（如圖二(d)所示）。當選定一個潛藏主題後，詞相鄰模型提供一個描述在這個主題下每一個詞出現的可能性的機率分布，而每一個潛藏主題本身有其發生的事前機率。

2. 模型參數量分析

在給定詞典 V（共 N 個詞，$\mathrm{V} = \{w_1, w_2, \ldots, w_N\}$）、訓練語料 D（共 M 篇文件詞，$\mathrm{D} = \{d_1, d_2, \ldots, d_M\}$）和假設潛藏主題數爲 K 的情況下，我們比較各主題模型的模型複雜度。機率式潛藏語意分析有每篇訓練文件產生每一個潛藏主題的機率 $P(T_k/d_m)$ 以及潛藏主題產生詞的機率 $P(w_n/T_k)$，共需 $N \times K + M \times K$ 個參數；詞主題模型則擁有每個詞主題模型產生潛藏主題的機率分布 $P(T_k/\mathrm{M}_{w_{n'}})$ 以及每一個潛藏主題產生每一個詞的機率分布 $P(w_n/T_k)$，共需 $2 \times N \times K$ 個參數；潛藏狄利克里分配則僅需要兩組參數 α 與 β 共 $K + N \times K$ 個參數；最後詞相鄰模型亦僅需 $K + N \times K$ 個參數，分別是每一個潛藏主題的機率 $P(T_k)$，以及每一個潛藏主題產生每一個詞的機率 $P(w_n/T_k)$。當訓練語料中所含的文件數小於詞典大小($M < N$)時，詞主題模型是四個主題模型中參數量最多的，但是若隨著收集的訓練語料越來越多機率式潛藏語意分析的參數量會呈現線性增加，而詞主題模型的參數量是固定的，所以當收集的訓練語料所含的文件數大過詞典大小($N < M$)時，機率式潛藏語意分析會需要最多的參數。潛藏狄利克里分配與詞相鄰模型則不論訓練語料大小，所需的參數量僅和潛藏主題個數與詞典大小有關。



(a) 機率式潛藏語意分析(PLSA)　　(b) 詞主題模型(WTM)



(c) 潛藏狄利克里分配(LDA)　　(d) 詞相鄰模型(WVM)

圖二、主題模型之圖形模型表示

3. 於大詞彙連續語音辨識運用之分析

「即時性」是當今語音辨識技術能否被廣爲使用的關鍵因素，本論文因此對於上述主題模型運用於語音辨識時之執行效能作概略分析。當機率式潛藏語意分析(PLSA)運用於語音辨識時，它最爲人所詬病的是使用期望值最大化法線上估測歷史詞序列的潛藏主題分布；雖然即時估測可以針對歷史詞序列重新計算獲得相對準確的主題分布，但這樣的過程實在過於耗費時間。對照於機率式潛藏語意分析，詞主題模型(WTM)使用線性組合的方式，直接將已解碼的歷史詞序列中每一個詞的詞主題模型線性結合，以此作爲歷史詞序列的主題分布。雖然使用詞主題模型所得到的歷史詞序列主題分布，如同機率式潛藏語意分析一樣，會受到訓練語料集的限制。但實際運用於語音辨識時，詞主題模型可以省去像機率式潛藏語意分析所需耗時的線上主題分布重估，具即時性之優點。另一方面，潛藏狄利克里分配(LDA)雖僅用少量的參數描述訓練語料集之主題分布特性，當直接對觀測到的歷史詞序列重估主題分布，亦可較不受訓練語料庫中文件的限制[11]。但當潛藏狄利克里分配被使用於語音辨識時，一樣遭受重估過於耗時的問題。就我們將語言模型使用於語音辨識實驗所作觀察，潛藏狄利克里分配是四者中最耗時的模型。最後，當詞相鄰模型(WVM)運用於語音辨識時，我們將歷史詞序列視爲由許多詞所組成的詞串，透過適當的機率轉換計算出在給定歷史詞序列中每一個詞下任一個候選詞出現的可能性，再如同詞主題模型以線性插補法的方式結合這些條件機率，以此做爲歷史詞序列的主題模型（參見式(7)）。其過程中雖然需透過一次的機率式轉換（參見式(6)），但是當實際運用於語音辨識時，詞相鄰模型在與機率式潛藏語意分析和潛藏狄利克里分配相較下，仍然擁有較佳的執行速度。

四、實驗結果與分析

（一）實驗設定

在語音特徵擷取部分，我們以梅爾率波器組(Mel-frequency Filter Bank)輸出爲基礎，使用異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)配合最大

化相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)，最後獲得 39 維語音特徵向量。另外，在辨識所需的聲學模型訓練上，考慮了中文語音結構，聲學模型由 22 個 INITIAL 模型、38 個 FINAL 模型（每個中文的音節都是由一個 INITIAL 及一個 FINAL 所組成）及一個靜音(Silence)模型組成，其中 INITIAL 模型會因其右邊可能接的 FINAL 模型種類而進一步細分成 112 個 INITIAL 模型[19]。我們最後總共使用了 151 個隱藏式馬可夫模型(Hidden Markov Models)來作爲這些 INITIAL-FINAL 聲學模型的統計模型。在隱藏式馬可夫模型中，每個狀態則依據其對應到的訓練語料多寡，以 2 到 128 個高斯統計分布來表示，不管男女性別都使用同一套聲學模型。聲學模型首先經由最大化相似度估測(Maximum Likelihood Estimation, MLE)訓練而得，再透過最小化音素錯誤訓練(Minimum Phone Error, MPE)以期望獲得最佳化聲學模型參數[20]。

本論文實驗使用的訓練與測試語料爲 MATBN 電視新聞語料庫[21]，是由中央研究院資訊所口語小組耗時三年與公共電視台合作錄製完成。我們初步地選擇外場採訪記者語料作爲實驗題材，將其中約 25 小時收錄於 2001 年 11 月至 2002 年 12 月期間的語料作爲聲學模型訓練語料，再由 2003 年的收錄語料中定義各約 1.5 小時做爲發展集語料（MATBN 發展集）以及測試集語料（MATBN 測試集），詳細資料集資訊如表二所示。更明確地，我們將由 MATBN 發展集中選定最佳模型參數，並將此參數運用於測試集語料，比較與討論各種主題模型的效能。

另一方面，背景三連語言模型(Trigram Language Model)訓練語料則是來自中央通訊社 2001 年至 2002 年的文字新聞語料，包含了約一億五千萬個中文字，經斷詞後約有八千萬詞。本論文實驗爲語言模型調適，我們由公視廣播新聞語料 2001、2002 與 2003 年的人工轉寫文件中篩選出約三千六百篇報導，約兩百萬個中文字，經斷詞後約有一百萬詞，作爲調適語料。詞典大小約爲七萬兩千詞。採用 SRI Language Modeling Toolkit[22] 訓練實驗所需要的三連語言模型。

論文將主題模型用於調適背景三連語言模型，其方式爲模型插補法。如式(2)所示，調整主題模型與背景三連語言模型影響的權重參數是先由 MATBN 發展集調整至最佳後，再用於 MATBN 測試集。語言模型效能的評估，是透過語言複雜度(Perplexity, PP)以及大詞彙連續語音辨識之辨識字錯誤率(Character Error Rate, CER)來達成。

實驗中我們將詞相鄰模型的訓練框 s 設定爲 2；另外，詞相鄰模型與詞主題模型需要給歷史詞序列中每一個詞的主題模型一個語言模型影響權重 γ_j，我們利用詞與詞之間的距離定義一個指數遞減函數，用來給定每一個詞的主題模型一個語言模型影響權重：

$$\gamma_j = \phi_j \prod_{s=j+1}^{i-1} (1-\phi_s) \tag{11}$$

當 $j = 2,\ldots,i-1$ 時 ϕ_j 是一個介於 0 到 1 的定值，而 ϕ_j 爲 1，並且此遞減函數也會滿足 $\sum_{j=1}^{i-1} \gamma_j = 1$。實驗中，我們將 ϕ_j 設爲 0.6。

再者，如同機率式潛藏語意分析一般，詞相鄰模型與詞主題模型亦可以假設潛藏主題產生每一個詞的機率不隨辨識過程變動，利用期望值最大化演算法調整歷史詞序列的主題機率分布，即調整式(7)或(8)中每一個歷史詞序列中詞的線性組合係數（或語言模

表二、資料集

	MATBN 發展集	MATBN 測試集	NOWnews 測試集
總句數	292	307	13,810
總詞數	16,106	16,494	1,075,409

表三、由詞相鄰模型(32 Topics)中選取出 4 個主題

Topic 8	Topic 13	Topic 14	Topic 23
詞(word) 權重(weight)	詞(word) 權重(weight)	詞(word) 權重(weight)	詞(word) 權重(weight)
主委陳菊 0.792	靜脈 1.202	平均地權 1.306	霍亂 0.752
發布新聞稿 0.750	顯微 1.002	公職人員財產申報 1.259	大腸直腸癌 0.681
總召柯建銘 0.630	切除 0.674	土地稅 0.704	沙門氏菌 0.471
副總裁陳師孟 0.625	肌瘤 0.668	菸酒稅法 0.489	口蹄疫 0.337
宜蘭縣長 0.564	腦炎 0.618	財稅 0.457	甲狀腺 0.303
副院長賴英照 0.550	子宮 0.501	修正草案 0.446	胃癌 0.298
立法院黨團 0.519	支氣管 0.500	財政收支劃分 0.428	徵狀 0.269
機要 0.495	縫合 0.463	購併 0.396	寄生 0.268
中央研究院院長 0.489	割除 0.367	暫行條例 0.383	皮膚癌 0.267
聯邦準備理事會 0.469	氣管 0.344	保險法 0.373	肺癌 0.234

表四、基礎實驗結果

baseline	MATBN 發展集		MATBN 測試集		NOWnews 測試集	
	CER(%)	PP	CER(%)	PP	CER(%)	PP
Trigram	20.22	667.23	20.08	682.10	null	808.76

型影響權重）。我們將利用指數遞減函數（式(11)）估測歷史詞序列之主題分布的方式以(ED)表示，而期望值最大化演算法估測的方式以(ML)表示之。

（二）實驗結果與分析

首先，我們由 32 個主題數的詞相鄰模型中取出 4 個潛藏主題，並且計算每個詞分別屬於不同潛藏主題時的主題分數(Topic Score)。其中某一個詞 w_i 隸屬於某一個潛藏主題 I_k 時的主題分數定義如下[23]：

$$TS(w_i, T_k) \equiv \frac{\sum_{m=1}^{M} c(w_i, d_m) P(T_k / d_m)}{\sum_{m=1}^{M} c(w_i, d_m)(1 - P(T_k / d_m))} \tag{12}$$

其中 $c(w_i, d_m)$ 爲詞 w_i 出現在文件 d_m 的次數， $P(T_k \mid d_m)$ 爲文件 d_m 詞產生潛藏主題 T_k 的機率。對於這 4 個潛藏主題，我們分別挑選出主題分數較大的 10 個詞彙，如表三所示。我們可以發現 Topic 8 傾向政黨政治新聞，Topic 13 收集醫學和醫療等相關資訊，Topic 14 關於政府稅收的主題資訊，最後 Topic 23 則是把疾病和病毒等名稱聚集在一起。由此實驗可知，詞相鄰模型本身對於訓練語料亦具備良好的非監督式分群之能力。

接著，我們比較各主題模型之語言複雜度(Perplexity, PP)。語言複雜度最早是由資訊理論發展而來，用來評估一個語言模型的好壞，其幾何意義爲語言模型產生一段文字的機率倒數再取幾何平均數，可視爲語言模型預測詞與詞接連的平均分支度。語言複雜度越小，表示所訓練的語言模型越具有預測詞產生的能力。如表四所示，背景三連語言模型在 MATBN 發展集所得的語言複雜度爲 667.23，而在 MATBN 測試集的語言複雜度爲 682.10。表五是當各種主題模型與背景三連語言模型結合後，作用於 MATBN 測試集的語言複雜度實驗結果；我們可以發現，各主題模型隨著主題數陸續增加語言複雜度也隨之降低。我們亦可由表五觀察到，在不同主題數設定時，不論是詞相鄰模型或是詞主題模型，其語言複雜度表現大都較潛藏狄利克里分配佳（有較低的語言複雜度值），亦較機率式潛藏語意分析好。另外，詞相鄰模型使用期望值最大化演算法估測歷史詞序列中每一個詞的主題模型之語言模型影響權重的方式（即 WVM(ML)）較使用指數遞減函數的方式（即 WTM(ED)，參見式(11)）有較低的語言複雜度值。但若對於詞主題模型而言，使用指數遞減函數的方式估測歷史詞序列中每一個詞的主題模型之語言模型影響權重（即 WTM(ED)）會較使用期望值最大化演算法（即 WTM(ML)）爲佳。另一方面，若比較詞相鄰模型與詞主題模型時，則可發現當潛藏主題數較小時詞相鄰模型(WVM(ML))有最低的語言複雜度，但隨著主題數漸漸增加詞主題模型(WTM(ED))的語言複雜度會快速下降。

當將上述這些主題模型與背景三連語言模型結合時，均能較僅使用背景三連語言模型時有明顯的語言複雜度降低。以最佳實驗設定而言，機率式潛藏語意分析有 23.1%、潛藏狄利克里分配有 21.4%、詞主題模型(WTM(ED))有 26.1%、詞相鄰模型(WVM(ML))有 24.2%的相對語言複雜度降低。

再者，我們比較各種主題模型與背景三連語言模型結合後，運用於大詞彙連續語音辨識時的辨識字錯誤率。如表四所示，在基礎實驗中，MATBN 發展集的字錯誤率爲 20.22%，MATBN 測試集的字錯誤率爲 20.08%。表五展現各主題模型與背景三連語言模型結合後，在不同潛藏主題數設定下的辨識詞錯誤之實驗結果。我們可以觀察到，當主題數設定爲 32 或 64 時各種主題模型可以分別獲得最佳辨識結果。以使用各種主題模型的最低字錯誤率來說，機率式潛藏語意分析有 4.1%、潛藏狄利克里分配有 5.2%、詞主題模型(WTM(ML))有 3.9%、詞主題模型(WTM(ED))有 5.5%、詞相鄰模型(WVM(ML))有 4.0%、詞相鄰模型(WVM(ED))有 5.0%的相對字錯誤率降低。我們發現詞相鄰模型與詞主題模型以線性結合的方式來估算歷史詞序列主題分布的方法運用於大詞彙連續語音辨識時皆可獲得不錯的實驗結果，當潛藏主題數設定爲 64 時詞主題模型(WTM(ED))

表五、各主題模型於 MATBN 測試集之實驗結果

MATBN 測試集	PLSA		LDA		WTM(ML)		WTM(ED)		WVM(ML)		WVM(ED)	
	CER(%)	PP	CER(%)	PP	CER(%)	PP	CER(%)	PP	CER(%)	PP	CER(%)	PP
8 topics	19.26	553.92	19.25	557.02	19.50	539.87	19.17	542.65	19.48	531.63	19.31	546.11
16 topics	19.40	547.03	19.11	550.35	19.44	529.73	19.19	533.41	19.49	528.60	19.18	540.45
32 topics	19.26	535.93	19.06	539.21	19.45	526.61	19.14	521.27	19.27	523.11	19.15	537.17
64 topics	19.29	530.85	19.24	537.03	19.29	523.88	18.98	509.18	19.37	516.75	19.22	530.73
128 topics	19.34	524.60	19.14	536.10	19.39	528.38	19.13	503.74	19.36	519.33	19.23	527.94

表六、不同歷史詞長度(L)之詞相鄰模型於 MATBN 測試集之實驗結果(CER(%))

L	1	2	4	8	16
8 topics	19.32	19.25	19.23	19.31	19.31
16 topics	19.24	19.17	19.18	19.18	19.18
32 topics	19.29	19.21	19.13	19.15	19.15
64 topics	19.39	19.28	19.21	19.20	19.22
128 topics	19.44	19.29	19.22	19.23	19.23

更可獲得最低的辨識字錯誤率。

進一步地，我們針對詞相鄰模型進行探討，當詞相鄰模型以線性結合的方式估算歷史詞序列主題分布的方法使用於語音辨識時，我們可以假設一個候選詞的出現僅與前 L 個詞有關，用以簡化計算複雜度。表六為假設候選詞出現的可能性僅與前 L 個詞相關時的實驗結果；結果顯示出，不同的 L 值皆大約在將主題數設為 32 或 64 時有最低的字錯誤率。而最低的字錯誤率是設定潛藏主題數共 32 個並且假設預測詞的出現僅與前 4 個詞相關的時候，其字錯誤率約為 19.13%。值得注意的是，當我們比較 L（歷史詞長度）為 8 與 16 時，辨識字錯誤率僅當潛藏主題數設為 64 時相差 0.02%，當比較完整的歷史詞序列與僅考慮前 16 個歷史詞時，在不同的潛藏主題數下皆獲得相同的辨識字錯誤率。我們在此推論，詞相鄰模型主要描述訓練語料中任意兩個鄰近詞的共同出現情況，所以距離候選詞較遠的詞僅能扮演輔助的角色，無法對候選詞可能出現與否有決定性的影響。

最後，我們比較各主題模型於同時期(Contemporary)測試文字語料集的語言複雜度。於是，我們收集了與 MATBN 測試集時期相近的今日新聞(NOWnews)文字新聞語料，由 2003 年 1 月至 4 月的新聞中挑選出共約五千六百多則新聞，包含約一萬三千則句子，以此做為同時期的測試集語料（NOWnews 測試集）。其基礎語言複雜度為 808.76，詳細語料資訊列於表二。在此，我們將 MATBN 發展集中調定的最佳參數運用於此測試集語料中。表七是各主題模型與背景三連語言模型結合後於 NOWnews 測試集的語言複雜

表七、各主題模型於 NOWnews 測試集之實驗結果(PP)

NOWnews 測試集	PLSA	LDA	WTM(ML)	WTM(ED)	WVM(ML)	WVM(ED)
8 topics	751.89	747.60	702.16	766.72	694.57	766.06
16 topics	742.71	738.61	692.20	762.50	690.77	765.59
32 topics	732.80	732.35	684.71	760.43	682.63	765.79
64 topics	726.17	731.58	679.17	760.62	674.04	761.42
128 topics	716.97	728.29	672.91	762.41	667.99	761.66

度實驗結果。雖說 NOWnews 測試集與語言模型調適語料屬同一時期之新聞語料，但是 NOWnews 測試集是文字新聞，而語言模型調適語料是廣播新聞語音轉寫文字，兩者在主題上有相當的相近度，但在語句或詞彙的使用上會不相同。實驗結果顯示，詞相鄰模型(WVM(ML))與詞主題模型(WTM(ML))對於這些文字新聞的主題預測能力優於機率式潛藏語意分析和潛藏狄利克里分配。相較於機率式潛藏語意分析和潛藏狄利克里分配，詞相鄰模型與詞主題模型希望在建立語言模型時不僅考慮文件或歷史詞序列之長距離語意資訊並且保有鄰近詞的詞彙資訊。故於此實驗中，雖然機率式潛藏語意分析也是以期望值最大化演算法來估測歷史詞序列的主題分布，但語言複雜度仍然高於使用相同方法估測歷史詞序列的詞相鄰模型(WVM(ML))與詞主題模型(WTM(ML))。在最佳的實驗設定下，詞相鄰模型(WVM(ML))可以獲得 17.4%的相對語言複雜度降低，而詞主題模型(WTM(ML))、機率式潛藏語意分析潛藏狄利克里分配分別有有 16.8%、11.3%與 9.9%的相對語言複雜度降低。

五、結論與未來展望

本論文提出一個嶄新的觀點－基於語言中詞與詞的關聯資訊來建構一個潛藏主題空間。我們嘗試於此空間中討論文件與詞之間的關係，提出詞相鄰模型(Word Vicinity Model, WVM)。本論文中討論詞相鄰模型之語言複雜度(Perplexity)，以及運用於大詞彙連續語音辨識時之辨識字錯誤率(Character Error Rate)。結果顯示詞相鄰模型相較於大部分主題模型擁有較低的語言複雜度；當運用於大詞彙連續語音辨識時，詞相鄰模型相較於基礎辨識率亦有 5.0%的相對進步率。未來，我們將研究詞相鄰模型運於不同領域的可用性，以及強化詞相鄰模型運用於大詞彙連續語音辨識時對於歷史詞序列的主題模型估測。

六、致謝

本研究承蒙國科會研究計畫 NSC 98-2221-E-003-011-MY3 、NSC96-2628-E-003-015-MY3、NSC97-2631-S-003-003 的部分補助，僅此致謝。

參考文獻

[1] F. Jelinek, "Up from trigrams! - the struggle for improved language models," in *Proc. of Eurospeech*, 1991.
[2] PF. Brown, VJ. Della Pietra, PV. deSouza, JC. Lai, and RL. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, 18(4):467–479,

December, 1992
[3] L. Saul and F. Pereira, "Aggregate and mixed-order Markov models for statistical language processing," in *Proc. of EMNLP*, 1997.
[4] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: topic mixtures versus dynamic cache models," Speech and Audio Processing, IEEE Transactions, 1999.
[5] J. Bellegarda, "Statistical language model adaptation: review and perspectives," in *Speech Communication*, 2004.
[6] R. Rosenfeld , "Two decades of Statistical Language Modeling: Where Do We Go From Here?", in *Proc. of the IEEE*, 2000.
[7] J.R. Bellegarda, *Latent Semantic Mapping: Principles and Applications*. Morgan and Claypool, 2007.
[8] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. of* SIGIR, 1999.
[9] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, 2001.
[10] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proc. of Eurospeech*, 1999.
[11] D.M. Blei, A.Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 2003.
[12] Y. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proc. of Interspeech*, 2005.
[13] T. Griffiths, "Gibbs sampling in the generative model of Latent Dirichlet Allocation," Technical Report.
[14] T. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. of the National Academy of Sciences*, 2004.
[15] H.-S. Chiu and B. Chen, "Word topical mixture models for dynamic language model adaptation," in *Proc. of ICASSP*, 2007.
[16] B. Chen, "Latent topic modeling of word co-occurrence information for spoken document retrieval," in *Proc. of ICASSP*, 2009.
[17] M. Steyvers and T. Griffiths, "Probabilistic topic models." In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Hillsdale, NJ: Erlbaum.
[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society*, 1977.
[19] B. Chen, J.-W. Kuo, and W.-H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. of ICASSP*, 2004.
[20] S.H. Liu, F.H. Chu, and B. Chen, "Improved MPE-Based Discriminative Training of Acoustic Models for Mandarin Large Vocabulary Continuous Speech Recognition," in *Proc.* of *ROCLING*, 2007. (in Chinese)
[21] H.-M. Wang, B. Chen, J.-W. Kuo and S.-S. Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics & Chinese Language Processing*, 2005.
[22] A. Stolcke, SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm/
[23] T.-H. Li, M.-H. Lee, B. Chen and L.-S. Lee, "Hierarchical Topic Organization and Visual Presentation of Spoken Documents Using Probabilistic Latent Semantic Analysis (PLSA) for Efficient Retrieval/Browsing Applications," in *Proc. of Eurospeech*, 2005.

Improving Translation Fluency with Search-Based Decoding and a Monolingual Statistical Machine Translation Model for Automatic Post-Editing

Jing-Shin Chang
Department of Computer Science
& Information Engineering
National Chi Nan University
1, Univ. Road, Puli, Nantou 545, TAIWAN
jshin@csie.ncnu.edu.tw

Sheng-Sian Lin
Department of Computer Science
& Information Engineering
National Chi Nan University
1, Univ. Road, Puli, Nantou 545, TAIWAN
s94321509@ncnu.edu.tw

Abstract

The BLEU scores and translation fluency for the current state-of-the-art SMT systems based on IBM models are still too low for publication purposes. The major issue is that stochastically generated sentences hypotheses, produced through a stack decoding process, may not strictly follow the natural target language grammar, since the decoding process is directed by a highly simplified translation model and n-gram language model, and a large number of noisy phrase pairs may introduce significant search errors. This paper proposes a statistical post-editing (SPE) model, based on a special monolingual SMT paradigm, to "translate" disfluent sentences into fluent sentences. However, instead of conducting a stack decoding process, the sentence hypotheses are searched from fluent target sentences in a large target language corpus or on the Web to ensure fluency. Phrase-based local editing, if necessary, is then applied to correct weakest phrase alignments between the disfluent and searched hypotheses using fluent target language phrases; such phrases are segmented from a large target language corpus with a global optimization criterion to maximize the likelihood of the training sentences, instead of using noisy phrases combined from bilingually word-aligned pairs. With such search-based decoding, the absolute BLEU scores are much higher than automatic post editing systems that conduct a classical SMT decoding process. We are also able to fully correct a significant number of disfluent sentences into completely fluent versions. The BLEU scores are significantly improved. The evaluation shows that on average 46% of translation errors can be fully recovered, and the BLEU score can be improved by about 26%.

Keywords: Translation Fluency, Fluency-Based Decoding, Search-Based Decoding, Statistical Machine Translation, Automatic Post-Editing

1 Introduction and Motivation

1.1 Fluency Problems with Statistical Machine Translations

Translation fluency of Machine Translation systems is a serious issue in the current SMT research works. With the research efforts for the past tens of years, the performances are still far from satisfactory. In translating English to Chinese, for instance, the BLEU scores [16] range only between 0.21 and 0.29 [22, 5, 17], depending on test sets and numbers of reference translations. Such translation quality is extremely disfluent for human readers. We therefore propose a statistical post-editing (SPE) model, based on a special monolingual SMT framework, for improving the fluency and adequacy of translated sentences.

The classical IBM SMT models [1, 2] formulate the translation problem of a source sentence F as finding the best translation E* from some stack decoded hypotheses, E, such that:

$$\begin{aligned} E^* &= \arg\max_E \Pr(E \mid F) \\ &= \arg\max_E \Pr(F \mid E) \times \Pr(E) \end{aligned} \quad (1)$$

where $\begin{cases} E\text{:target sentence} \\ F\text{:source sentence} \end{cases}$ and $\begin{matrix} \Pr(F \mid E) : Translation\ \ Model\ \ (TM) \\ \Pr(E) : Language\ \ Model\ \ (LM) \end{matrix}$

The $\arg\max_E$ operation implies to generate candidate target sentences E of F so that the SMT model can score each one, based on the TM and LM scores and select the best candidate. The process of candidate generation is known as the decoding process. The conventional decoding process is significantly affected by the TM and LM scores; only those candidates that satisfy the underlying criteria of the TM and LM will receive high scores. Unfortunately, to make the SMT computationally feasible, the TM and LM are highly simplified. Therefore, the candidates are not really generated based on target language grammar, but based on the model constraints. For instance, the classical SMT model does not prefer word re-ordering with long distance movement. Such candidates are then not generated regardless of the possibility that the target grammar might prefer them.

1.2 LM and Decoding

There are three directions to improve the translation fluency with the classical SMT model, Equation (1). Firstly, we can improve the Translation Model (TM) to fit the source-target transfer process. Secondly, we can improve the Language Model (LM) to respect the target language grammar. Finally, we could try to generate better and much more fluent candidates in the decoding process so that the TM and LM can select the real best one from fluent candidates, rather than from junk sentences.

The research communities normally focus on the TM and LM components by assuming that there are good ways to generate good candidates for scoring. Actually, most attention is paid to the Translation Model (TM); LM and decoding were not gaining the same weight. In particular, people tend to think that the candidate generation process guided by the highly simplified TM and LM will eventually generate good candidates.

Unfortunately, to make the computation feasible, the classical SMT models have very low expressive power in the Translation Model (TM) and Language Model (LM) components. It formulates the TM in terms of the *fertility* probability, lexical *translation* probability and *distortion* probability [1, 2]. A word-based 3-gram model is usually used as the language model (LM). Longer n-grams are used at higher training cost and severe data sparseness.

In fact, the candidates of the target sentence, which are hidden in the $\arg\max_E$ operator, are generated as a stochastic process in most SMT today. Starting from a particular state, the next word is predicted based on a local n-gram window within a distance allowed by the distortion criterion; the possible paths are exploited using stack decoding, beam search or other searching algorithms. The candidates generated in this way thus may be only "piecewise" consistent with the target language grammar, but may not be really globally

grammatical or fluent. This means that the TM and LM are not scoring a complete sentence but some segments pasted by the n-gram LM. It is then not likely to be fluent all the time.

This decoding process therefore sometimes falls into the "garbage-in and garbage-out" situation. No matter how well-formulated the TM and LM may be, if the stochastically generated candidates do not include the correct and fluent translation, the system will eventually deliver a garbage output, that is, a disfluent sentence, as the *best* one. This kind of error is known as searching error. Because the TM and LM have limited expressive power to describe the real criteria that carry the generation process, the decoding process might only generate noisy sentence segments and thus disfluent sentences for scoring. This could lead to bad performance in terms of BLEU score or human judgments.

Phrase-based SMT had partially resolved the expressive power issue of TM and LM by using longer word sequences. However, the acquisition of "phrases" has its own problems. In particular, most phrase-based SMT acquires the phrase pairs by conducting bilingual word alignment first. Adjacent words are then connected in some heuristic ways [12, 13, 14, 15], which do not have direct link with the source or target grammar, to form the "phrases". The phrases generated in this way normally do not satisfy any global optimization criteria related to the target grammar, such as maximizing the likelihood of the target language sentences. The quality of such phrases is therefore greatly affected by the word alignment accuracy; and, the phrases for the target language side may not really respect the target grammar. Under such circumstances, a huge number of noisy "phrases" will be introduced and significantly enlarge the searching space. The stochastically generated phrase sequences thus may not correspond to good candidate sentences either.

To summarize, the application of word-for-word or phrase-to-phrase translation (with "noisy" phrases) plus a little bit local word/phrase re-ordering in classical SMT might not generate fluent target sentences that respect the target grammar. In particular, many target specific lexical items and morphemes cannot be generated through this kind of models. If they do, they may be generated in very special ways. This could be a significant reason why the SMT models do not work well after the long period of research.

The implication is that we might have to examine the $\arg\max_{E}$ operation, that is, the *decoding* or *searching* process, in the classical SMT models more carefully. We should try decoding method that respect target grammar more, instead of following the criteria set forth by the TM and LM of the SMT model, which encode highly simplified version of the target grammar. Only with a decoding process that respect the target grammar, will the system generate fluent candidates at the first place before submitting the candidates to the TM and LM for scoring.

Furthermore, a phrase-based language model, instead of word-based n-gram model for the target side may improve the fluency of machine translation further since more context words can be consulted, if the "phrases" are not noisy. To avoid a huge number of noisy source-dependent phrases that might be harmful for fluency and searching, such phrases may better be trained from a target corpus, instead of being acquired from bilingually word-aligned chunks.

1.3 Statistical Post-Editing Model Based on Monolingual SMT

Instead of developing new models for the TM and LM, an alternative to improve the translation fluency is to cascade an Automatic Post-Editing (APE) module to the translation output of an MT/SMT system. While the classical SMT models may not be suitable for directly *generating* fluent translation, due to the limited expressive power of the TM and LM and search errors of the decoding process, an SMT or its variant may be sufficient for re-ranking hypotheses in the automatic post editing purposes, if appropriate hypotheses generation mechanism is available. Actually, we can regard a post-editing process as a translation process from disfluent sentence to fluent sentence. This is particularly true if the disfluency is limited to local editing operations like *insertion* of target specific morphemes, *deletion* of source-specific function words, and *lexical substitution* from many possible lexical choices. These kinds of errors are often seen in MT/SMT systems. Inspired by the above ideas, this paper propose a statistical post-editing (SPE) model based on a monolingual SMT paradigm for improving the translation fluency of an MT system, instead of improving the TM directly.

In this SPE model, the searching or decoding is a fluency-based search. We search fluent translations, based on the lexical hints of the disfluent sentence, from a large target text corpus or from the Web. Therefore, all candidates will be fluent ones. The best hypotheses re-ranked best by the SPE model will then serve as the post-edited version of the disfluent sentence. Sometimes, a searched sentence may not have a high translation score to justify itself as an appropriate translation. For instance, the target sentence pattern may be correct but different lexical choices have been made. In this case, automatic local editing is applied to the weakest alignments to incrementally patch the target sentence pattern with right target lexical items. By combining the grammatical (and fluent) sentence pattern of the searched sentence and the right lexical items from the disfluent sentence, the disfluent translation could be repaired to a fluent one incrementally. This may include some local insertion, deletion and lexical substitution operations over phrase pairs that are unlikely to be translation of each other.

To really improve the fluency incrementally, the local editing process is applied in a manner that will monotonically increase the likelihood of the incrementally repaired sentence. To respect the target grammar further, the repair is phrase-based. In other words, phrase-based n-gram language model (n=1) is used in the translation score so that the likelihood of the repaired target sentence is incrementally increased during the local editing process.

In parallel with the development of our work, a few APE systems were also proposed [7, 20, 21, 8] with good results. Publicly available SMT systems (like Portage PBMT, Moses, etc.) are used directly as the post-editing module. They are trained using human post-edited target sentences with their un-edited MT outputs to learn the translation knowledge between disfluent ('source') and fluent ('target') sentences [20]. Alternatively, they may be trained using standard parallel corpora (Europarl, News Commentary, Job Bank, Hansard, etc.) where the disfluent sentences are generated using a rule-based MT (like SYSTRAN) or other SMT [21].

Therefore, these works require substantial human post-editing costs to train the SMT. Or they need a sizable parallel corpus for training, which may not be available to many language pairs. In addition, it requires an RBMT or SMT pre-trained for translating the source corpus, which may not be available to many language pairs. Most importantly, these frameworks use

the same decoding process as well as the TM and LM of the original SMT to generate their post-editing hypotheses. Therefore, the previously discussed performance issues that apply to classical SMT will also apply to such APE modules. The cascade of an SMT as an APE module might imply the use of a system with low BLEU performance to correct the outputs with low BLEU scores. The improvement could thus be substantially limited. This may be seen from the fact that the contribution of the APE becomes negligible as the training data is increased [21].

In contrast, we discard the stochastic decoding process, which might generate disfluent hypotheses, but search a large corpus for highly similar sentences to the disfluent sentence, and thus will have raw hypotheses with high BLEU scores. Additional local editing will further improve the fluency. Furthermore, our proposal can generate interesting error patterns automatically using the target language corpus alone. Therefore, the APE module can be constructed without a real MT system (although it would be better to have one in order to correct the specific errors of a specific system.). The following sections will discuss the formulation in more details.

2 Problem Formulation for SPE

In our work, we propose to adopt a Statistical Post-Editing (SPE) Model to translate disfluent sentences into fluent versions. Such a system can be regarded as a "disfluent-to-fluent" SMT. As will be seen later, it can be trained with a Monolingual SMT Model. Given a disfluent sentence E' translated form a source sentence F, the automatic post-editing problem can be formulated as finding the most fluent sentence E* from some candidate sentences E such that:

$$\begin{aligned} E^* &= \arg\max_E \Pr(E \mid E') \\ &= \arg\max_E \Pr(E' \mid E)\Pr(E) \end{aligned} \quad (2)$$

As usual, we will refer Pr(E'|E) as the translation model (TM), and Pr(E) as the language model (LM) of the SPE model. We thus encountered the same SMT problems to formulate the TM, LM and the decoding (or searching) process.

2.1 Order-Preserved Translation Model

The automatic post-editing problem is intuitively easier than SMT since we can assumes that the disfluency is due to some local editing errors, such as mis-insertion or mis-deletion of function words, and wrong lexical choices. Under this assumption, we can formulate the TM as:

$$\begin{aligned} &\Pr(E' \mid E) \\ &= \sum_A \Pr(E', A \mid E) \\ &\approx \max_A \Pr(E', A \mid E) \\ &\approx \Pr(E', A_s \mid E) \\ &= \prod_{E_p = A_s(E'_p)} \Pr(E'_p \mid E_p) \end{aligned} \quad (3)$$

In Eqn. (3), phrase-aligned phrase pairs are represented by E'p and Ep for the disfluent and fluent versions, respectively. We assume that the most likely alignment A_s, among all generic alignment pattern A, between E' and E is an "order-preserved" or "sequential" alignment between their constituents. We further assume that this most likely alignment has much higher probability than other alignments such that we don't have to sum over all generic alignment patterns. In the post-editing context, this assumption may be reasonable if the disfluency results from simple local editing operations. In particular, if we are using phrase-based alignment, the word order within the phrases can be ignored. The order preservation assumption will be even more reasonable. We therefore assume that the TM is the product of the probabilities of sequentially aligned target phrase pairs. The phrase segmentation model for dividing E or E' into phrases will be further detailed later when discussing the target phrase-based LM. Given the segmented phrases, the best sequential alignment can easily be found using a standard dynamic programming algorithm for finding the "shortest path".

The TM for the SPE model is special in that the training corpus can be easily acquired from a large monolingual corpus with fluent target sentences. Generating a disfluent version of the fluent monolingual corpus automatically based on some error model of the translation process will make this possible. One can then easily acquire the model parameters for translating disfluent sentences into fluent ones through a similar training process for a standard SMT. In comparison with standard SMT training, which requires a parallel bilingual corpus, the monolingual corpus is much easier to acquire.

2.2 Target Phrase-Based Language Model

To respect the fluency of the target language in the decoding process, the language model score Pr(E) should be evaluated based on long target language phrases, Ep, instead of target words. The "phrases" should also be defined independent of source-language in order not to introduce a huge number of noisy phrases as PBSMT normally did. The proposed LM for the current SPE, which is responsible for selecting fluent target segments, is therefore a phrase-based *unigram* model, instead of the widely used word-based *n-gram* model. In other words, we have

$$\Pr(E) = \prod_{Ep \in E} \Pr(Ep).$$

To avoid source-language dependency, we also decided not to define target phrases in terms of chunks of bilingually aligned words. Instead, the best target phrases are directly trained from the monolingual target corpus by optimizing the phrase-based unigram model. In other words, the best phrase sequence $\vec{p}^*$ for an n-word sentence w_1^n, will be the sequence, among all possible phrase segmentation, p_1^m, such that:

$$\vec{p}^* = \arg\max_{p_1^m} \Pr\left(p_1^m \mid w_1^n\right) = \arg\max_{p_1^m} \prod_i \Pr\left(p_i\right).$$

Fortunately, extracting monolingual phrases using the phrase-based uni-gram model can be done easily. The training method is just like the word based uni-gram word segmentation model [4], which was frequently used in Chinese word segmentation tasks. Unsupervised training is easy for this. Upon convergence, a set of well-formed phrases can be acquired. (This set of phrases will be called a phrase example base, PEB. Phrases in the PEB will be used later in the Local Editing Algorithm for post-editing.)

Since a phrase trained in this way can be longer than a 3-gram pattern, the modeling error could be reduced to some extend. Furthermore, the number of such phrases will be much smaller than those randomly combined phrases acquired from word-aligned word chunks. As a result, the estimation error due to data sparseness will be significantly reduced too. Unlike the rare parallel bilingual training corpus, the amount of such target language corpora is extremely large. Therefore, fluent phrases can be extracted easily. With phrases as the basic lexical unit, SPE model will reduces to

$$E^* = \arg\max_{E} \prod_{Ep=As(Ep')} \Pr(Ep' \mid Ep) \Pr(Ep) \quad (4).$$

Since a phrase can cover more than 3 words, the selected phrases might be more fluent than word trigrams. Such phrases will fit target grammar better and therefore will prefer more fluent target sentences in general.

2.3 Search-Based Decoding for Fluency

One key issue that causes disfluency is the decoding process used in classical SMT. Most decoding process regard target sentence generation as a stochastic process, and only local context of finite length window is consulted while decoding. Therefore, the target sentences generated in this way are usually not fluent. Our work proposes to search fluent translation candidates from a huge target sentence base or from web documents, instead of using traditional decoding methods to generate the translation candidates. Since the large corpus and the Web documents are produced by native speakers, the target sentences thus searched are most likely fluent with high BLEU scores.

Our current work simply used a heuristic matching score to extract a set of candidate sentences for a disfluent sentence. The candidates are then re-ranked using the translation score defined by the SPE model. The best candidate will be regarded as the post-edited version of the disfluent sentence if the translation score is higher than a threshold. Otherwise, it will be locally edited to incrementally increase its translation score. The matching score is simply the number of identical word tokens in two sentences, which is normalized by the average length of the two sentences. In other words, it is the percentage of word matches between two sentences.

We searched the candidate translations from the Academia Sinica Word Segmentation Corpus, ASWSC-2001 [6], as well as Chinese webpages indexed by Google. (We assume that the target language is Chinese.) Different query strings will result in different returned pages. Totally, we have tried 4 models for searching:

(1) Model **C**: search the corpus (only) for Top-N hypotheses (N=20). (The length difference must not be greater than two words.)
(2) Model **C+W**: search the corpus and the web for additional N hypotheses by submitting the complete disfluent target sentence as-is to Google.
(3) Model **C+W+P**: including partial matches against substrings of the disfluent target sentence, where 1~L-1 words in the disfluent sentence are successively deleted and then submitted as query strings to the search engine. (L: number of words in disfluent sentence)
(4) Model **C+W+Q**: adjacent words in the deleted disfluent sentence are quoted as a single query token before submission so that the search engine will match more exactly.

Even with such a heuristic search, a substantial number of fluent sentences similar to the disfluent sentences can be found for re-ranking and local editing.

2.4 Local Editing

If exact translation is found during searching, the searching process itself is exactly a perfect translation process. If highly similar sentences are found, simple lexical substitution or automatic post-editing [9, 11] might patch the searched fluent sentences into correct translations. Some previous works for automatic post editing have been restricted to special function words, such as the English article 'the/a' [9, 10], the Japanese case markers and Chinese classifier or particle 'de' [18]. The automatic post-editing model here is intended to resolve general editing errors that are frequently made by a machine translation system.

Briefly, the best sentence E^*_{eb} in the searched candidates will be output as the translation of the disfluent translation E' if the translation score associated with the SPE model is higher than a threshold. (The set of candidate translation sentences is called its example base, thus the subscript 'eb'.) Otherwise, the automatic local editing algorithm will find the weakest phrase alignments and fix them one-by-one to maximize the translation score.

An alignment phrase pair <Ep', Ep> is said to be "weak" if its local alignment score Pr(Ep'|Ep) x Pr(Ep) is small and thus contributes little to the global translation score for the sentence pair <E', E>. When the weakest pair, (Ep'- | Ep-) with the lowest local alignment score is identified, we should try to replace Ep-, the "most questionable phrase" in the fluent (yet incorrect) example sentence E, with some candidates that would make the patched example sentence more likely to be the translation of E'.

There are some reasons why the alignment (Ep'- | Ep-) is the weakest. First of all, Ep- might not be the right phrase, and should be replaced by Ep'- to make the fluent sentence E also the correct translation of E'. Second, Ep'- might not be the correct translation of some source phrase. In this case, the most likely translation(s) of Ep'-, called Ep+, should be used to replace Ep-. Third, Ep- is a more appropriate phrase than Ep+. In this case, it should be retained and next weakest alignment pair be repaired.

As a result, potential candidates for replacing Ep- will include Ep'-, Ep+ and Ep- itself. The best substitution will be the phrase that maximizes Pr(Ep'|Ep) x Pr(Ep). Actually, many phrases in the PEB can be a more fluent version of Ep'-. Currently, the 20 best matches will play the role of Ep+ during local editing. And the local editing algorithm will successively edit weaker alignments until the (monotonically increasing) translation score is above some threshold. The algorithm is outlined as follows.

Local Editing Algorithm

Input : E' and E^*_{eb}

Step 1 : Find the weakest alignment entry in E' from the $< E', E^*_{eb} >$ alignment.

$$Ep'- = \arg\min_{Ep' \in E'} \Pr(Ep' \mid Ep) \Pr(Ep)$$

Step 2 : Identify $Ep-$ that is the phrase in E^*_{eb} aligned with $Ep'-$.

$$Ep- = \text{align}(Ep'-)$$

Step 3 : Find the fluent phrase $Ep+$ of $Ep'-$ from PEB.

$$Ep+ = \text{PEB}(Ep'-)$$

Step 4 : Select the best substitution among Ep'-, Ep+ and Ep- which maximize the translation score:

$$E^*_{ps} = \arg\max_{E_{ps} \in \{Ep'-, Ep+, Ep-\}} \Pr(E' \mid E) \Pr(E)$$

Step 5 : Cut $Ep-$ from E^*_{eb} and paste Eps to E^*_{eb}.

$$E^*_{eb} = E^*_{eb} - (Ep-) + (Eps)$$

(Repeat until the translation score Pr(E'|E)xPr(E) reaches some threshold.)

Constrained Decoding

Note that, local editing is applied only to a local region of the example sentence based on the disfluent sentence. Intuitively, those sentences searched from a text corpus or from the Web corpus will be much more fluent than stochastically combined sentences from the SMT decoding module. Even if local editing is required, the repair will be quite local. The search space for repairing will be significantly constrained by words in the most likely example sentence. Such a searching and local editing combination can thus be regarded as a *constrained decoding*. The searching error can thus be reduced significantly in comparison with the large search space of the decoding process of a typical SMT.

2.5 Generating Faulty Sentences

The TM parameters can actually be trained from an E'-to-E monolingual Machine Translation System, where E' can be derived by applying to E some commonly found editing operations in the SMT translation process. The operations might include the insertion of target specific lexicon, deletion of source specific lexicon, local reordering of words and substitution of lexical items.

In the current work, we apply three kinds of editing operations to the fluent sentences in a monolingual corpus to simulate frequently found errors in an MT system. The fluent and its disfluent versions are then phrase segmented so that the sentences are represented by phrase tokens (instead of word tokens). Such fluent-disfluent (E-E') target sentence pairs are then trained using the GIZA++ alignment tools [12, 13, 14, 15]. Upon convergence, the translation model between the sentences to be post-edited and their correct translation can readily be acquired.

The three editing operations include:

(1) **Insertion**: The insertion errors will occur when an MT system translates a source word into a target word while it should not be translated. For instance, the English infinitive "to" need not be translated into any Chinese word most of the time. But the bilingual dictionary may indicate the possibility to translate it into "去" (chu). We therefore automatically insert the Chinese words to simulate such an error.

(2) **Deletion**: The deletion error occurs when a target specific word is not generated in the translation. For instance, the Chinese classifiers have no correspondence in the English language. We therefore delete the following classifiers from fluent Chinese sentences to create instances with deletion errors: '個', '隻', '枝', '位', '顆', '棵'.

(3) **Substitution**: When a translation system chooses a wrong lexical item, a typical substitution error will occur. To simulate the substitution errors, Chinese words in the fluent sentences are lookup against an English-Chinese dictionary. Chinese words that are also the translation of the English word are then substituted to simulate the substitution error. For instance, '問題' is a Chinese translation for the English word 'problem'. But 'problem' also has other translations, like '習題' and '疑難'. These words are therefore used to simulate the substitution errors. In our simulation, the top-30 most frequently used Chinese words are adopted to simulate the substitution errors.

With disfluent sentences created from fluent sentences with the above frequently encountered translation errors, an automatic statistical post-editing model can readily be trained using state-of-the-art alignment tools.

3 Experiments

To see the performance of the current SMT-based SPE model, about 300,000 word segmented Chinese sentences from the Academia Sinica [6] was used as our target sentence corpus. The corpus has about 2,450,000 word tokens, and the vocabulary size is about 83,000 word types. 10% of the sentences are used as the test set and 90% are used for training. The 3 types of errors are applied to the testing sentences independently. For each error type, 100 sentences are randomly selected for evaluating automatic post editing.

The performance is evaluated in terms of two criteria. The first criterion is the number (percentage) of fully corrected disfluent sentences from the test set. By fully corrected, we mean that the sentence corrected by the statistical post editing (SPE) system is completely the same as its original fluent version. Table 1 indicates the performance in terms of the error correction capability.

Error types	Searching Models			
	C	C+W	C+W+P	C+W+Q
Substitution	21	23	32	34
Deletion	28	39	46	62
Insertion	40	43	47	47
Average	30	35	42	48

Table 1. Number of fully corrected sentences with different searching models (N=100)

Note that, even with the very simple minded searching method, the SPE was able to correct, on average, about 48% of the faulty sentences to their fluent version if the search space is sufficiently large (with the C+W+Q searching model). The performance increases with the search space. And the performance is increased at most by 62%, 121% and 17.5 %, respectively for the substitution, deletion and insertion errors when the Web corpus is included to the search space. Obviously, the substitution is the hardest to resolve while insertion error seems to be easier to resolve.

The second evaluation criterion is the improvement in the BLUE score with respect to the un-corrected test sentence. Table 2 shows the BLEU scores for the various searching models. The first column labeled as E'(ts) lists the BLEU scores for the test sentences that has not been post-edited. By searching for fluent translation and applying local editing, the BLEU scores are improved with increasing search space. The best performance is to increase the BLEU scores by 15%, 38% and 26% respectively for the three types of errors. On average, the improvement is about 26%, which is substantial. On the other hand, the absolute changes are 9.4, 22.8 and 16.9 points in BLEU score, respectively.

Error Types	BLEU Scores				
	E'(ts)	C	C+W	C+W+P	C+W+Q
Substitution	0.637	0.656	0.676	0.737	0.731
Deletion	0.598	0.686	0.750	0.781	0.826
Insertion	0.646	0.762	0.780	0.810	0.815

Table 2. BLEU Scores for Various Searching Models

Note that, with search-based decoding, the absolute BLEU scores are much higher than automatic post editing systems that simply cascade a classical SMT module to the output of an MT/SMT [20, 21, 8]. Although the experiment settings are not the same and thus cannot be compared directly, the results to have higher absolute BLEU scores can be expected since searched sentences are almost always fluent, whether they are post-edited or not.

Obviously, with the same training corpus, the search space and the searching method play important roles in improving the performance. The inclusion of the web corpus does improve the performance significantly. It was reported in [19] that well formulated query strings can effectively improve searching accuracy. Therefore, by using better searching strategy, part of the translation problems for fluent translation might be resolved as a searching and automatic post-editing problems. Currently, a statistical searching model specific for the fluency-based decoding is being developed.

4 Concluding Remarks

In this paper, we propose not to generate sentence hypotheses for APE systems by using conventional SMT decoding process, since such a decoding process tends to lead to an open-ended search space. It is not easy to generate fluent sentence hypotheses under such circumstances due to the large search error. We propose to search sentence hypotheses, from a large target text corpus or from the web, based on the words in the disfluent translations, since the potential candidates will mostly be fluent. A statistical post-editing model is also proposed to re-rank the searched sentences, and a local editing algorithm is proposed to automatically recover the translation errors when the searched sentence is not a good

translation. With the SPE, the local editing algorithm tries to maximize the translation score for each local editing. It therefore improves the translation fluency incrementally. Since the TM can be trained from an automatically generated fluent-disfluent parallel corpus, training such a system is easy. The evaluation shows that, on average, 46% of translation errors can be fully recovered, and the BLEU score can be improved by about 26%. The absolute BLEU is also high with the search-based decoding process in comparison with conventional decoding process.

References

[1] Brown, Peter F., J. Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, "A statistical approach to machine translation." *Computational Linguistics*, 16(2):79–85, 1990.

[2] Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation." *Computational Linguistics*, 19(2):263–311, 1993.

[3] Chang, Jing-Shin and Chun-Kai Kung, "A Chinese-to-Chinese Statistical Machine Translation Model for Mining Synonymous Simplified-Traditional Chinese Terms," *Proceedings of Machine Translation Summit XI*, pages 81-88, Copenhagen, Denmark, 10-14, September, 2007.

[4] Chiang, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin and Keh-Yih Su, "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of ROCLING-V*, pp. 123-146, Taipei, Taiwan, R.O.C., 1992.

[5] Chiang, David, "A Hierarchical Phrase-Based Model for Statistical Machine Translation," *Proc. ACL-2005*, pages 263–270, 2005.

[6] CKIP 2001, *Academia Sinica Word Segmentation Corpus*, ASWSC-2001, (中研院中文分詞語料庫), Chinese Knowledge Information Processing Group, Acdemia Sinica, Tiapei, Taiwan, ROC, 2001.

[7] Dugast, L., J. Senellart, P. Koehn, "Statistical Post-Editing on SYSTRANS's Rule-Based Translation System," *Proceedings of the Second Workshop on Statistical Machine Translation*, 2nd WSMT, pp. 220-223, Prague, Czech Republic, June 2007.

[8] Isabelle, P., G. Goutter, M. Simard, "Domain Adaptation of MT Systems through Automatic Post-Editing," *Proceedings of MT Summit XI*, pp. 255-261, Copenhagen, Denmark, 10-14 Sept. 2007.

[9] Knight, Kevin, and Ishwar Chander, "Automated Post-Editing of Documents," in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 779-784, CA, USA, 1994.

[10] Lee, J., "Automatic Article Restoration," in *Proc. HLT-NAACL 2004 Student Research Workshop*, Boston, MA, 195-200, May, 2004.

[11] Llitjós, Ariadna Fontós, and Jaime Carbonell, "Automating Post-Editing to Improve MT Systems," in *Automated Post-Editing Workshop*, AMTA, Boston, USA, August 12, 2006.

[12] Och, Franz Josef, Christoph Tillmann, and Hermann Ney, "Improved Alignment Models for Statistical Machine Translation," in *Proc. EMNLP/WVLC*, 1999.

[13] Och, Franz Josef and Hermann Ney, “A comparison of alignment models for statistical machine translation.” In *Proc. COLING ’00: The 18th International Conference on Computational Linguistics*, pages 1086–1090, Saarbrücken, Germany, August, 2000.

[14] Och, Franz Josef and Hermann Ney, “Improved statistical alignment models.” In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447, 2000.

[15] Och, Franz Josef and Hermann Ney, “The alignment template approach to statistical machine translation.” *Computational Linguistics*, 30:417–449, 2004.

[16] Papineni, K., S. Roukos, T. Ward, and W. J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” In *Proceedings of ACL-2002*, 40th Annual Meeting of the Association for Computational Linguistics pp. 311—318, 2002.

[17] Shen, Wade, Brian Delaney and Tim Anderson, “The MIT-LL/AFRL IWSLT-2006 MT System,” *Proc. of the International Workshop on Spoken Language Translation (IWSLT) 2006*, pp. 71-76, Kyoto, Japan, 27 November 2006.

[18] Shia, Min-Shiang, *Using Phrase Structure and Fluency to Improve Statistical Machine Translation*, Master Thesis, Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC, June, 2006.

[19] Shih, Shu-Fan, *A Query Augmentation Model for Answering Well-Defined Questions*, Master Thesis, Department of Computer Science and Information Engineering, National Chi Nan University, Taiwan, ROC, July, 2007.

[20] Simard, M., G. Goutter, P. Isabelle, “Statistical Phrase-Based Post-Editing”. *Proceedings of NAACL-HLT 2007*, pp. 508-515, Rochester, NY, April 2007.

[21] Simard, M., N. Ueffing, P. Isabelle, R. Kuhn, “Rule-Based Translation with Statistical Phrase-Based Post-Editing”. *Proceedings of the Second Workshop on Statistical Machine Translation*, 2nd WSMT, pp. 203-206, Prague, Czech Republic, June 2007.

[22] Zhou, Yu, Chengqing Zong, and Bo Xu, “Bilingual Chunk Alignment in Statistical Machine Translation,” In *Proceedings of IEEE International Conference on Systems, Man & Cybernetics (SMCC2004)*, Hague, Netherlands, 2004.

Minimally Supervised Question Classification and Answering based on WordNet and Wikipedia

張至 Joseph Chang 顏孜羲 Tzu-Hsi Yen

蔡宗翰 Richard Tzong-Han Tsai*

元智大學資訊工程學系

Department of Computer Science and Engineering,

Yuan Ze University, Taiwan

{s951533, s940635}@mail.yzu.edu.tw, thtsai@saturn.cse.yzu.edu.tw

*corresponding author

摘要

在此篇論文中，我們提出一個自動將問題分類至現有詞網 (WordNet) 中之細分類方法。為此，我們利用維基百科之特性以及其文章標題，建立大規模語意實體分類表，包含了1,581,865個實體。為了表現我們研究之效用，我們建構了一個基於冗餘原則的自動問題回答系統，並透過所提出的問題分類方法來增進其效能。實驗結果顯示所提出的方法能夠有效地提升問題分類與回答的精確率。

Abstract

In this paper, we introduce an automatic method for classifying a given question using broad semantic categories in an existing lexical database (i.e., WordNet) as the class tagset. For this, we also constructed a large scale entity supersense database that contains over 1.5 million entities to the 25 WordNet lexicographer's files (supersenses) from titles of Wikipedia entry. To show the usefulness of our work, we implement a simple redundancy-based system that takes the advantage of the large scale semantic database to perform question classification and named entity classification for open domain question answering. Experimental results show that the proposed method outperform the baseline of not using question classification.

關鍵詞： 自動問題回答，問題分類，辭彙語意資料庫，辭網，維基百科

Keywords: question answering, question classification, semantic category, WordNet, Wikipedia.

1. Introduction

Question classification is considered crucial to the question answering task due to its ability

to eliminating answer candidates irrelevant to the question. For example, answers to person-questions (e.g., *Who wrote Hamlet?*) should always be a person (e.g., *William Shakespeare*). Common classification strategies includes semantic categorization and surface patterns identification. In order to fully benefit from question classification techniques, answer candidates should be classified the same way as questions.

Surface patterns identification methods classifies questions to sets of word-based patterns. Answers are then extracted from retrieved documents using these patterns. Without the help of external knowledge, surface pattern methods suffer from limited ability to exclude answers that are in irrelevant semantic classes, especially when using smaller or heterogeneous corpora.

An other common approach uses external knowledge to classify questions to semantic types. In some previous QA systems that deploy question classification, named entity recognition (NER) techniques are used for selecting answers from classified candidates. State-of-the-art NER systems produce near human performances. Good results are often achieved by handcrafted complex grammar models or large amount of hand annotated training data.

However, most high performance NER systems deal with a specific domain, focus on homogeneous corpora, and support a small set of NE types. For example, in the Message Understanding Conference 7 (MUC-7) NER task, the domain is "Airplane crashes, and Rocket/Missile Launches" using news reports as the corpus. There are only three NE classes containing seven sub classes: ORG, PERSON, LOCATION, DATE, TIME, MONEY, PERCENT. Notice that in the seven subclasses, only three of them are NEs of physical objects, others are number based entities. This is apparently insufficient for candidates filtering for general question answering. Owing to the need of wider range NE types, some of the later proposed NE classes construct of up to 200 sub classes, but NER systems targeting these types of fine-grained NE classes may not be precise enough to achieve high performance.

The amount of supported classification types greatly influences the performance of QA systems. A coarse-grained classification achieving higher precision, may still be weak in excluding improper answers from further consideration. A fine-grained classification may seem a good approach, but the cost of high-precision classification may be too high to produce actual gain in QA systems.

Moreover, in open domain QA, answers are not necessarily NEs nor can they be captured by using simple surface patterns. Using a small set of NE types to classify questions has its limits. We randomly analyzed 100 question/answer pairs from the Quiz-zone Web site (http://www.quiz-zone.co.uk/), only 70% of them are NEs. This shows being able to classify common nouns is still very important in developing QA systems.

In order to support more general question anwering, where the answer can be NEs and common nouns, we took the approach of using finer-grained semantic categories in an existing lexical database (i.e., WordNet). WordNet is a large scale, hand-crafted lexical ontology database widely used in solving natural language processing related tasks. It provides taxonomy of word senses and relations of 155,327 basic vocabularies that can be used as an semantic taxonomy for entity classification. However, in the later sections of this paper, we will show that WordNet leave room for improvement in question classification and

answer validation, and more entities, especially NEs, are needed to achieve reasonable coverage for answer candidates filtering.

With this in mind, we turn to Wikipedia, an online encyclopedia compiled by millions of volunteers all around the world, consisting articles of all kinds. It has become one of the largest reference tool ever. It is only natural that many researchers have used Wikipedia to help perform the QA task.

Using WordNet semantic categories and rich information from Wikipedia, we propose an minimally supervised question classification method targeting at the 25 WordNet lexicographer's files for question classification. Experimental results show promising precision and recall rates. The method involve extending WordNet coverage and producing the training data automatically from question/answer pairs, and training a maximum entropy model to perform for classification.

The rest of the paper is organized as follows. In the next section, we review related work in question classification and question answering. In Section 3 we explain in detail the proposed method. Then, in Section 4 we report experimental results and conclude in Section 5.

2. Related Work

Text Retrieval Conference (TREC) has been one of the major active research conferences in the field of question answering. The early tasks in the question answering track in TREC focuses on finding documents that contain the answer to the input question. No further extraction of exact answers from the retrieved documents is required.

In an effort to foster more advanced research, the TREC 2005 QA Task focuses on systems capable of returning exact answers rather than just the documents containing answers. Three types of questions are given, including FACTOID, LIST, and OTHER. For every set of questions a target text is also given as the context of the set of questions. LIST questions require multiple answers for the topic, while FACTOID questions required only one correct answer. Therefore, many consider LIST questions are easier.

More recent TREC QA Tasks focuses on complex, interactive question answering systems (ciQA). In ciQA Tasks, fixed-format template questions are given (e.g. What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?). Complex questions are answerable with several sentences or clauses. (e.g. United States arrested 167 people - including 26 Mexican bankers) The design of an interactive query interface is also a part of this task. In this paper, we focus on the issue of classifying questions in order to effectively identify potential answers to FACTOID and LIST questions.

More specifically, we focus on the first part of question answering task, namely identifying the semantic classes of the question (and answer) that can be used to formulate an effective query for document retrieval and to extract answers in the retrieved documents. The body of QA research most closely related to our work focuses on the framework of representing types of questions and automatic determination of question types from the given question. Ravichandran and Hovy [2002] proposed a question classification method that does not rely on external semantic knowledge, but rather classifies a question to different sets of

surface patterns, e.g. *ENTITY was born in ANSWER*, which requires *ENTITY* as an anchor phrase from the given question and impose no constraint on the semantic type of *ANSWER*. In contrast, we use a sizable set of question and answer pairs to learn how to classify a given question into a small number of types from the broad semantic types in the existing lexical knowledge base of WordNet.

In a study more closely related to our work, Ciaramita and Johnson [2003] used WordNet for tagging out-of-vocabulary term with supersense for question answering and other tasks. They discovered it is necessary to augment WordNet by employing complex inferences involving world knowledge. We propose a similar method *WikiSense*[1]that uses Wikipedia titles to automatically create a database and extend WordNet by adding new Wikipedia titles tagged with supersenses. Our method, which we will describe in the next section, uses a different machine learning strategy and contextual setting, under the same representational framework

Once the classes of the given questions have been determined, typical QA systems attempt to formulate and expand the query for each type of question or on a question by question basis. Kwok et al. [2001] proposed a method that matches the given question heuristically against a semi-automatic constructed set of question types in order to transform the question to effectively queries, and then extract potential answers from retrieved documents. Agichtein, Lawrence, and Gravano [2004] used question phrases (e.g., "*what is a*" in the question "*What is a hard disk?*") to represent the question types and learn query expansion rules for each question type. Prager et al. [2002] describe an automatic method for identifying semantic type of expected answers. In general, query expansion is effective in bringing more relevant document to the top-ranked list. However, the contribution to the overall question answering task might be marginal only. In contrast to the previous work, we do not use question types to expand queries, but rather use question types to filter and re-rank potential answers, which may contribute more directly to the performance of question answering.

Indeed, effective explicit question classification is crucial for pinpointing and ranking answers in the final stage of answer extraction. Ravichandran and Hovy [2002] proposed a method for learning untyped, anchored surface patterns in order to extract and rank answers for a given question type. However, as they pointed out, without external semantic information, surface classification suffers from extracting answer of improper class. Example shows a *where-is* question (e.g. *Where is Rocky Mountains?*) may be classified to the pattern "*ENTITY in ANSWER*" ("*Rocky Mountains in ANSWER*"), but with the retrieved text "...took photos of Rocky Mountains in the background when visiting...", the system may mistakenly identifies "*background*" as the answer. Intuitively, by imposing a semantic type of *LOCATION* on answers, we can filter out such noise (*background belongs* to the type of *COGNITION* according to WordNet). In contrast, we do not rely on anchor phrases to extract answers but rather use question types and redundancy to filter potential answers.

Another effective approach to extract and rank answers is based on redundancy. Brill, Lin, Banko, Dumais and Ng [2001] proposed a method that uses redundancy in two ways. First, relevant relation patterns (linguistic formulations) are identified in the retrieved documents, redundancies are counted. Second, answer redundancy is used to extract relevant

[1] The data of WikiSense will be made available to the public in the near future

answers. Distance between answer candidates and query terms are also considered in the proposed method through re-weighting. In our QA system, we use a similar approach of answer redundancy as our base line.

In contrast to the previous research in question classification for QA systems, we present a system that automatically learns to assign multiple types to a given questions, with the goal of maximizing the probability of extracting answers to the given question. We exploit the inherent regularity of questions and more or less unambiguous answer in the training data and use semantic information in WordNet augmented with rich named entities from Wikipedia.

3. Proposed Methods

In this section, we describe the proposed method for supersense tagging of Wikipedia article titles, minimally supervised question classification, and a simple redundancy based QA system for evaluation.

3.1 Problem Statement and Datasets

We focus on deploying question classification to develop an open domain, general-purpose QA system. Wikipedia titles, Wikipedia categories and YAGO are used in the process of generating WikiSense. For question classification, the 25 lexicographer's files in WordNet (supersenses) are used as the targeting class tagset. Both WordNet and WikiSense are used to generate the training data for classifying questions.

person	cognition	time	event	feeling
communication	possession	attribute	quantity	shape
artifact	location	object	motive	plant
act	substance	process	animal	relation
food	state	phenomenon	body	group

Table 1. The 25 lexicographer's files in WordNet, or supersenses.

At run time, we continue to use both WikiSense and WordNet for answer candidates filtering. Either the Web is used as the corpus, and Google is used as the information retrieval engine.

```
1)   Generate Large Semantic Category from Wikipedia titles (WikiSense)
         (Section 3.2.1)
2)   Training of Question Classifier using WikiSense and WordNet
         (Section 3.2.2)
3)   Redundancy QA System with Question Classification
         (Section 3.3)
```

Fig 1. Out line of the proposed method for QA system construction

3.2 Training Stage

The training stage of the proposed QA system consists of two main steps: generation of large scale, semantic category using Wikipedia (WikiSense) and training of fine-grained question classifier using WikiSense and WordNet. Figure 1 shows the steps of our training process and QA system.

3.2.1 Automatic Generation of Large Scale Semantic Category from Wikipedia

In the first stage of the training process (Step (1) in Figure 1), we generate a large scale, finer-grained supersense semantic database from Wikipedia. Wikipedia currently consists of over 2,900,000 articles. Every article in Wikipedia is hand tagged by volunteers with up to a few dozens of categories. There are 363,614 different categories in Wikipedia, some used in many articles, while many are used in only a handful of articles. These categories are a mixed bag of subject areas, attributes, hypernyms, and editorial notes. In order to utilize the information provided in Wikipedia categories, Suchanek, Kasneci, and Weikum [2007] developed YAGO as an ontology with links from Wikipedia categories to WordNet senses, thereby resolving the ambiguities that exist in category terms (e.g., *Capitals in Asia* is related to *capital city*, while *Venture Capital* is related to *fund*).

Although YAGO only covered 50% (182,945) of the Wikipedia categories, these categories covers of substantial part of Wikipedia articles. By using this characteristic in combination with YAGO, we use voting to heuristically determines which of the 25 WordNet lexicographer files the titles belongs to. Figure 2 shows the algorithm for categorizing Wikipedia titles using its Wikipedia categories and YAGO.

```
procedure WikiSense(Wikipedia, YAGO, WordNet)

      Declare Tags as list
      Declare Results as list

      for each Article in Wikipedia:
           Title := title of Article
(1)        Initialize Vote as an empty dictionary
           for each Category in Article:
(2)             if Category is supported by YAGO:
(3a)                 WordNetSense = YAGO(Category)
                     Append WordNetSense to Tags
(3b)                 WordNetSuperSense = WordNet(WordNetSense)
(4)                  Vote[WordNetSuperSense]++

           Class := superSense with most votes in Vote
(5)        append <Title, Class, Tags> to Results
(6)   return Results
```

Fig 2. Generation of WikiSense using Wikipedia titles/categories and YAGO

For every articles in Wikipedia, we use a dictionary to keep track of which supersense has the highest votes (Step (1)). In Step (2), all the category in the article are checked if they are supported by YAGO. Supported categories are than transformed to WordNet senses through YAGO in Step (3a). The transformed senses are than transformed again by WordNet to its corresponding supersense in Step Step (3b), and the supersense is voted once (Step (4)). Once all categories has been checked, title and its supersense with the highest votes is recorded, we also recored all the transformed WordNet senses for future uses (Step (5)). After all the articles in Wikipedia are processed, all the recorded results are returned in Step (6). In the entire process, WordNet is only used to transform a word sense to its supersense (lexical file).

We show the classification process and results of three example titles in Wikipedia in Table 2. None of these titles are in the WordNet vocabulary.

Wiki Title	Zenith Electronics
Categories	Consumer_electronics_brands, Electronics_companies_of_the_United_States, Companies_based_in_Lake_County_Illinois, Amateur_radio_companies, Companies_established_in_1918, Goods_manufactured_in_the_United_States
Senses	company#1 (3), electronics_company#1 (1), good#1 (1), 1:trade_name#1 (1)
Supersense	**noun.group (4)**, noun.attribute (1), noun.communication (1)
Wiki Title	Paul Jorion
Categories	Consciousness_researchers_and_theorists, Artificial_intelligence_researchers, Belgian_writers, Belgian_sociologists, Belgian_academics
Senses	research_worker#1 (2), writer#1 (1), sociologist#1 (1), academician#3 (1)
Supersense	**noun.person (5)**

Wiki Title	Hsinchu
Categories	Cities in Taiwan
Senses	city#1 (1)
Supersense	**noun.location (1)**

Table 2. Example of Wikipedia titles classification for generating WikiSense

3.2.2 Minimally Supervised Question Classification

In the second and final stage in the training process (Step (2) in Figure 1), we use WordNet and the previously introduced WikiSense to automatically create training data. Figure 3 shows the training algorithm for constructing question classification method. We use the Maximum Entropy Model to construct a single classifier with multiple outcomes (Step (1)). The input of this stage includes a semantic database to determine the outcomes and a set of question/answer pairs. For each question/answer pairs, we first determine whether the answer is listed in the input semantic database, unsupported question/answer pairs are neglected (Step (2)). In Step (3), a listed answer is transform into its supersense using semantic database as outcome(Step (3a)), features are extracted from question (Step (3b)). Finally, extracted features and transformed outcome is used as an event to train the classifier in Step (4). After all the listed question/answer pairs has been processed, the trained classifier is returned.

```
procedure QC Train(SemanticCategory, QASet)

(1)         Declare Classifier as Maximum Entropy Model

            for each <Q, A> in QASet:
(2)                 if A is not supported by SemanticCategory:
                        continue
(3a)                Outcome  := SemanticCategory(A)
(3b)                Features := ExtractFeatures(Q)
(4)                 Classifier.AddEvent(Features, Outcome)

            Classifier.Train()
(5)         return Classifier
```

Fig 3. Minimally Supervised training method of question classifier.

Most of the concepts in WordNet are basic vocabularies. Only few name entities can be found in WordNet, whereas Wikipedia contains a large amount of NEs. For instance NEs like “Charles Dickens” (writer) is in both WikiSense and WordNet vocabulary, while “Elton John” (singer), “Brothers in Arms” (song) or “Ben Nevis” (mountain) can only be found in WikiSense. However, WordNet, being handcrafted, still have much higher accuracy on basic words and phrases. Therefore we use both WikiSense and WordNet to cover common nouns as well as NEs.

There are three main features used in the training stage: (1) the supersense of NEs

found in the given question (2) the question phrase of the given question (3) any words in the given question.

Question	Named Entity Class	QuestionPhrase
In kilometres, how long is the Suez Canal?	noun.artifact	how-long
The action in the film "A View To A Kill" features which bridge?	noun.communication	which-bridge
Which famous authour was married to Anne Hathaway?	noun.person	which-author

Table 3. Example questions and features

At runtime, classification outcomes with probability higher than a threshold are retrieved. The value of the thresholds are set to a number of multiples uniform-distribution probability. In Section 4, we show experimental results of the proposed methods performed at different threshold.

3.3 Redundancy Based Question Answering System

We use Google as our document retrieval engine to search the entire Web. Only the snippets of the top 64 retrieval results are used. After retrieving snippet passages, we take advantage of the large amount of retrieved text to extract candidate and rely on redundancy to produce the answer. Previous work shows that answer redundancy is an effective technique for the QA task (Brill et al. [2001]).

Once answer candidates are extracted and redundancy counted, candidates are re-ranked based on question classification results. We retain and make use of several predicted question types (with probability higher than a threshold), in other words, the given question may be classified to multiple classes. This is reasonable due to the characteristic of our class tagset. Consider the question "*Where were Prince Charles and Princess Diana married?*". It may be answered with either name of a city (*London),* or name of a church (*St Paul's Cathedral)*, therefore the question type could be either LOCATION or ARTIFACT. After the passages are retrieved, answer candidates are extracted and classified using WordNet and WikiSense. Finally, we re-rank the 20 most frequent candidates by order the candidates in descending order of question type probability, and then by frequency counts. Finally, we produce the top n candidates as ouput.

4. Experimental Results Evaluation

In this section r, we describe experimental settings and evaluation results. In Section 4.1, we describe in detail the experimental settings and evaluation matrices. Then evaluation results and analysis of WikiSense and question classification are discussed in Section 4.2 and Section 4.3. Finally, we report the performance of the classifier on a simple redundancy based QA system and evaluate its effectiveness in Section 4.4.

4.1 Experimental Setting and Evaluation Matrices

In the first experiment, we explain and analysis the result and coverage of WikiSense, which is then used in the second experiment to classify questions in addition to WordNet.

We collected 5,676 question/answer pairs as the training data from the Quiz-zone Web site (http://www.quiz-zone.co.uk/), an online quiz service with popular culture and general knowledge questions designed to be answered by human. To evaluate our method, one tenth of the question/answer pairs is separated from the training data as the evaluation data. Correct classes of the questions are labeled by human judges in order to evaluate the performance of question classification.

We then used the proposed minimally supervised training method to generate two question classifiers based on different database setting. In the first experiment, we only used WordNet to generate data to train the first classifier (baseline), and then compared the classifier with the second classier trained on both WordNet and WikiSense. The purpose is to show the amount of improvement contributed by WikiSense, if any. Since WordNet is constructed by human, we consider it to have higher precision. Therefore, WordNet is used when conflicting arises between WikiSense and WordNet. The results of both classifier are presented and compared in term of recall and precision rates.

4.2 WikiSense

An implementation of the proposed method classifies about 55% of all titles in Wikipedia, resulting a large scale, finer-grained, supersense semantic category containing 1,581,865 entities.

Unclassified titles are usually caused by articles with little or no categories so their semantic type can not be accurately determined. However, the result does not imply the classification method has low coverage. Unlike most offline encyclopedias, Wikipedia is an ongoing collaborative work. Thousands of new and unfinished articles are created by volunteers or robots daily. The Wikipedia editorial principle state that every Wikipedia article should belong to at least one category, therefore uncategorized titles usually belongs to articles still in the early stage of development (called “stubs” in the Wikipedia community).

4.3 Question Classification

In this section, we report the evaluation results on using the trained classifier to classify questions. Figure 4 shows the results of the two classifiers in terms of recall and retrieval size at different level of threshold (in multiples of 0.04, the average probability). At same recall performance, the lower retrieval size results in higher precision. As Figure 5 shows, higher precision is achieved with higher threshold, trading off recall. Notice that the recall of both classifiers gradually decreases when threshold increases from one to five times of uniform probability. Above threshold 5, recall of both classifiers decreases rapidly. Considering recall being crucial to question classification task in order to prevent early elimination of the correct answer candidates, we focus our analysis on thresholds lower than 5. We can see that the

precision increases for both classifiers as threshold increases. The combined classifier was able to achieve slightly higher recall and higher precision of 9% at threshold of 2 times of uniform probability.

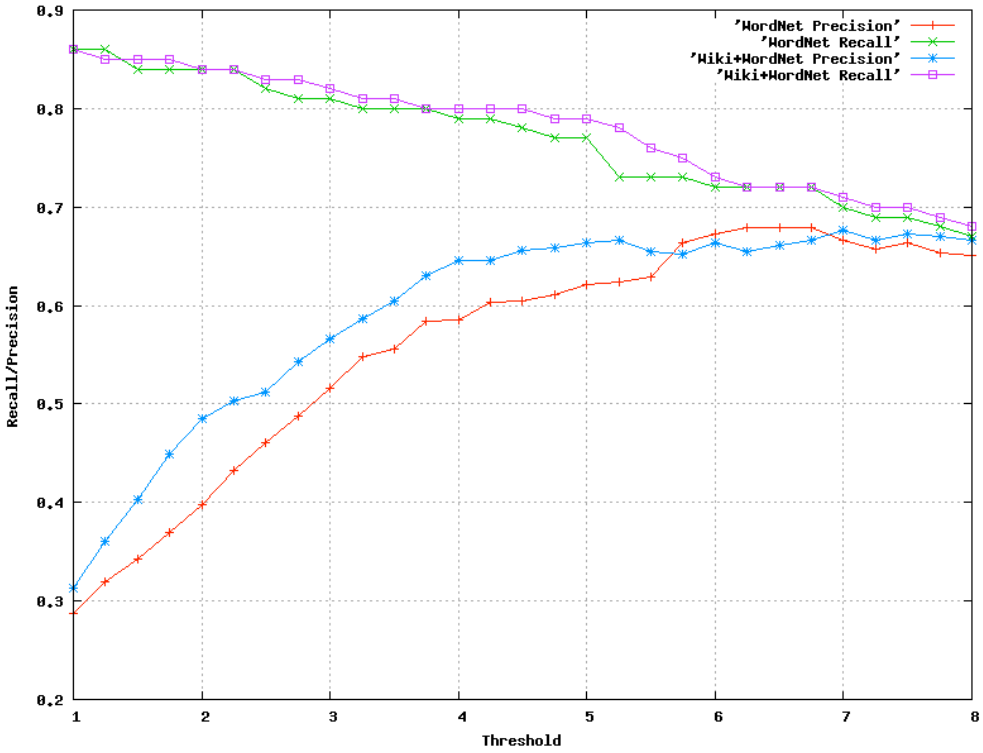


Fig 5. Performance in terms of precision and recall at different threshold.

4.4 Question Answering

In this experiment, we first run our QA system with out any question classification as our baseline. We then run the same system on the same evaluation dataset using two different question classifier, one trained by WordNet and the other trained on WordNet plus WikiSense.

Threshold	Top 1	MRR
Baseline	34%	0.451
1.0	39%	0.476
1.5	40%	0.482
2.0	42%	0.501
2.25	42%	0.503
2.5	35%	0.457

(a) WordNet

Threshold	Top 1	MRR
Baseline	34%	0.451
1.0	43%	0.492
1.5	43%	0.503
2.0	44%	0.509
2.25	43%	0.512
2.5	35%	0.457

(b) WikiSense + WordNet

Table 4. Top 1 precision and MRR result of deploying the 2 classifiers

Table 4 lists Top 1 precision and MRR of our baseline system and the system with the two classifiers at varying thresholds. As we can see, by including question classification, both systems performed better than baseline. With the enhancement of WikiSense, results in Table 4(b) achieve significantly higher MRR and top 1 precision comparing to system with a classifier trained on WordNet only (see Table 4(a)). The best performance of both MRR and top 1 precision was achieved by the system with both WikiSense and WordNet. At threshold of 2.25, the MRR was higher than the baseline by 0.061, and top 1 precision is higher by 9%.

5. Conclusions

Many future research directions present themselves. For example expanding the coverage of WikiSense using other characteristics of Wikipedia, such as internal link structure, article contents, information boxes and Wikipedia templates, minimally supervised training for automatically supersense tagging on Wikipedia title, and a more complex QA system that take full advantage of finer-grained classification.

In summary, we have introduced a method of minimally supervised training for fine-grained question classification using an automatically generated supersense category (WikiSense) and WordNet. The method involves supersense tagging of answers to generate training data, and using Maximum Entropy model to build question classifiers. We have implemented and evaluated the proposed methods using a simple redundancy based QA system. The results show the method substantially outperforms the baseline of now using question classification.

References

[1] E. Agichtein and S. Lawrence and L. Gravano, Learning to Find Answers to Questions on the Web, ACM Transactions on Internet Technology (TOIT), volume 4, pp. 129-162, 2004

[2] E. Brill and J. Lin and M. Banko and S. Dumais and A. Ng, Data-Intensive Question Answering, In Proceedings of the Tenth Text REtrieval Conference (TREC), pp. 393-400, 2001

[3] M. Ciaramita and M. Johnson, Supersense Tagging of Unknown Nouns in WordNet, Conference on Empirical Methods on Natural Language Processing (EMNLP), pp. 168-175, 2003

[4] C. Fellbaum, Wordnet: An Electronic Lexical Database, ISBN: 026206197X, May 15, 1998

[5] C. Kwok and O. Etzioni and D. S. Weld, Scaling question answering to the web, ACM Transactions on Information Systems (TOIS), Volume 19, Issue 3, pp. 242-262, 2001

[6] MetaWeb Technologies, Freebase Wikipedia Extraction (WEX) version June 16, 2009, http://download.freebase.com/wex/ , 2009

[7] J. Prager and J. Chu-Carroll and K. Czuba, Statistical answer-type identification in open-domain question answering, Proceedings of the second international conference on Human Language Technology Research, pp. 150-156, 2002

[8] F. M. Suchanek and G. Kasneci and G. Weikum, Yago - A Core of Semantic Knowledge, 16th international World Wide Web conference (WWW), 2007

[9] D. Ravichandran and E. Hovy, Learning Surface Text Patterns for a Question Answering System, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 41-47, July 2002

應用句型結構與部份樣本樹於對話行爲之偵測
Dialogue Act Detection Using Sentence Structure and Partial Pattern Trees

梁維彬、蕭育丞、吳宗憲
Wei-Bin Liang, Yu-Cheng Hsiao, and Chung-Hsien Wu

國立成功大學資訊工程學系
Department of Computer Science and Information Engineering, National Cheng Kung University
E-mail: liang@csie.ncku.edu.tw ychsiao9@gmail.com chwu@csie.ncku.edu.tw

摘要

本論文提出一使用部份樣本樹及句型結構於對話行爲之偵測。爲了建構具強健性的對話行爲偵測模型，我們針對語音辨識之輸出語句，使用部份樣本樹來產生多重候選句，以避免語音辨識錯誤所衍生句子錯誤之問題。而後再經由剖析器得到候選句所對應之語法規則。而再針對每一類對話行爲所包含的規則做句型分類，來降低對話行爲之間的混淆。最後，利用潛在對話行爲矩陣來描述語法規則和意圖之間的關係。另外，在對話系統應用中，我們採用部份觀察馬可夫決策程序從對話歷程中訓練出之最佳對話策略，以增進對話系統的可用性。在實驗中，我們建立一個旅遊資訊諮詢對話系統，作爲實際應用測試平台。而在測試時，分別就每項對話行爲做測試。相較於應用語義表格(semantic slot)方法達到之 48.1%正確率，本論文所提之方法可得到整體正確率爲 81.9%，提升了 33.8%的正確率。由實驗可知論文所提之方法在實際應用上能有明顯的效能提升。

Abstract

This paper presents a dialogue act detection approach using sentence structures and partial pattern trees to generate candidate sentences (CSs). A syntactic parser is utilized to convert the CSs to sentence grammar rules (SRs). To avoid the confusion between dialogue intentions, the *K*-means algorithm is adopted to cluster the sentence structures of the same dialogue intention based on the SRs. Finally, the relationship between these SRs and the intentions is modeled by a latent dialogue act matrix. Moreover, for the application to a travel information dialogue system, optimal dialogue strategies are trained using the partially observable Markov decision process (POMDP) for robust dialogue management. In evaluation, compared to the semantic slot-based method which achieves 48.1% dialogue act detection accuracy, the proposed approach can achieve 81.9% accuracy, with 33.3% improvement.

關鍵詞：對話行爲、部份樣本樹、句型結構、部分馬可夫決策程序

Kewords: Dialogue act, partial pattern tree, sentence structure, POMDP

一、 緒論

在本論文中，我們以建構一套旅遊資訊查詢系統爲主要目標。因此，以下我們將對國內外的針對對話系統之相關研究做一文獻回顧。首先，針對旅遊資訊對話系統的相關研究[1]，在國外方面，有美國麻省理工學院(MIT)的餐廳導覽系統[2]、美國電信電報公司

(AT&T)的線上服務系統[3]、日本國家資訊通信科技研究機構(NICT)的旅遊導覽系統[4]，以及 Philips 公司所開發的火車時刻票價查詢系統[5]。在國內方面，台大有銀行電話查詢系統[6]、交大有汽車導覽系統[7]、工研院則有智慧型總機、氣象查詢系統的實現[8]，和成大智慧型醫療服務對話系統[9]。在意圖偵測部份的相關研究，Choi 等學者[10]將對話語料中每句話標記其意圖，及對整個對話過程標記其爲對話過程的開始、結束、正在對話等，再使用機器學習(machine learning)的方法來建立其模型以判斷意圖，但這一部分的方法對於語音辨識錯誤造成系統回應錯誤的問題並未考慮。而在對話管理部份，目前的研究有應用有限狀態機(finite state machine) [4]與部份觀察馬可夫決定程序(partial observation Markov decision process, POMDP)[11]來實現。這一部分的研究主題爲系統與使用者互動中，先判斷使用者的意圖，再對此意圖作出最適當的回應。

近幾年來，口述語言對話系統已經有顯著的進步，尤其是建構於填表式(slot-filling)的資料庫查詢方法已進入應用的階段。然而，因爲語音辨識錯誤造成表格填入錯誤，進一步使得自然語言理解產生錯誤及誤判使用者的意圖，導致系統回應錯誤。這類型的問題尚未成功地解決。因此，如何有效地在具語音辨識錯誤的條件下仍能得到好的意圖偵測結果是我們的研究主要目標。除此之外，我們亦希望能得到良好的人機互動，所以一個有效的系統回應機制以避免對話發散之窘境，讓使用者不致於對系統產生排斥甚至厭惡進而提高其可行性，也是我們所考量的部份。

在本論文的其他段落安排如下。第二節描述我們爲了本實驗所蒐集的旅遊相關資訊語料及其對話、語義類別、對話行爲和對話行爲所對應的行動標記。然後，第三節介紹論文核心的對話行爲偵測模型與其訓練方法將被描述。下一段的第四節爲對話管理決策的訓練。第五節的實驗說明了對話行爲偵測器中語音辨識器元件的訓練、數種對話行爲行爲偵測比較實驗和相關統計資料。最後，結論與未來展望將被討論於第六節。

二、 語料收集

2.1 語料錄製與標記

在實驗室環境下，使用 audio-technica AT9940 數位錄音麥克風，以 16 位元 16KHz 的取樣頻率將發音人的語料錄於單聲道。錄音情境爲，發音人面對電腦自行輸入譯文、操作錄音及辨識過程，錄音計劃負責人操作修改回應的部分。總共收錄到 144 個對話回合(dialogue turn)，總數爲 1,586 句的語料。錄音完成後，以人工方式進行對話語料相關資料標計(如 Dialogue 編號和 Turn 編號)、對話行爲(dialogue act, DA)。

2.2 語義類別(Semantic Class)

在以填表(slot-filling)方式爲基礎的對話系統中，若關鍵字詞過於散亂將間接導致系統效能不彰的問題，相較之下，若將關鍵字提升到語義類別，不僅可在對話行爲偵測上可避免偵測類別過多的問題，對於語料庫的擴增與維護也有較良好的管理。因此，在語料庫的資料分析過程中，我們將標記爲關鍵字的詞彙作進一步的整理，即如表 1 中所呈現的內容。最後，我們蒐集的語料庫總共包含 27 種語義類別。

2.3 對話行爲(Dialogue Act)和其系統回應行動(Action)

當語者說出了一句話，這句話本身的文字有其意義，但語者之所以會說出這句話有各種目的。這樣就可用對話來做一個行動，這就是對話行爲。舉例來說，「請問安平古堡的

票價是多少？」這句話被表達出來的 DA 爲詢問票價。因此，語言學家稱所有類型的溝通行爲爲"對話行爲"[12]。完成語料收集後，依據系統提供的任務(task)，我們分析對話語料並且在每一種系統任務中設計了該任務所含括的表單(slot)值及每個表單可填入值，如表 2 所示。在我們所收錄的語料中，可分爲三大任務，分別爲查詢系統服務、查詢景點相關資訊和查詢交通相關資訊，其中第 j 個任務所包含的 DA 數表示爲：

$$\text{Task}_j\text{的 DA 數} = \prod_{i=1}^{\text{Task}_j\text{包含的 slot 數}} \text{Slot}_i\text{可填入值數} - 1 \quad (1)$$

其中-1 是因爲未填值可能在其他任務中有填值，例如：任務 1 的“我想查詢高鐵”和任務 3 的 “跟我說高鐵的時刻表”。 其他的意圖包括歡迎、結束、無意圖，總共 38 個。當對話系統偵測出使用者的 DA 後，對話系統應作出合理系統回應行爲以達到和使用者之間的互動。因此，我們根據 DA 的內容整理出如圖 2 最右欄位所示的系統回應。大致可分爲系統詢問未填值的表格資訊和回答資訊的行動，總共有 20 種 [13]。

表 1：語義類別範例及其對應的關鍵詞彙範例。

編號	語義類別	詞彙	編號	語義類別	詞彙	編號	語義類別	詞彙
1	日期	日,月,星期,禮拜	10	感謝語	謝謝,感謝	19	開車	開車
2	城市	台北,台中,台南	11	疑問詞	什麼,怎麼	20	高鐵	高鐵
3	地點	安平古堡,成功大學	12	肯定	好,沒錯,了解	21	火車	火車
4	…	…	13	…	…	22	…	…

表 2：系統任務分類及其表單和可填入值之對照表範例。

Task	Category	Slot	可填入值				DA個數
1	查詢系統服務	系統服務	查詢地點	查詢車站	查詢服務	未填值	4-1=3
2	查詢景點相關資訊	地點表單	有地點	無地點			2*7-1=13
		地點資訊	地址	介紹	…	未填值	
2	查詢交通相關資訊	交通方式	公車	高鐵	火車	…	5*2*2-1=19
		目的地	有目的地	無目的地			
		出發地	有出發地	無出發地			
無	其他	歡迎					3
		…					
總計							38

表 3 ：DA 列表、例句和其對應的 Action。

Task	DA 編號	DA	Example	Action(編號)
1	1	查詢服務	你有什麼服務可以查詢？	回答系統能提供的服務(1)
	…	…	…	…
2	4	地點	安平古堡。	詢問使用者想查詢地點的何種資訊(4)
	5	查詢地址有地點	我想查詢安平古堡的地址。	回答地點地址資訊(5)
	…	…	…	…
3	17	公車有出發地有目的地	怎麼搭公車從成功大學前往赤崁樓？	回答搭乘公車方式(12)
	…	…	…	…
	35	有出發地無目的地	我從台北出發。	詢問使用者想要何種交通方式(11)
無	36	歡迎	您好。	系統無任何反應行動(18-1)
	37	結束	謝謝。	系統無任何反應行動(18-2)
	38	無意圖	無此種語料	系統無任何反應行動(18-3)

三、 系統架構

本論文的系統架構如圖 1 所示，虛線以上爲使用者部份，包括發音 U 和接收系統回應訊息所產生的聲音 U′；而虛線以下爲對話系統部分，主要分爲三個部份，包括輸入處理(input processing)、對話管理(dialogue management, DM)和輸出處理(output processing)。在輸入處理部分，語者的發音 U 經由麥克風傳送至自動語音辨識器(automatic speech recognizer, ASR)並產生辨識結果 W，此一辨識結果將送至口述語言理解(spoken language understanding, SLU)單元進行潛在對話行爲偵測而得到第 c 類對話行爲 DA_C，而 SLU 也是本論文的核心，我們將逐一介紹 SLU 的訓練方法和採用的技術。然後，DA_C 將傳送到對話管理並根據由 POMDP 所訓練而得的對話策略(strategy)和對話意圖歷史記錄(dialogue act history) DA_H 來採取合適的回應(action) a_t。當系統做出回應後，系統將從我們蒐集而來的旅有資訊資料庫(travel information database)中查詢對應的資料並輸出文字 Context_t 至語音合成器(text-to-speech synthesizer, TTS)產生語音資訊 U′傳達給使用者。以下我們將逐一介紹各個部份。

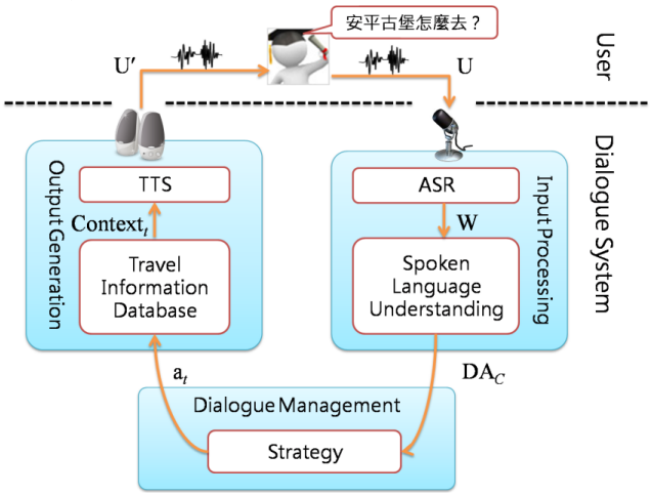


圖 1 ：對話系統架構圖。

3.1 口述語言理解(SLU)

在潛在對話行爲偵測過程中，系統必須根據使用者發音 U 和對話意圖歷史記錄 DA_H 來偵測最佳意圖 DA*，則此偵測法則(detection criterion)定義爲式子：

$$\text{DA}^* = \underset{\mathbf{DA}}{\text{argmax}}\, P(DA_C \mid \text{U}, DA_H) \tag{2}$$

其中 **DA** 爲所有可能的 DA 集合，DA_C 爲發音 U 被辨識爲第 c 類 DA。第 i 種可能辨識字串 W_i 爲 U 經過辨識後所得到的可能辨識結果文字。然而，我們僅取最佳的辨識結果 $\hat{W}$，因此式子(2)改寫爲式子(3)，進一步展開爲式子(4)：

$$\begin{aligned}\text{DA}^* &= \underset{\mathbf{DA}}{\text{argmax}} \sum_{W_i} P(DA_C, W_i \mid \text{U}, DA_H) \\ &\approx \underset{\mathbf{DA}}{\text{argmax}} \max_{\mathbf{W}} P(DA_C, W_i \mid \text{U}, DA_H) \qquad (3) \\ &= \underset{\mathbf{DA}}{\text{argmax}} \max_{\mathbf{W}} P(DA_C \mid W_i, \text{U}, DA_H) P(W_i \mid \text{U}, DA_H) \qquad (4)\end{aligned}$$

假設辨識結果 W_i 與輸入語音 U 代表相同意義且語音辨識結果 W_i 與對話意圖歷史記錄

DA_H 獨立，則我們可以得到式子(5)。此外，P($DA_C|W_i$)爲經由貝氏決策法則(Bayes' decision rule)而來，因此我們將式子進一步改寫爲式子(6)：

$$\mathrm{DA}^* = \underset{\mathbf{DA}}{\operatorname{argmax}} \max_{\mathbf{W}} P(DA_C \mid W_i) P(DA_C \mid DA_H) P(W_i \mid \mathrm{U}) \tag{5}$$

$$= \underset{\mathbf{DA}}{\operatorname{argmax}} \max_{\mathbf{W}} \frac{P(W_i \mid DA_C) P(DA_C)}{P(W_i)} P(DA_C \mid DA_H) P(W_i \mid \mathrm{U}) \tag{6}$$

其中， $P(W_i)$可被省略，$P(DA_C)$在本論文假設爲均等事前機率(equal prior)，所以最後得到式子：

$$\mathrm{DA}^* \approx \underset{\mathbf{DA}}{\operatorname{argmax}} \max_{\mathbf{W}} P(W_i \mid \mathrm{U}) P(W_i \mid DA_C) P(DA_C \mid DA_H) \tag{7}$$

其中 $P(W_i|\mathrm{U})$爲自動語音辨識器從語音 U 所得到的辨識字串 W 的機率；$P(W_i|DA_C)$爲辨識字串被偵測爲第 c 個 DA 的偵測機率(probability of DA detection)，即對話行爲偵測的部份；$P(DA_C|DA_H)$爲對話意圖歷史(dialogue history)機率，用來防止系統跳脫使用者的意圖，例如使用者正在查詢高鐵時刻表，系統卻進入旅遊景點的相關資訊查詢功能。

3.2 對話行爲之偵測(Dialogue Act Detection)

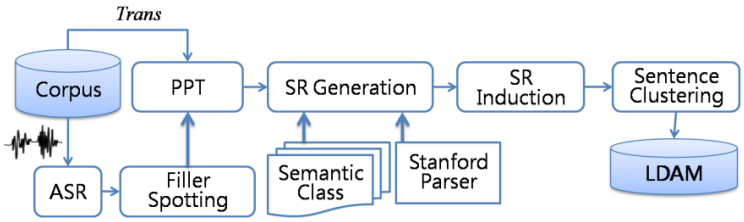


圖 2：對話行爲偵測的潛在語義矩陣模型訓練流程。

圖 2 展現了本論文 DA 偵測所使用的潛在意圖矩陣模型的訓練流程。無論文字類型的譯文檔或者譯文檔所對應的語音皆在訓練過程中被採用。爲了克服惱人的語音辨識錯誤問題，部份樣本樹用來產生多條候選句。在語句規則(sentence rule, SR)產生步驟裡，候選句會參考語義類別進行替換，然後使用 Stanford 剖析器[14]進行規則的產生。我們把語料庫中所有能產生的語句規則和所有的 DA 以矩陣模式建構兩者之間的關係，並進行歸納(induction)的動作。由於不同的 DA 之間可能擁有相同的語句規則，爲了避免 DA 之間的混淆，我們採用修改過的 K-means 演算法對同一類別 DA 之句型進行分類。我們將逐一詳細說明各流程步驟。

3.3 部份樣本樹(Partial Pattern Trees, PPT)

在本研究中，我們將對話語句視爲數個功能性詞彙(optional phrase, OP)與至少一個主要關鍵詞(main phrase, MP)之間的組合。因此我們從一句複雜的句子中擷取出部份樣本句(partial pattern, PP)，可看做類似對這句句子作部份分解，而 MP 則隱含著語句的語義，爲了保留語句的主要語義，因此每一句 PP 必需包含著 MP。相反的，OP 則可能在辨識結果中因刪除型錯誤(deletion error)而被省略。所以 PP 就是完整句子因爲某些 OP 在辨識結果中被刪除型的句子，而 PPT 是根據 PP 所建立的語句模組。

當在辨識對話語句時，一個常遭遇到的問題就是語句模組在建立時並沒有考慮到對話語

句常雜夾著一些無意義的贅語(例如：”嗯”和”喔”)和各種可能的語音辨識錯誤，這個問題進而造成 DA 偵測錯誤。在某些研究[15][16]中顯示出，一個詞彙出現在這些不流利的贅語之後的平均機率會小於出現在無贅語之後的情況。這指出我們很難去預測，當一個詞彙出現在多餘贅語之後的機率。因此根據這些觀察，我們將省略掉贅語或沒被辨識出來的詞彙而產生的句子，稱之爲 PP。更進一步探討，部份樣本樹一個很重要的應用，就是針對替代性錯誤做修正，而之前的研究對於錯誤的回復(recovery)通常都是利用句型規則在眾多的候選句中找出最符合語法句型的句子[17][18]。然而這些方法所產生的句子在句型上雖然非常符合，但是卻可能在語義上的意義是不足的。所以針對此點我們所定義的每一句部分樣本句，都需保留原句中的主要關鍵詞以維持句子的語義，然而功能性詞彙則有可能被省略。因此根據上述觀察，我們提出了利用 PP 來建立效能更佳的語句模組，在這裡我們將句子 $Trans_i$ 視爲一連串的 OP 與一 MP 的組合，表示成：

$$Trans_i = \{OP_1^i, OP_2^i, ..., OP_{NB_i}^i, MP^i, OP_{NB_i+1}^i, ..., OP_{NB_i+NA_i}^i\} \tag{8}$$

其中 NB_i 和 NA_i 分別爲在 MP 之前與 MP 之後的 OP 數。根據上述定義，PP 爲包含 MP^i 的子序列，其中每一個 OP 都有可能被省略，在本文中，被省略的 OP 我們替換成 Filler。替換成 Filler(F)的原因是我們想保持句子原本的句型，且也可以假設爲語音辨識錯誤時可能會發生的情況。舉例說明若有一句子爲”ABC”且 A，C 爲 OP 而 B 爲 MP 彙，則共有四句 PP 分別爲”ABC”，”ABF”，”FBC”和”FBF”。

3.4 填充字擷取(Filler Spotting)

即使 ASR 的辨識結果採用詞圖(word-graph)[19]爲基礎的重新計分(rescoring)方式，依然可能存在語音辨識錯誤的問題。這是因爲 word-graph 在重新計分時倘若辨識結果中夾雜著其他無助於 SLU 效能的文字，則除了從譯文檔產生 PP，我們也將譯文檔所對應的語音進行辨識以增加句型的多樣性(diversity)，彌補文字語料無法產生的句型，使得 SLU 單元能辨認更多種類的句型。所以我們對每個語音辨識結果的詞彙作填充字擷取。本論文採用統計方式的卡方檢定(χ^2-test)來進行填充字擷取。我們記錄被正確辨識詞彙所對應的分數，接著對每個詞彙計算其分數的平均值 (mean) 與標準差(standard deviation) σ。而填充字擷取的依據，我們使用卡方檢定，數學式爲：

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - \quad)^2}{\sigma} \tag{9}$$

藉此判斷語音辨識結果的每個詞彙要接受或拒絕，拒絕的詞彙我們替換成 Filler。替換成 Filler 的原因是我們想保持句子原本的句型，且也可以假設爲語音辨識錯誤時可能會發生的情況。

3.5 句型規則產生(Sentence Rule Generation)

在句型規則產生部分，首先，我們需要一個語義剖析器來處理訓練語句，並建立對應的語義樹狀結構，得到其句型規則，本研究利用史丹佛大學所研究開發的剖析器來達成此目的。史丹佛的剖析器[14]是基於 PCFG (Probabilistic Context Free Grammar) 的觀念所建立而成的剖析器。所謂的 PCFG 是一種隨機語言模型 (Stochastic Language Models, SLM)，而 SLM 的主要目的之一是根據訓練語料的統計資料來提供足夠的機率資訊以運用在語音辨識的構句處理上，不僅能有效提高的辨識正確率，更可藉由搜尋路徑的限制，節省計算時間，而應用在文句剖析上則能提供正確性較高的句法結果。關於史丹佛剖析

器主要的核心概念可以參考文獻[20][21]。圖 2 流程中的句型規則產生，在剖析前，我們先利用事先定義好的語義類別將語句中的詞彙替換成語義類別，再透過史丹佛剖析器得到剖析結果。替換的目的是降低句型規則的複雜度，讓相同語義的詞彙屬於同一條規則，例如：「NP → NN 安平古堡」和「NP → NN 億載金城」皆屬於「NP → NN 地點」這條規則。圖 3 範例是語句「怎麼去安平古堡」經過語義替換爲「疑問詞 路線 地點」，經由剖析器可得到一顆文法樹，其包含的句型規則包括：(1) Root → IP、(2) IP → VP、(3) VP → ADVP VP、(4) ADVP → AD 疑問詞、(5) VP → VV 路線 NP 和(6) NP → NN 地點。我們便以這六條規則代表這句話。

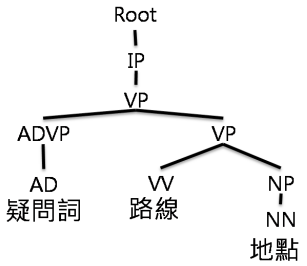


圖 3：剖析得到的文法樹範例。

3.6 句型規則歸納(Induction)

假設從所有語料中得到的句型規則可被表示爲維度爲 L 的規則向量 **Rule**，每個維度對應著一條句型規則。則此規則向量和所有的 **DA** 可構成一個矩陣來建立句型規則與意圖之間的關係，此關係可定義爲：。

$$\mathbf{\Phi}_{L\times Q}=\begin{bmatrix}\phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,Q}\\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,Q}\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{L,1} & \phi_{L,2} & \cdots & \phi_{L,Q}\end{bmatrix} \tag{10}$$

其中$\mathbf{\Phi}_{L\times Q}$是維度爲 $L\times Q$ 的文法結構資訊矩陣，L 代表訓練語料所有句型規則的個數，Q 代表意圖的總數。矩陣中每個元素 $\phi_{l,q}$ 代表著第 l 條文法規則 $Rule_l$在第 q 個 DA 中所佔的重要性。因此本研究中定義 $\phi_{l,q}$的估計法如下：

$$\phi_{l,q}=(1-\varepsilon_l)P(Rule_l \mid DA_q) \tag{11}$$

其中，$P(Rule_l \mid DA_q)$是該條規則佔該句語法結構的比重，該項可以寫爲：

$$P(Rule_l \mid DA_q)=\frac{C(Rule_l, DA_q)}{\sum_k C(Rule_k, DA_q)} \tag{12}$$

且 C($Rule_l$, DA_q)表示句型規則 $Rule_l$ 出現在 DA_q 中的次數。另外，(1-ε_l) 是利用量度文字亂度 (Entropy) 的方法來量度某條規則在該語料中是否具有鑑別性並賦予該元素的權重，則ε_l可定義爲：

$$\varepsilon_l=-\frac{1}{\log Q}\sum_{q=1}^{Q}\frac{C(Rule_l, DA_q)}{\sum_{i=1}^{Q} C(Rule_l, DA_i)}\log\frac{C(Rule_l, DA_q)}{\sum_{i=1}^{Q} C(Rule_l, DA_i)} \tag{13}$$

3.7 語句分群(Sentence Clustering)

使用者的 DA 可能以不同語句表達，不同語句意味著他們可能蘊含著不同的句型規則，因此，同一個 DA 下可能包含著多種句型規則導致和其他意圖之間造成混淆。例如：兩段屬於同樣 DA_1 的語音分別包含語句規則{1,2,3}和{4,5,6}，另有一段屬於 DA_2 的語音包含語音規則{3,4,7}，則此語音可能會被誤判爲 DA_1。因此，爲了避免意圖之間的混淆，語句需要分群。對於傳統的分群方法，如 *K*-means 演算法，必須計算各資料點和 centroid 之間的距離。然而，句型規則的 centroid 不具有任何物理意義。因此爲適應本論文的需求，我們先將屬於第 q 個 *DA* 的第 i 個譯文檔 $Trans_i$ 表示爲：

$$\Phi_q^{Trans_i} = (\phi_{1,q}\delta_1, \phi_{2,q}\delta_2, \ldots, \phi_{L,q}\delta_L) \tag{14}$$

其中 $\phi_{l,q}$ 的定義等同於上述句型規則歸納矩陣 $\mathbf{\Phi}_{L\times Q}$ 中的 $\phi_{l,q}$ 。而 δ_l 指出若 $Trans_i$ 使用到 $Rule_l$，則其值爲 1，反之爲 0。另外，我們選擇一個特殊函式作爲最大化群之內相似度，此特殊函式表示爲：

$$(G_1, G_2, \ldots, G_K)^* = \arg\max \sum_{k=1}^{K} \sqrt{\sum_{\Phi_q^{Trans_i}, \Phi_q^{Trans_j} \in G_k} Similarity(\Phi_q^{Trans_i}, \Phi_q^{Trans_j})} \tag{15}$$

其中 K 爲欲分群之數量，G_k 表示屬於第 k 群的譯文檔集合，而 *Similarity*(•)爲兩則譯文檔之間的相似度計算，我們採用 Cosine Measure 數學式表示爲：

$$Similarity(\mathbf{\Phi}_q^{Trans_i}, \mathbf{\Phi}_q^{Trans_j}) = \frac{\mathbf{\Phi}_q^{Trans_i} \cdot \mathbf{\Phi}_q^{Trans_j}}{\left\|\mathbf{\Phi}_q^{Trans_i}\right\| \cdot \left\|\mathbf{\Phi}_q^{Trans_j}\right\|} \tag{16}$$

3.8 潛在對話行爲矩陣模型(Latent DA Model, LDAM)

經由部份樣本句與填充字擷取後，我們將得到的句型規則與意圖句型分類後的類別建立關係矩陣，而產生的句型規則比原本訓練的文字語料所產生的句型規則包含更多意涵，所以我們稱此矩陣爲潛在意圖矩陣且定義爲：

$$\mathbf{LDAM}_{L\times M} = \begin{bmatrix} \nu_{1,1} & \nu_{1,2} & \cdots & \nu_{1,M} \\ \nu_{2,1} & \nu_{2,2} & \cdots & \nu_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{L,1} & \nu_{L,2} & \cdots & \nu_{L,M} \end{bmatrix} \tag{17}$$

其中意圖矩陣 $\mathbf{LDAM}_{L\times M}$ 爲一個維度爲 $L\times M$ 的文法結構資訊矩陣，L 代表訓練語料經過語句規則產生步驟後所有的句型規則個數，M 代表意圖句型群聚後的類別總數。矩陣中每個元素 ν_{lm} 代表著第 l 條句型規則在第 m 個 DA 中所佔的重要性。因此本研究中定義 ν_{lm} 的估計法如下：

$$\nu_{lm} = (1-\varepsilon_l) P(Rule_l \mid DA_m) \tag{18}$$

其中 $P(Rule_l|DA_m)$是該條規則佔該句語法結構的比重，該項可以寫爲：

$$P(Rule_l \mid DA_m) = \frac{C(Rule_l, DA_m)}{\sum_k C(Rule_k, DA_m)} \tag{19}$$

而 C($Rule_l$, DA_m)表示句型規則 l 出現在第 m 個意圖句型分類後的類別中的次數。另外，(1-ε_l)是利用量度文字亂度 (Entropy) 的方法來度量該條規則在語料中的鑑別性，當作矩陣中該元素的權重，ε_l 可定義爲：

$$\varepsilon_l = -\frac{1}{\log C}\sum_{c=1}^{C}\frac{C(Rule_l, DA_m)}{\sum_{i=1}^{C} C(Rule_l, DA_i)}\log\frac{C(Rule_l, DA_m)}{\sum_{i=1}^{C} C(Rule_l, DA_i)} \tag{20}$$

其中大寫 C 爲由上述 K-means 分群而得最終 DA 數量，而 DA_C 所表示的就是 **LDAM** 第 c 個 column 的內容。最後，我們得到 SLU 偵測語義所需要的模型。

3.9 對話行爲型態偵測

在對話行爲 DA 偵測中， $P(W_i \mid DA_C)$項因爲一個句子包含語義成分及語法成分，所以我們對語音辨識的結果進一步拆解爲：

$$P(W \mid DA_c) \approx P(\mathbf{Rule}_W \mid DA_c)P(SC_W \mid DA_c) \tag{21}$$

其中 $\mathbf{Rule}_W$ 爲記錄辨識結果 W 所使用的句型規則。則 $\mathbf{Rule}_W$ 對潛在意圖矩陣中第 c 個類別的相似度，我們採用 Cosine Measure 來計算，定義爲：

$$P(\mathbf{Rule}_W) \mid DA_c) = \frac{\mathbf{Rule}_W^T \bullet DA_c}{\|\mathbf{Rule}_W\| \times \|DA_c\|} \tag{22}$$

而 SC_W 爲將辨識結果 W 轉換爲語義類別的函式。在語義成分分數的計算，我們經由統計文字語料，得到每個對話行爲 DA 出現每個意圖類別的機率，數學式可寫爲：

$$P(\mathrm{SC}_W \mid DA_c) = \prod_{w_n \in W} P(\mathrm{SC}_{\mathrm{j}}^{w_n}) \tag{23}$$

其中 $P(\mathrm{SC}_{\mathrm{j}}^{w_n})$ 表示辨識字串 W 的第 n 個字 w_n 屬於第 j 個語義的機率。這可以從語料庫中離線預先估測而得。

3.10 對話行爲歷史記錄

對話行爲歷史記錄方面的目的是爲避免使用者在查詢其中一項任務時，在尚未完成任務卻因爲 ASR 錯誤造成對話行爲誤判而轉爲詢問使用者其他任務的內容。假設 DA_t 定義爲目前語音所得到的 DA，即 DA_C，而過去歷史紀錄 DA_t^{t-1} 定義爲 DA_H，倘若我們假設對話行爲只與前一個對話行爲有關，則數學表示法可定義爲：

$$P(DA_t = DA_C \mid DA_1^{t-1} = DA_H) = P(DA_t \mid DA_1, DA_2, \cdots, DA_{t-1}) = P(DA_t \mid DA_{t-1}) \tag{24}$$

四、 對話管理決策

對話管理是基於對話行為偵測結果而採取適當回應以與使用者進行互動，而適當回應仰賴於對話策略的規劃。在本研究中，我們採用 POMDP 作為對話策略規劃的工具。POMDP 將系統所處的狀態視為隱含變數(hidden variable)，因此必須使用一個相信函數(belief function)來假設系統所處的狀態並定義了五個變數值組(tuples) {**S**,**A**,**R**,**T**,**O**}，分別代表狀態組成的集合 **S** 並用一個相信函數(belief function) b 來控制(maintain)，在本研究中，定義為前文所提的 DA；回應使用者的方式所組成的集合 **A**；獎勵函數 $R(s,a)=r$，表示在狀態 s 採取回應使用的方式 a，系統所得到的獎勵為 r；轉移機率 $T(s,a,s')=P(s_{t+1}=s'|s_t,a_t)$為系統在時間點 t，在狀態 s 採取使用的方式 a，而在時間點 t+1，狀態將會變成 s'的機率；觀察(observation)所組成的集合 **O**，描述著 POMDP 能接收的訊息而觀察機率 $P(o'|s',a)$表示系統在時間點 t 採取使用的方式 a，及在時間點 t+1 系統所處的狀態 s'，所觀察到的觀察的機率。綜合以上變數，POMDP 應用於本研究的概念圖如圖 4 所示。

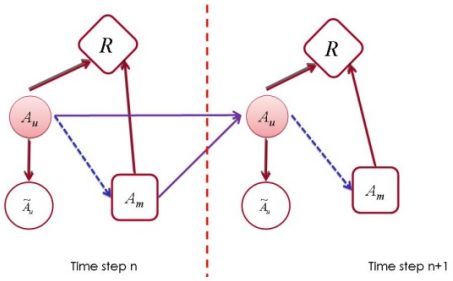


圖 4：馬可夫決策程序概念圖。

在相信函數的更新部份，我們引用文獻[11]所推導的公式。

$$b'(s') = P(s'|o',a,b) = k \cdot P(o'|s',a)\sum_{s\in S} P(s'|s,a)b(s) \tag{25}$$

其中 $b'(s')$ 為更新的相信函數， $P(o'|s',a)$ 為觀察機率， $P(s'|s,a)$ 為轉移機率， $b(s)$ 為相信函數。部分觀察馬可夫決策程序的最佳值函數(Optimal value function)為：

$$V^*(b) = \max_{a\in A}[\sum_{s\in S} r(s,a)b(s) + \gamma \sum_{o',s,s'} p(o'|s',a)p(s'|s,a)b(s)V(b'(s'))] \tag{26}$$

4.2 對話策略學習

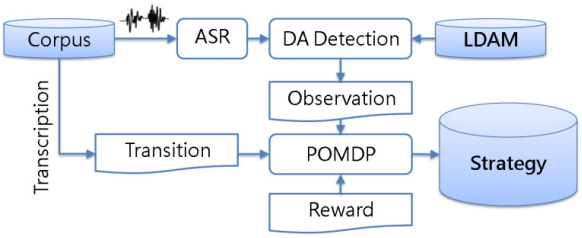


圖 5：對話策略訓練流程圖。

在訓練對話策略之前，我們必須定義系統的狀態、觀察及獎勵函式。在本論文中，對話策略學習的狀態和狀態所對應的回應行為為第二節所提到的 slot 和 action 的部份。而對話策略訓練流程圖如圖 5 所示。首先，從我們收錄的語料中，由文字的每個對話(dialogue)得到轉移(transition)機率，及從文字相對應的音檔經由對話行為偵測後得到觀

察機率，及定義獎勵函式爲：

$$r=\begin{cases}+10 & ,if\ \text{系統採取正確回應}\\ -10 & ,if\ \text{系統採取錯誤回應}\\ -5 & ,if\ \text{重複詢問問題}\\ -100 & ,if\ \text{系統採取正確回應歡迎發生在開始之外}\\ +100 & ,if\ \text{系統結束}\end{cases} \tag{27}$$

在觀察機率，假設觀察爲對話行爲型態偵測後的假設結果，可表示爲：

$$P(o'\mid s',a)\equiv P(DA_{o'}\mid DA_{s'},a) \tag{28}$$

假設觀察與上一次系統回應無關，所以觀察機率爲：

$$P(DA_{o'}\mid DA_{s'},a)\cong P(DA_{o'}\mid DA_{s'})=\begin{cases}P(DA_C\mid S,DA_H)(1-p_{err_C})\\ (1-P(DA_C\mid \mathrm{U},DA_H))\cdot\dfrac{p_{err_C}}{\left|\mathrm{DA}_u\right|-1}\end{cases} \tag{29}$$

其中 P_{errc} 透過訓練語料求得每個對話行爲型態的正確率。此觀察機率包含著對話行爲型態偵測分數 $P(DA_C|\mathrm{U}, DA_H)$及此對話行爲型態偵測結果的可信度$(1\text{-}P_{errc})$。最後我們經由 POMDP 軟體[22]訓練我們的對話策略。

五、 實驗

5.1 自動語音辨識器(ASR)的建立

ASR 的基本建立步驟，包含了聲學特徵參數的萃取(feature extraction)、聲學模型訓練(acoustic model, AM)和語言模型(language model)的訓練。我們採用劍橋大學所開發的工具 HMM Tool Kit(HTK)來建立本研究的 ASR。爲了建立一個較爲可靠的 ASR，我們先採用麥克風語料庫 TCC300 進行種子(seed) AM 訓練，包括 115 個右相關(right-context dependent) initial 次音節和 38 個獨立(right-context independent) final 次音節，分別使用 3 個和 5 個狀態(state)，每個狀態最多 32 個 mixtures。，再以我們所錄製而來的旅遊相關語料進行調適(adaptation)。特徵參數爲 39 爲度的梅爾倒頻譜特徵參數(MFCC)，其中預強調係數爲 0.97，其餘相關參數設定可參考 HTK Book 說明。在語言模型部分，我們採用 TCC300 語料庫來建立語言模型的部分，並嘗試進行調適。然而，經由實驗發現，語言模型在本系統似乎作用性不高，這是因爲我們所收錄的旅遊相關語料內容對於 TCC300 而言屬於微量，因此許多旅遊景點相關字詞無論在 uni-gram 和 bi-gram 都只是極小值。就語音辨識器而論，我們所建立的語音辨識器對於所蒐集的旅遊相關語料庫有高達 84.33%的正確率。經語者調適後，更可達 93.12%的正確率。

5.2 對話語料之分析

將收集而來的語料，經過整理挑選出適合的語料作爲訓練語料之用，語料總共有 144 個對話回合，總數爲 1586 句，爲了了解對話用句分佈情形，本論文針對語料做了兩種分析，分別爲對話之意圖分佈圖和對話的長度分析。在表 3 中，我們定義了 38 種 DA 而圖 6(a)呈現出語料裡各種 DA 分佈的情形。由分佈圖可知，使用者會根據他的需求查詢他所想要的資訊，所以每個意圖出現頻率不同，而「結束」出現頻率遠高於其他意圖說明了在對話結束時使用者習慣性說告別用語。說明了語料裡對話回合次數分佈的情形，

所謂對話回合次數即是一次對話需要來回多少次(turns)。就如圖 6(b)所示，每一次對話的句數大都分布在 3~15 句之間。而 3~5 句代表著使用者只使用一項任務便結束系統，其他則表示使用者可能同時查詢了好幾項資訊才結束系統。

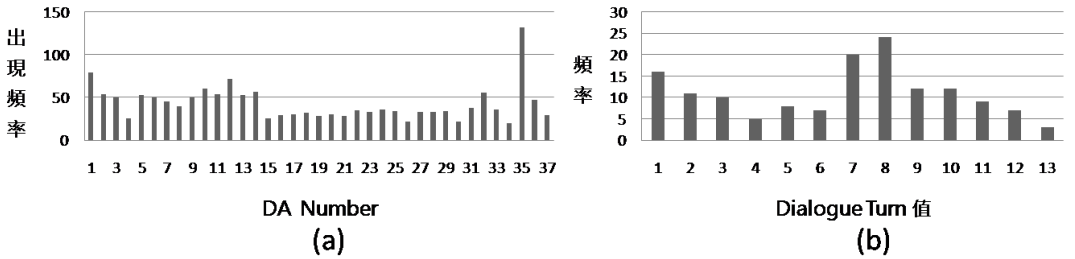


圖 6：對話意圖分佈和對話回合次數分佈。

5.3 系統評估分析

系統評估方面，我們把評估方式分爲二種，第一種是對各類 DA 做評估，觀察 DA 偵測模組對於每類 DA 的運作效能。第二種是對話回合(turn)次數的評估，此評估目的在於了解 POMDP 對於整個對話流程其影響結果。評估語句總數爲 912 句。評估 DA 偵測模組時，分別考慮(1)僅使用語義表格(semantic slot, SS)，依據使用者填入的 semantic slot 判斷使用者的意圖，(2)使用 SS 和史丹佛剖析器(Stanford Parser, SP)建立潛在意圖矩陣 LDAM 的偵測方式，(3)使用 SS、SP 和部分樣本樹(PPT)建立 LDAM 的偵測方式和(4)使用 SS、SP、PPT 和句型分類(sentence clustering, SC)建立 LDAM 的偵測方式四種對話行爲模組的評估方法。各方法的對話行爲偵測準確率如圖 7，平均正確率分別爲 49.6%、76.2%、81.6%和 82.9%。圖 7 標示爲 DA 的欄位爲 37 種 DA，缺少「無意圖」DA 是因爲無法收集無意圖 DA 的句子，故無法評估其準確率，無意圖 DA 是用來當系統無法判斷使用者的意圖時所做出的回應。評估方法(1)的結果，雖然在某些 DA 上能有可接受的表現，但我們可以發現屬於任務 3 的表現出現落差，因爲 DA 中有幾種彼此會互相混淆，造成很多意圖的準確率是 0%，例如詢問交通方式的 DA。使用評估方法(2)的結果，相較於(1)有明顯的改善，但依然無法解決因爲語音辨識錯誤而造成 DA 偵測的問題。使用評估方法(3)的結果，在測試時，語句的辨識結果必須經過 PPT 產生樣本候選據來進行意圖偵測，我們可以看到標示爲「無出發地有目的地」的 DA 有非常顯著的改善，這類的句子如「怎麼到安平古堡」容易在評估方法(1)和(2)被判別爲其他類似的對話行爲型態，如「火車無出發地有目的地」DA。最後，評估方法(4)，這個組合方式即爲論文中 LDAM 模型的訓練方式，經實驗證明，在所有的方法中，我們所建構的 LDAM 模型是可行的。另外，查詢地點(2)、查詢車站(3)和系統歡迎(36)因爲問題本身簡單，以致於四種評估方法皆有相同的效能。在查詢交通方式的火車和高鐵部分，因爲本身的關鍵字詞重疊性太高，以致於沒有明顯的改善表 4 爲評估由 POMDP 訓練而得的對話策略模式對於我們所蒐集語料的效用，我們可以發現 POMDP 的確能降低對話回合次數，但由於收錄的語料辨識結果還算正確且對話行爲型態判斷大部份也是正確，所以降低的對話回合次數沒有顯著的提升。在另一方面，本來對話行爲型態偵測錯誤的句子，若是用人工訂定的回應可能會產生奇怪的回應，例如使用者詢問地址而系統卻回答票價，若使用 POMDP 則可能產生較爲正確的回應，例如上述例子中，系統回應將變爲詢問使用者意圖。雖無減少對話回合次數但卻使得使用者覺得系統回應更爲人性化，所以 POMDP 確實有它的效能。

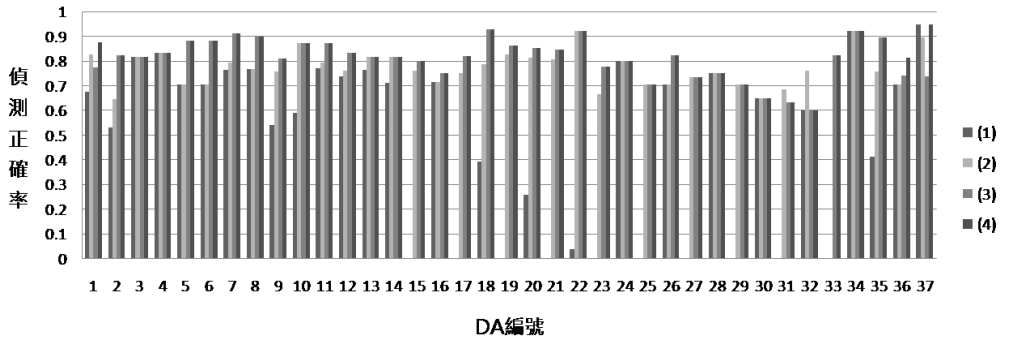


圖 7：各方法對話行爲型態準確率比較：(1)僅使用 SS 的偵測，(2)使用 SS 和 SP 偵測，(3)使用 SS、SP 和 PPT 的偵測，(4)使用 SS、SP、PPT 和 SC 的偵測。

表 4 ：對話回合次數之比較。使用 POMDP 訓練而得的策略降低了對話回合數。

	without POMDP	with POMDP
Average #(Turns)	8.6	8.4

六、結論與未來展望

本論文提出了利用部份樣本句與句型規則建構出潛在意圖矩陣，藉此判斷對話行爲型態，並有效改善因語音辨識錯誤造成對話行爲型態錯誤的問題。此外本論文也加入對話歷史的概念，考慮對話語義脈絡來幫助對話理解。爲了增加人機之間的互動，在系統決策管理方面也運用 POMDP 以求取最佳策略，使得系統產生最佳回應，減少系統與使用者之間無法完成任務的情況。透過實驗的證明，本論文所提出的方法在潛在對話行爲偵測上平均準確率爲 81.9%，相較於單純使用表單填格方式的意圖偵測平均準確率 48.1%，使用本方法可提升了 33.8%之準確率，本論文所提出的方法的確是有效的方法。雖然本論文所提出的方法可達到不錯的成效，但仍有不少地方有待改進，以下我們將逐一說明可改進的地方：(1)如何自動找尋應用領域的對話行爲型態以減少人爲介入，爲理解系統另一個研究議題。(2)在本論文中所使用的 POMDP，可以進一步將狀態假設爲許多不同值域的集合，使系統能更細緻地與使用者互動。(3)可以定義獎勵函數，根據不同的填值狀況作不同的獎懲。

參考文獻

[1] X.-D Huang, Alex Acero, H.-Wd Hon, “Spoken Language Processing”, Prentice-Halln, Inc. 2001

[2] Ji.-J. Liu, Y.-S. Xu, S. Seneff, and Victor Zue, “Citybrowser Ⅱ: A multimodal restaurant guide in Mandarin”, in Proc. International Chinese Spoken Language Processing, 2008.

[3] AT&T(2002) How May I Help You? [Online]. Available: http://www.research.att.com/~algot/hmihy/

[4] C. Hori, K. Ohtake, T. Misu, H. Kashioka, S. Nakamura, “Dialog Management using Weighted Finite-State Transducers”, Interspeech, 2008

[5] S. Bennacef, L. Devillers, S. Rosset, and L. Lamel, “Dialogue in the RAILTEL Telephone-Based System”, in Proc. of ICSKP’96, vol. 1, pp. 550-553, 1996

[6] C.-J. Lee, E.-F. Huang, and J.-K. Chen, “A Multi-keyword Spotter for the Application of the TL

Phone Directory Assistant Service", in Proc. Workshop on Distributed System Technologies & Applications, pp. 197-202, 1997

[7] 蔡金翰, "語音對話系統和對話策略之研究," 國立交通大學電信工程學系碩士論文, 2005

[8] T.-H. Chiang, C.-M. Peng, Y.-C. Lin, H.-M. Wang and S.-C. Chieh, "The Design of a Mandarin Chinese Spoken Dialogue System", in Proc. COTEC'98, Taipei 1998, pp.E2-5.1~E2-5.7

[9] 陳銘軍, 葉瑞峰, 吳宗憲, "以知識概念模型爲基礎之多主題對話管理系統", in Proc. ROCLING XV, Hsinchu, Taiwan, 2003.

[10] W.-S. Choi, H. Kim, and J.-Y. Seo, "An Integrated Dialogue Analysis Model for Determining Speech Acts and Discourse Structures," the Institute of Electronics, Information and Communication Engineers (IEICE), 2005

[11] J. D. Williams, and Steve Young, "Partially Observable Markov Decision Processes for Spoken Dialog Systems," Computer Speech and Language, 2007.

[12] David R. Traum, "Speech Act for Dialogue Agents," Kluwer Academic Publishers, 1999.

[13] Y.-C. Xiao, "MHMC Annotation of MHMC Travel Corpus," 2009. [Online]. Avaliable: http://chinese.csie.ncku.edu.tw/~liang/MHMC_Annotation_of_Travel_Corpus.pdf

[14] Stanford Parser [Online]. Available: http://nlp.stanford.edu/software/lex-parser.shtml

[15] E. Shriberg and A. Stolcke, "Word Predictability After Hesitations: A Corpus-Based Study", in Proc. International on Conference Spoken Language Processing (ICSLP), pp. 1868-1871, 1996.

[16] M Siu, M. Ostendorf, and H. Gish, "Modeling Disfluencies in Conversational Speech" , in Proc. International on Conference Spoken Language Processing (ICSLP), vol 1, pp. 386-389, 1996.

[17] T.R. Niesler and P.C. Woodland, "Variable-Length Category N-gram Language Models", Computer, Speech and Language, vol. 21, pp. 1-26, 1999.

[18] J. S. Hamaker, "Towards Building a Better Language Model for Switchboard: the POS Tagging Task," in Proc. International Conference on Acoustics, Speech, and Signal Processing(ICASSP), pp. 579-582, 1999.

[19] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," IEEE Trans. on Speech and Audio Processing, vol. 9, no. 3, pp. 288-298, 2001

[20] Dan Klein, and C. D. Manning, "Fast Exact Inference with a Factored Model for Natural Language Parsing," in Advances in Neural Information Processing Systems, 2003.

[21] Dan Klein, and C. D. Manning, "Accurate Unlexicalized Parsing," in Proc. the 41st Meeting of the Association for Computational Linguistics, pp. 423-430, 2003.

[22] POMDP 軟體 [Online]. Available: http://staff.science.uva.nl/~mtjspaan/software/approx/

資源受限運算環境下華英混雜語音辨識系統

Mandarin/English Mixed-Lingual Speech Recognition System on Resource-Constrained Platforms

洪維廷　Wei-Tyng Hong
元智大學通訊工程學系
Department of Communications Engineering
Yuan Ze University

陳弘啓　Hong-Ci Chen
元智大學通訊工程學系
Department of Communications Engineering
Yuan Ze University

廖宜斌　I-Bin Liao
中華電信研究所
Telecommunication Laboratories, Chunghwa Telecom

王文俊　Wern-Jun Wang
中華電信研究所
Telecommunication Laboratories, Chunghwa Telecom

摘要

本論文提出結合華語、英語語音模型進行混合關鍵詞辨識，並且於資源受限制的情況下進行全整數運算。我們將所處理過後的定點化特徵參數經過 RASTA 濾波器[1]，成爲定點化 RASTA 特徵參數。而針對關鍵詞詞彙結構的相關性，我們建立一個樹枝狀的搜尋架構，並且利用光束搜尋法，減少辨識語料中音框所需的音節節點，進而減少搜尋空間，並且維持一定的辨識率。

針對辨識語料所產生的華語、英語語音模型相似度分數差異，我們提出偏差補償值應用於英語模型，並提出改變前置詞搜尋機制。增加可忽略搜尋前置詞路徑與無關詞垃圾模型路徑，以因應使用者沒有按照使用規則說出定義內前置詞的情形，藉此測試對於主詞的辨識率。

關鍵詞：　詞典樹、光束搜尋法、垃圾模型

Keywords：Lexicon Tree、Beam Search、Garbage Model

一、緒論

目前在自動語音辨識的發展課題上，大部分仍以單一語言構成的語音爲主，混合語言的語音研究仍則較爲少見。在 2006 年，國內成功大學電機工程研究所黃建霖等人[2-3]，利用聲學與文脈分析應用於多語(multilinguality)語音辨識單元的產生，並且利用融合技術(Fusion)結合聲學相似度及前後文脈分析，有效提升多語語言的辨識率。在 2008 年，國內成功大學電機工程研究所李奇峰等人[4]發展一套多語言辨識單元集技術，在原有的語音辨識系統前端，建立一個有效的混語語音模型，並且將系統實現於 PDA 之嵌入式系統裝置上。在 2004 年時工研院資通所所發表的「中英文混雜關鍵詞萃取技術」[5]，提出一個中英文辨識分數的偏差補償值應用於中英文混雜辨識，去克服中英文聲學模型間的差異性，並且在關鍵詞認證方面，則提出一個正規化的驗證機制，而從驗證分數裡找出一組最好的結果。

上述所提到的多語語音辨識即是結合多個單一語言(monolingual)模型，於前端判斷所說詞彙的語言種類，之後再進行多語語音辨識，而在本論文中的混合語音辨識研究，則是辨別華英語混合的關鍵詞詞彙，並且於辨識時灌進華語、英語語音模型，利用每個時間音框中所累積的相似度分數，去判別出最佳的辨識結果。

在於關鍵詞的萃取方面，可以分爲關鍵詞模組與無關詞模型。在 1985 年，Higgins 與 Wohlford[6]利用連續語音辨識應用於關鍵詞萃取(keyword Spotting)，並且定義填充模型(filler template)去表示無關詞部份。在 1989 年，Rohlicek 等人[7]提出在關鍵詞模組狀態中的機率密度函數加入權重值去建立無關詞模組。在 1990 年，Rose 與 Paul 提出[8]使用單音節來當作無關詞模組。近幾年來，陸續提出利用 HMM Base 的次音節關鍵詞與無關詞模組應用於關鍵詞的辨識。在無關詞模組方面，在 1996 年，Caminero 等人[9]提出加入垃圾模型(Garbage Model)，並且套用鑑別分析準則於言語判別(utterance verification)。在辨識結構上，加上垃圾模型於文法規則上，可以降低對於關鍵詞誤判的機率。

當系統的無關詞與關鍵詞模組加大時，所需要的搜尋時間將會大大增加，尤其是花費大量資源在搜尋路徑方面，這對於應用於資源受限的智慧型裝置有極大的挑戰，所以必須利用較快速有效的方法，去達成一定水準的辨識結果。在 1990 年，Frank k.Soong 等人[10]提出利用樹枝狀結構下的快速搜尋法找尋 N-Best 的辨識結果，應用於連續語音辨識。在 2004 年，Xie Lingyun 等人[11]，提出在每個時間音框點，利用動態刪減去調節搜尋範圍，並且應用於大詞彙連續語音辨識系統。

本論文模擬於資源受限的環境下做全整數運算。首先，建構前置詞(Prefix)與主詞(Main)的詞典樹(Tree Lexicon)，利用詞彙內字與字的連結關係，佈置成樹狀結構，爾後當成搜尋時的依據。當測試語料進來時，我們利用光束搜尋法(Beam Search)去找出每個語料的每個音框(Frame)可能存活的節點，並且經過排序剔除分數值較低的節點，最後找出一組最佳路徑當成辨識結果。而爲了使系統更加有延展性，我們改變前置詞的搜尋機制，加入可忽略搜尋前置詞路徑，測試使用者忘記說定意內前置詞時，可以繞過此路徑保留對於主詞的辨識效能；加入無關詞垃圾模型，測試使用者說錯定義內前置詞時，錯誤路徑可以被吸收，藉此保留對於主詞的辨識率。

二、華英語混合辨識系統

華語、英語混合辨識的系統流程圖如下圖 2-1。

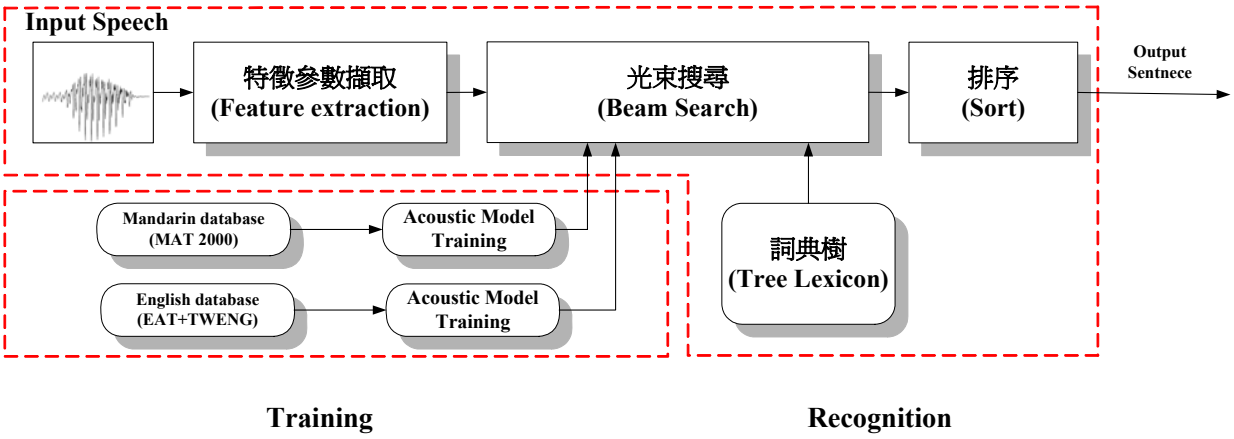


圖 2-1：華語、英語混合辨識流程圖

根據上圖的流程圖，我們的系統主要分爲兩個部份，分別爲訓練階段與辨識階段。其中訓練階段是在 Linux 下做浮點數運算，利用隱藏式馬可夫模型分別訓練出華語語音模型與英文語音模型。在測試階段，爲了符合於資源受限的環境，所以我們於 PC 上進行全整數運算。當一個測試語料進來時，我們先做語音訊號的特徵參數擷取，之後進入光束搜尋法的程序，結合詞典樹與匯入華語、英語語音模型進行關鍵詞萃取。並從所記錄的節點路徑中，經過語音模型相似度分數的累積與排序，尋找出最佳的前三名辨識結果。

2.1 隱藏式馬可夫模型

隱藏式馬可夫模型是一種以機率統計的方式來做辨識的模型，辨識語音的度量是計算從模型產生的機率值大小。一般常用於狀態觀測機率的機率密度函數爲高斯混合模型(Gauss Mixture Model)。但我們爲了因應在智慧移動裝置上面的資源限制，所以採用拉普拉斯分佈(Laplacian Distribution)做爲我們的狀態觀測機率。其 probability density function 定義如下式：

$$Lap(x,u,v)=\frac{1}{2v}\exp\left(-\frac{|x-u|}{v}\right) \tag{1}$$

其中 u 爲 location parameter，v 爲 scale parameter。假設一個維度爲 D、特徵值爲 x 的語音訊號，假設在每個特徵參數值之間是獨立的關係。我們以此狀態中所有混合數(mixture)最大的機率值來代表隱藏式馬可夫模型第 j 個狀態的觀測機率：

$$\begin{aligned} b_j(x) &= \max_m \left\{ \prod_{d=1}^{D} Lap\left(x_d; u_{j,m,d}; v_{j,m,d}\right) \right\} \\ &= \left\{ \prod_{d=1}^{D} \frac{1}{2v_{j,m,d}} \exp\left(-\frac{|x_d - u_{j,m,d}|}{v_{j,m,d}}\right) \right\} \end{aligned} \tag{2}$$

其中$u_{j,m,d}$和$v_{j,m,d}$分別爲隱藏式馬可夫模型第 j 個狀態上第 m 個混合數(mixture)之第 d 維度的 location parameter 和 scale parameter。假設 k 滿足下列式子：

$$k = \arg\max_{m} \left\{ \prod_{d=1}^{D} Lap\left(x_d; u_{j,m,d}; v_{j,m,d}\right) \right\} \tag{3}$$

$b_j(x)$的 log 式子可表示如下：

$$\log\left(b_j(x)\right) = C_{j,k} - \sum_{d=1}^{D} \frac{\left|x_d - u_{j,k,d}\right|}{v_{j,k,d}} \tag{4}$$

其中$C_{j,k}$爲和 HMM 第 j 個 state 第 k 個 mixture 的參數相關的數值(和特徵值 x 無關)，可預先算出作爲另一個模型參數。接下來要進行的是降低計算量和防止溢位(overflow)：

1 經由適當的定點化技術，將$x_d, u_{j,k,d}, v_{j,k,d}, c_{j,k}$轉換成 16-bit 整數$x_d^{'}, u_{j,k,d}^{'}, v_{j,k,d}^{'}, c_{j,k}^{'}$，如此可保證$\log(b_j(x))$爲 16-bit 整數範圍；另外在 Beam Search 模組中 Viterbi Search 使用之 log-likelihood 累加陣列(accumulated array)則以 32-bit 整數宣告，這樣可降低累加陣列 overflow 的機率。

2 將除法運算拿掉，令整數$\beta_{j,k,d} = \frac{2^s}{v_{j,k,d}^{'}}$，經過適當式子重整，可以由下列的全整數運算式子逼近：

$$\log(b_j(x)) \approx C_{j,k}^{'} - \left[\sum_{d=1}^{D} \beta_{j,k,d} \times \left| x_d^{'} - u_{j,k,d}^{'} \right| \right] >> s \tag{5}$$

其中 >> 爲 bit-shift right 運算元。

2.2 詞典樹

在建構整個詞典樹的過程中，包含了前置詞(Prefix Word)與主詞(Main Word)。在本論文中，所有詞彙皆爲華語、英語或華英語混雜的語料，其中前置詞的個數爲 10 個，主詞個數爲 200 個。在此，我們舉出一個範例，例子中前置詞有〝Email〞與〝查詢〞，而主詞有〝元智大學〞、〝元智大學通訊所〞與〝元智大學通訊所辦公室〞等五個詞彙，利用樹狀結所構成的辨識詞彙集合如圖 2-2 所示。T 代表詞彙的終結。

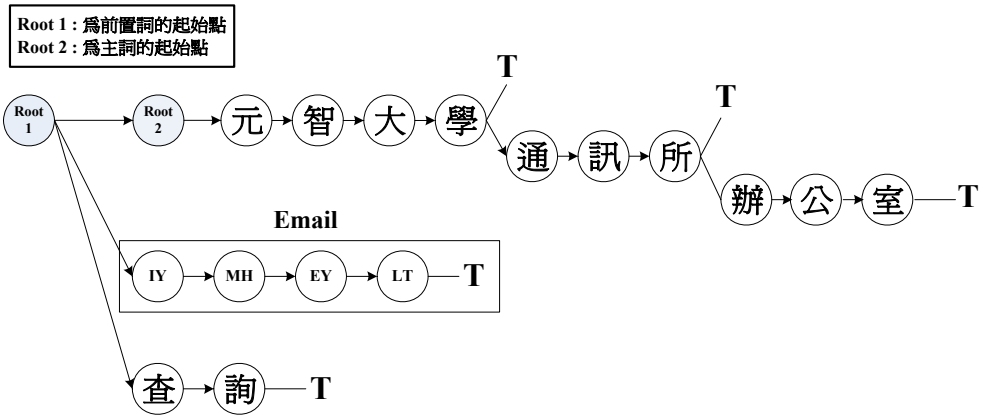


圖 2-2：樹狀結構圖(一)

依據上圖 2-2 的方式可以將關鍵詞詞庫建立成一個樹狀結構的詞典樹，其中華語音節點對應 411 個音節碼，而英語音節點則是對應 134 個右相關音素模型(Phone Model)，之後將一階動態搜尋器的搜尋空間由華語的 411 音節碼與英語的 134 個右相關音素模型展開所可以連結的音節點節點。如果爲華語節點，則有 7 個狀態，包含基本音節之 Initial 2 個與 Final 4 個狀態，跟一個可以跳過的靜音狀態來描述音與音之間可能存在/不存在的靜音，如下圖 2-3 所示。如果爲英語節點，則有 3 個狀態，包含 2 個基本音節狀態與 1 個靜音狀態，如圖 2-4 所示。在搜尋的法則上，華語部份於 Final 最後一個狀態限制只可以連結靜音狀態或者華語部分的 Initial 第一個狀態或者英語部份第一個狀態；英語部份則是在狀態 2 之後只可以連結靜音狀態或者華語部份 Initial 第一個狀態或者英語的第一個狀態。

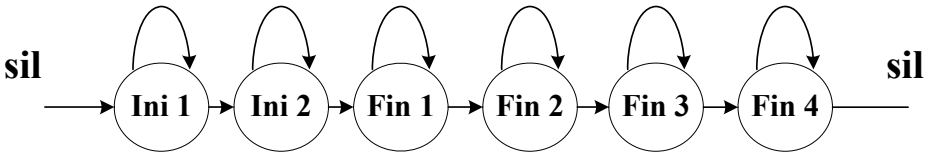


圖 2-3：華語節點中之狀態圖

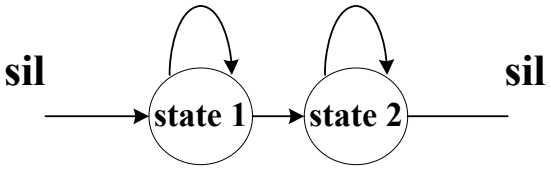


圖 2-4：英語節點中之狀態圖

2.3 光束搜尋法實做

在光束搜尋法設計實作方面我們考慮音框中語音節點會屬於靜音節點、華語節點、英語節點與垃圾模型節點四種情形。四種不同的節點會根據詞典樹與文法限制去展開可連結的節點，並且依據節點內的語音模型相似度分數大小，通過光束搜尋法的篩選，產生光束寬度(BeamWidth)。以下我們利用虛擬碼(Pseudo Code)去描述系統內一個音框資訊搜尋語音節點的演算法：

*Step*1：t=0 的音框進來。(t=音框 index)

*Step*2：依照佈於詞典樹裡所有前置詞詞彙的第一個音節去計算華語或英語語音模型的相似度分數，並且與靜音模型相似度分數進行分數上的排序。

*Step*3：t=1 的音框進來，依據 *Step*2 所產生的音節個數，重複執行 *Step*5 與 *Step*6，並進入 *Step*7 產生新的音節個數。

*Step*4：t=2 到 t=T 的音框進來，依據 *Step*7 所產生的音節個數，重複執行 *Step*5 與 *Step*6。

*Step*5：節點判別爲靜音節點，當節點內標號爲-1 時，則生長佈於詞典樹裡所有前置詞詞彙的第一個音節點；標號不爲-1 時，則依據詞典樹中此標號之後可以連結的節點進行連結。

*Step*6：節點判別爲華語節點、英語節點或垃圾模型節點，如果此節點未生成前一個語音狀態，則生成；如果非最後一個語音狀態，則產生下個語音狀態。如果已經到最後一個語音狀態，則判斷是否爲前置詞或主詞的終結點，並且可以連結到靜音節點。此時節點如果不爲主詞的終結點，則此節點可以繼續連結其他華語節點或英語節點。

*Step*7：累積 *Step*5 與 *Step*6 所產生的節點，進行靜音、華語、英語、垃圾模型節點間語音模型相似度分數間的排序，並且經過光束搜尋法的篩選，剔除分數較小的節點。

*Step*8：最後一個音框內的語音節點經過相似度分數排序過後，進入辨識程序。

在此我們也舉出一個例子，說明詞典樹與光速搜尋法如何結合應用於系統中。我們設定總共有 6 個詞彙，包含了 3 個前置詞與 3 個主詞，共有 15 個節點。依據詞彙內音節佈成的樹狀結構如下圖 2-5：

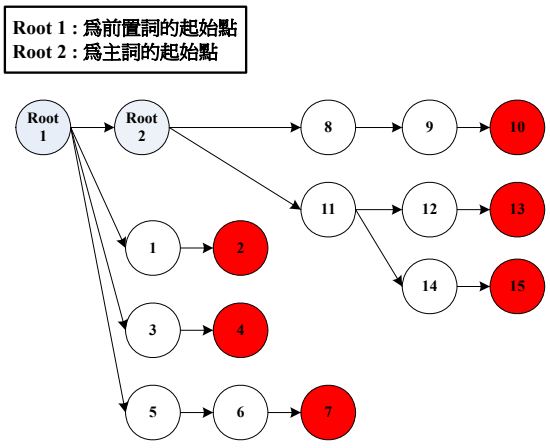


圖 2-5：樹狀結構圖(二)

其中數字 1~15 代表詞彙內的音節，紅色代表此詞彙的結束節點。下圖 2-6 以圖 2-5 的詞典樹爲依據所產生的光速搜尋法路徑。當測試語料進來時，先生長前置詞、靜音與垃圾模型節點，再生長主詞節點，並且依照節點內狀態所對應的語音模型相似度分數大小，進行排序與篩選。在此，圓圈間的連結所代表的是節點內狀態的轉移情形。而此例子中，每個節點有 3 個狀態數，爲 S0、S1、S2。而靜音與垃圾模型節點各爲一個狀態數。

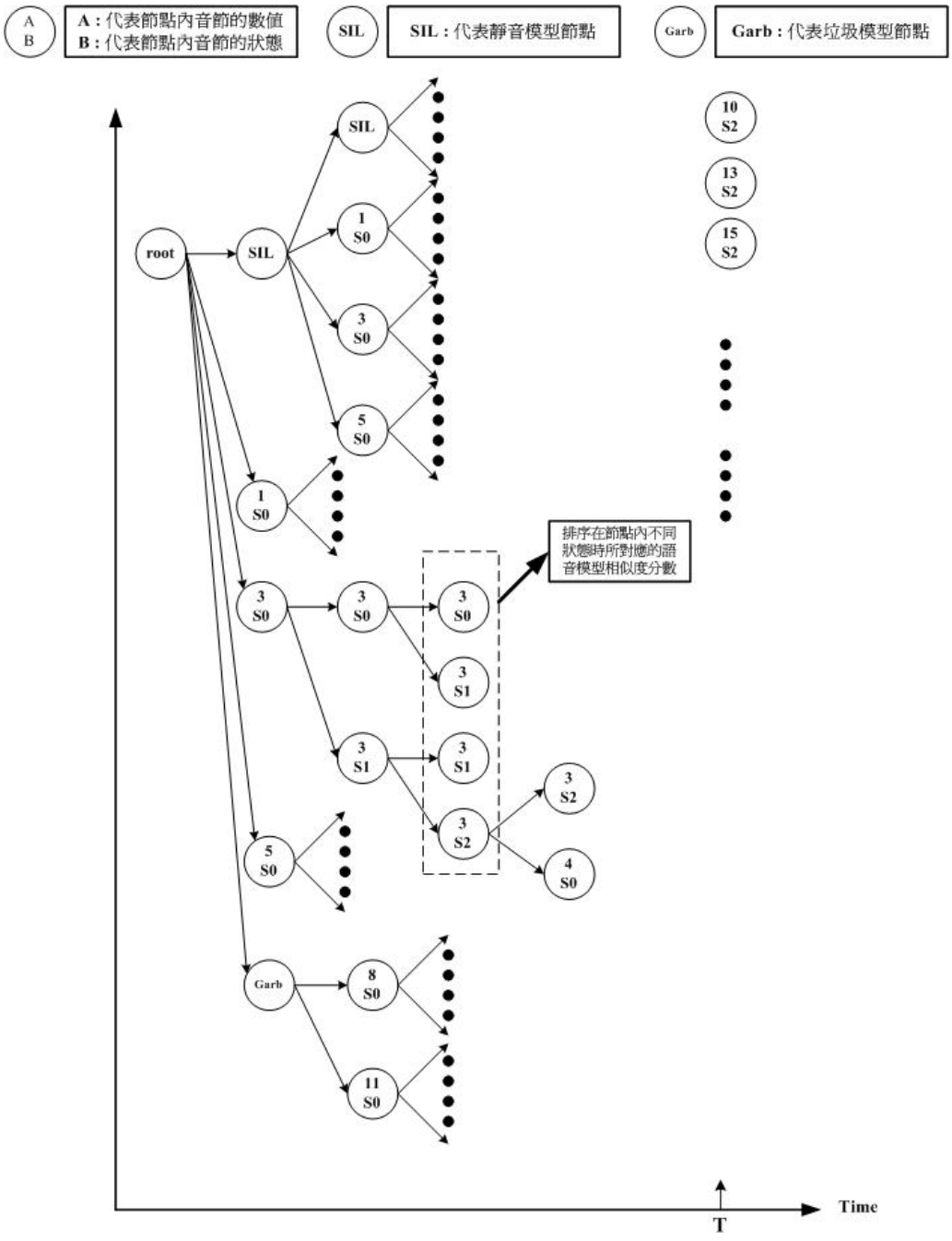


圖 2-6：光束搜尋法路徑圖

2.4 華英語混合關鍵詞萃取

所謂關鍵詞的萃取，是指在特殊辨認詞彙下，事先對此特殊詞彙選取若干個關鍵詞，在辨識的時候只要將預先定義好的關鍵詞萃取出來。這部分我們主要是介紹華英語混合的語音辨識，重點是放在關鍵詞的辨識上面，其中關鍵詞包含了前置詞與主詞的辨識，例如：〝*查詢 元智大學*〞。在一些特殊情況下，例如使用者忘記說或說錯定義內前置詞時，針對於主詞關鍵詞萃取辨識還是我們的目的，這時候我們就在搜尋機制中加入可忽略搜

尋前置詞的路徑或無關詞垃圾模型路徑，去降低對於主詞的破壞性，保留對於主詞的辨識率。

而關於搜尋機制的改變，我們主要針對前置詞搜尋機制方面做探討。首先，於前置詞詞庫中加入可忽略搜尋前置詞的路徑，此路徑並非定義內的前置詞，而是爲一個靜音模型的路徑。對此我們測試使用者忘記說定義內前置詞時對於主詞的辨識率。爲了使系統更有彈性與效能，接下來我們在前置詞詞庫中加入無關詞垃圾模型路徑，去抵抗當使用者說錯定義內前置詞的情況。而整個前置詞搜尋路徑的改變，如下圖 2-7 所示。因爲關鍵詞辨認的技術比連續語音辨識來的廣泛且穩定，所以我們將對華英語混合的關鍵詞辨識做以下更詳細的討論。

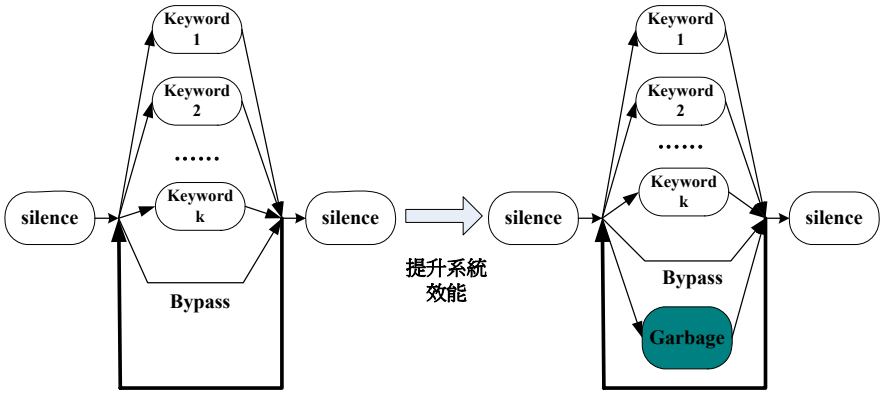


圖 2-7：前置詞搜尋機制圖

2.4.1 關鍵詞模組

關鍵詞即根據特殊辨識詞彙的不同，而去事先定義、選取的辨識標的。在前置詞方面，有尋找、打電話給、Email...等等日常生活中人與人之間會交際互動的詞彙；而在主詞方面，則以中英文人名爲主。我們採用次音節中的右相關音素模型(RCD)串連來產生關鍵詞模組，作爲聲學層次的辨識，這樣可以根據特殊辨識詞彙的不同，而前置詞或者主詞的加減，不需重新訓練模型，使的系統更有彈性。

2.4.2 無關詞垃圾模型

無關詞垃圾模型類似於是填充模型，利用填充模型的混淆來拉下無關詞的分數，進而增加辨識率。而本論文中的無關詞垃圾模型是將華語訓練語料中全部聲母部分所有音框特徵參數資訊重新訓練成一個單一新的聲母模型，只有一個狀態數且混合數爲 32，而也將韻母所有音框特徵參數資訊重新訓練成一個單一新的韻母模型，同樣也設定爲一個狀態數混合數爲 32。匯集新的聲母、韻母模型成爲我們所定義的垃圾模型，並且把它加入到所定義的前置詞詞庫路徑之中。雖然我們無法預測使用者的說話情形，但我們必定知道在每句句子之中必定含有我們要辨識的對象，即關鍵詞的存在。在辨識的過程之中我們可以利用連結字音的辨識方法將關鍵詞與無關詞結合成一個辨識單元作處理，達成我們辨識關鍵詞的目的。

2.4.3 關鍵詞萃取的排列

關鍵詞的萃取架構，在前置詞部分，可以包含關鍵詞模組或者無關詞模型；主詞部份則有關鍵詞模組。在本論文中，我們舉出二種主要出現於關鍵詞萃取中搜尋機制改變的情形。假設定義 2 個前置詞分別爲〝**查詢**〞與〝**Email**〞與定義 3 個主詞，分別爲〝**元智大學**〞、〝**台灣大學**〞與〝**交通大學**〞。第一種情況如圖 2-8，

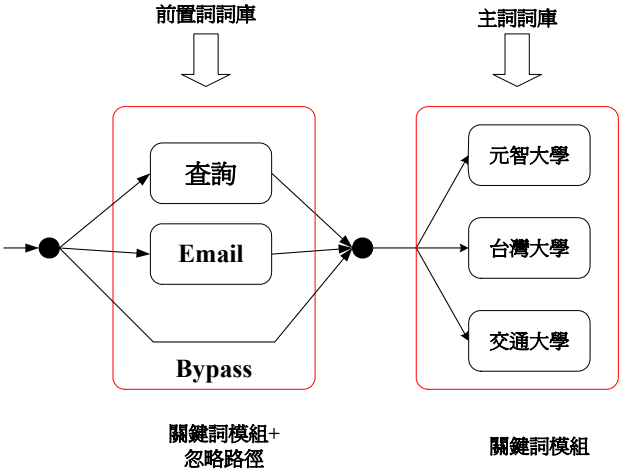


圖 2-8：於前置詞中加入可忽略搜尋前置詞路徑圖

在前置詞詞庫中加入了可忽略前置詞路徑，測試使用者忘記說前置詞時，是否可以讓前置詞搜尋機制繞過此路徑，保留住對於主詞的辨識。第二種情形如圖 2-9，

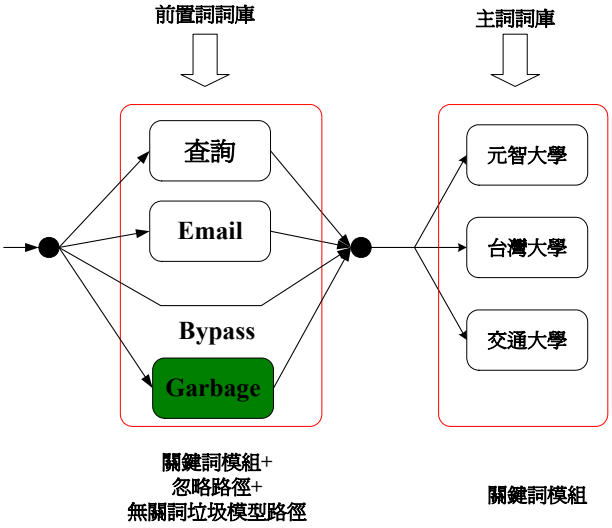


圖 2-9：於前置詞中加入無關詞垃圾模型路徑圖

於前置詞詞庫中加入可忽略前置詞路徑與無關詞垃圾模型路徑，可應用於當使用者說錯定義內前置詞或忘記說定義內前置詞情況時，不會大幅破壞對於主詞的辨識率。

2.5 語音模型的補償

華語與英語的訓練語料來自不同的錄音方式與環境，加上辨認單元和參數量不同造成模型解析度不同，所以當測試語料進來時，我們必須去觀察所產生的華語或英語模型間相似度分數的差距並且作補償的動作。在此，我們會分別對英語模型與無關詞垃圾模型做偏差值補償。

2.5.1 華英語混合關鍵詞的偏差補償

華語語音模型與英語語音模型因訓練參數量不同，所以當要辨識的語音進來時，所產生的華語、英語相似度分數在分布空間上必然不同，進而在分數上產生一定的落差。經過統計與觀察之後，兩者間的差距如圖 2-10 所示。

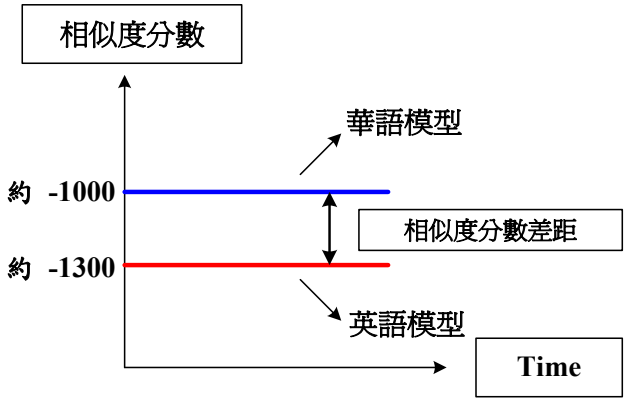


圖 2-10：華英語相似度分數差距圖

因爲相似度分數的差異所以造成在搜尋關鍵字辨識時，容易將華語語音節點與英語語音節點互相判別錯誤，所以爲了改善此判別錯誤的情形，我們在進行搜尋語音節點的時候，對每個英語模型以每個音框爲單位，乘上固定的偏差補償值，如下數學式(6)所示：

$$L_E^{'} = L_E \times 0.7896 \tag{6}$$

其中 $L_E^{'}$ 爲乘過偏差補償值後的英語模型相似度分數值。

2.5.2 無關詞垃圾模型的偏差補償

在本論文之中，華語模型與無關詞垃圾模型都是利用相同訓練語料所訓練而成，只是差別在聲母與韻母的狀態數多寡。在無關詞垃圾模型方面，我們只使用聲母、韻母各一個狀態數，所以訓練出來的模型會比較無鑑別性。而當無關詞垃圾模型與華語模型的相似度分數值進行競爭的時候，我們發現因爲無關詞垃圾模型的相似度分數值小於華語模型，兩者之間有一段差距，造成即使使用者說錯前置詞時，搜尋路徑還是會連帶影響到對於主詞的判斷，導致主詞辨識錯誤。所以我們對無關詞垃圾模型的相似度分數乘上一個靜態偏差補償值，如下數學式(7)、(8)所示：

$$L_{ini}^{'} = L_{ini}^{'} \times 0.824 \tag{7}$$

$$L_{fin}^{'} = L_{fin}^{'} \times 0.8205 \tag{8}$$

其中 L'_{ini} 與 L'_{fin} 爲經過偏差補償值後的聲母、韻母模型相似度分數值。

三、實驗語料

實驗所需的華語訓練語料爲「中華民國計算語言學會」所提供的 2000 人「國語語音資料庫」(Mandarin speech database Across Taiwan, 簡稱 MAT2000)[12]，此語料庫是透過公眾電話網路所錄製，取樣頻率和位元數分別爲 8KHz 和 16 位元。總共選取 84737 句語料當做華語模型訓練語料。而英語的訓練語料爲台灣腔英文資料庫(English Across Taiwan，EAT)[13]加上台灣腔式英語訓練語料(TWENG)。台灣腔英文資料庫分爲麥克風語料與電話語料，其中電話語料可細分爲固定式電話(PSTN)語料與行動電話(GSN)語料。取樣頻率和位元數分別爲 8kHz 和 16 位元，選取 90475 句訊練語料。台灣腔式英文少量語料，其取樣頻率和位元數也是爲 8kHz 和 16 位元，這部分我們選取 9535 句語料當成訓練語料。

實驗中測試語料爲 ME_Speech Corpus，是由 12 位男生與 12 位女生利用 10 個前置詞與 200 主詞所混合組成的關鍵詞辭彙錄製而成。總共有 4800 句語料，其中第一部份 2400 句語料爲包含了前置詞與主詞混合而成的關鍵詞，另外第二部份的 2400 句爲第一部份字句刪除前置只留下主詞的詞彙。

四、實驗分析

本系統中語音信號之取樣頻率爲 8kHz。每個輸入音框(frame)之特徵參數是由 12 維的「梅爾刻度式倒頻譜參數」(Mel-scale cepstrum)及對應之 12 維「倒頻譜差量參數」(delta Mel-scale cepstrum)加上 1 維「差量對數能量」(delta log energy)與 1 維「差差量對數能量」(delta delta log energy)所構成 26 維度的特徵參數，此參數並經過 RASTA 濾波器用於降低通道效應。辨認搜尋單元採用次音節模型，華語部分由 100 個 2 狀態的右相關聲母(initial)模型、38 個 4 狀態的韻母(final)模型與 1 個狀態的靜音(silence)模型組成。而英語部份由 134 個 2 狀態右文相關英語音素(Phone)模型與 1 個狀態的靜音(silence)模型組成。每一個狀態皆爲高斯混合模型，而協方差矩陣則是假設對角矩陣來代表，用於降低運算量。模型的高斯混合(mixture)數目則是根據訓練語料的多寡估算產生，其中華語聲母模型和韻母模型最多有 8 個高斯混合數，靜音模型則由 16 個高斯組成的高斯混合模型；而英語音素模型最多有 16 個高斯混合數，靜音模型則由 64 個高斯組成的高斯混合模型。辨認搜尋採用光束搜尋(beam search)模式之關鍵詞辨認，辨認核心皆經定點化並適合 PDA 級資源下運作。

4.1 改變搜尋機制

實驗中，我們提出兩種搜尋路徑加入系統的搜尋機制中。第一，在前置詞中加入忽略搜尋前置詞路徑，第二，在前置詞中加入無關詞垃圾模型路徑。在這兩種情況下，我們去測試系統對於主詞的辨識效能。以下我們分類爲四種搜尋機制，如下表一：

表一、搜尋機制

	英文模型偏差補償	忽略搜尋前置詞路徑	無關詞垃圾模型路徑
搜尋機制 1			
搜尋機制 2	✓		
搜尋機制 3	✓	✓	
搜尋機制 4	✓	✓	✓

4.2 辨識率比較

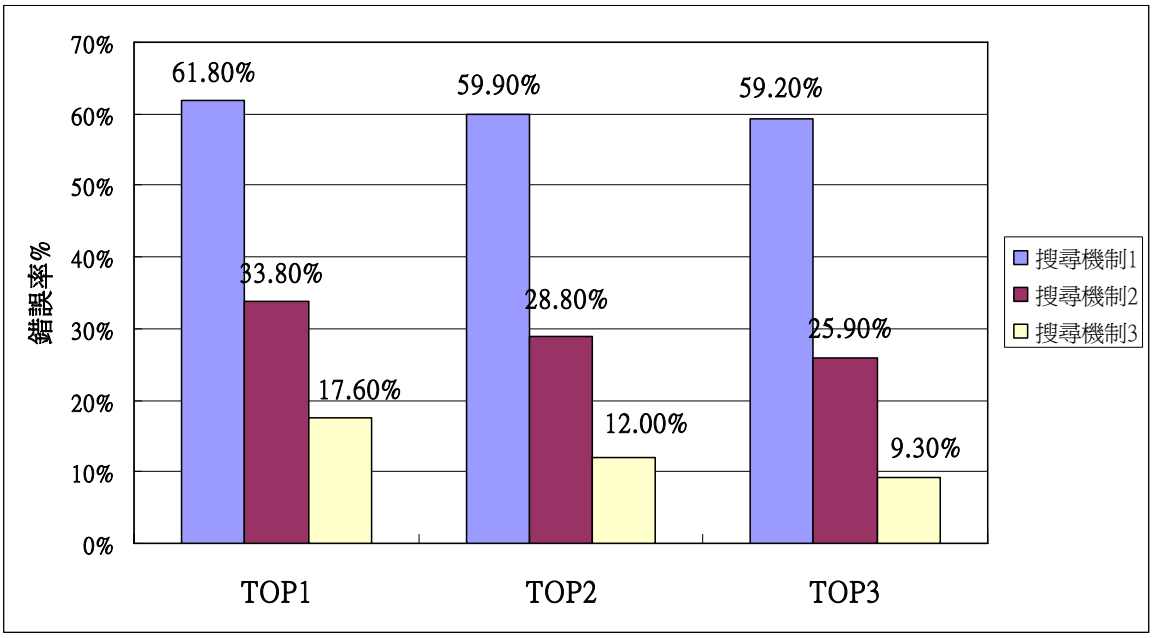


圖 4-1：加入偏差補償值與可忽略前置詞搜尋路徑的比較

首先我們先實驗測試語料對於有無英語模型偏差補償值與有無忽略搜尋前置詞路徑的效果，如上圖 4-1。有對英語模型相似度分數乘上偏差補償值的搜搜尋機制 2 與搜尋機制 3 都比搜尋機制 1 好上許多。在第一名的相對錯誤改善率方面，搜尋機制 2 比搜尋機制 1 改善了 45.2%，而搜尋機制 3 比搜尋機制 1 改善了 71.5%。而針對有加上忽略搜尋前置詞路徑，在第一名的相對錯誤改善率方面，搜尋機制 3 比搜尋機制 2 改善了 47.9%。

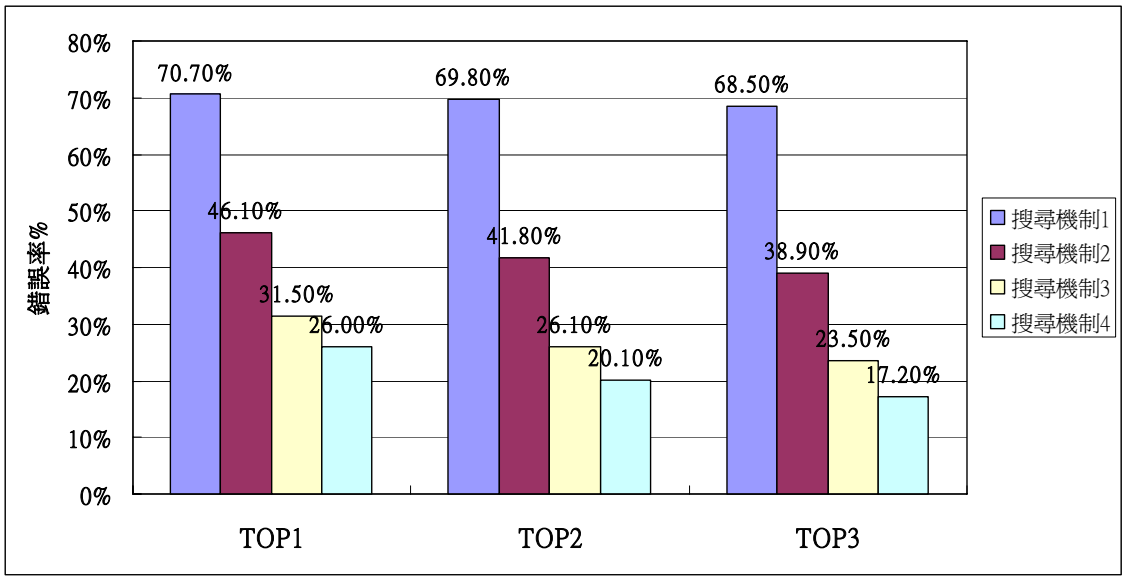


圖 4-2：加入偏差補償值、忽略前置詞搜尋路徑與無關詞垃圾模型路徑比較

接下來我們把 10 個前置詞中的 5 個華語前置詞全部刪掉，製造出語料中有說錯前置詞的情形，並且對於 4 個搜尋機制做實驗，如圖 4-2。對於有英文模型偏差補償值的搜尋機制 2 與 3，比較搜尋機制 1 在第一名的錯誤改善率爲 34.8%、55.4%，而加入忽略前置詞路徑後，搜尋機制 3 比搜尋機制 2 於第一名的錯誤率改善了 31.7%。再加上無關詞垃圾模型路徑之後，搜尋機制 4 比搜尋機制 3 在第一名的錯誤率改善了 17.5%。由此可知，當系統中搜尋機制越來越完整後，可針對使用者沒按規定說定義內前置詞的情形下，對主詞還有一定的辨識效能。

五、結論

在本論文中，我們先建立出一個詞典樹的架構，並且利用樹枝狀結構的概念，把辨識詞庫內每個辭彙的音節依序佈成樹狀。之後，我們把光束搜尋法應用於語音節點的搜尋與篩選，累積每個時間點的存活語音節點，進行相似值大小的排序。於實驗中，對測試語料所計算出的英語模型相似度分數乘上一個偏差補償值，可以拉近與華語模型產生出的相似度分數距離，並且大幅提升對於主詞辨識率。而當測試語料中含有無定義內前置詞的字句時，於前置詞搜尋機制中加入忽略搜尋前置詞路徑，讓測試語料於無前置詞的音框階段，語音節點可以進入此路徑，並且不破壞對於主詞語音節點的連結。

爲了讓系統的搜尋機制更加有彈性，我們於系統中加入無關詞垃圾模型路徑並且搭配無關詞垃圾模型的偏差補償值。改變定義內前置詞的數量，讓原本的測試語料有了說錯定義內前置詞的情形發生。系統在測試語料含有錯誤前置詞音框時，可以使所產生的錯誤節點路徑被無關詞垃圾模型路徑所吸收。經由實驗可得，有了此路徑對於說錯定義內前置詞的情形下，對於主詞的辨識率仍有一定的效果。

誌謝 本研究承中華電信研究所提供計畫費補助，謹誌謝忱。

參考文獻

[1] H. Hermansky, N. Morgan, "RASTA Processing of Speech," *IEEE Transactions SAP*, vol. 2, pp. 578-589, Oct 1994.

[2] C. L. Huang, C-H Wu, "Phone Set Generation Based on Acoustic Contextual Analysis for Multilingual Speech Recognition," *ICASSP*, vol. 4, pp. 1017-1020, 2007.

[3] C. L. Huang, C-H Wu, "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis," *Computers, IEEE Transactions*, vol. 56, pp. 1225-1233, 2007.

[4] Po-Yi Shih, Jhing-Fa Wang, and Hsiao-Ping Lee, "Acoustic and Phoneme Modeling Based on Confusion Matrix for Ubiquitous Mixed-Language Speech Recognition," *SUTC*, pp. 500-506, June 2008.

[5] 遊山銳, 簡世傑等人, "中英文混雜關鍵詞萃取技術," *TEPS*, pp. 66-79, 2004.

[6] A. L. Higgins, R. E. Wohlford, "Keyword Recognition Using Template Concatenation," *ICASSP,* vol. 10, pp. 1233-1236, 1985.

[7] J. R. Rohilcek, W. Roukos, and H. Gish, "Continuous Hidden Markov Models for Speaker Independent Word Spotting," *ICASSP*, pp. 627-630, 1989.

[8] R. Rose, D. Paul, "A Hidden Markov Model Based Keyword Recognition System," *ICASSP,* vol. 1, pp. 129-132, 1990.

[9] J. Caminero, C. de la Torre, L. Villarrubia, C. Martin, and L. Hernandez, "On-line Garbage Modeling with Discriminant Analysis for Utterance Verification," *ICSLP*, vol. 4, pp. 2111-2114, Oct 1996.

[10] T. Svendsen, F. K. Soong, and H. Pumhagen, "Optimizing Baseforms for HMM-Base Speech Recognition," *Proceedings of EuroSpeech*, pp. 783-786, 1995.

[11] X. Lingyun, D. Limin, "Efficient Viterbi Beam Search Algorithm Using Dynamic Pruning," *ICOSP*, vol. 1, pp. 699-702, 2004.

[12] H. C. Wang, F. Seide, C. Y. Tseng, and L. S. Lee, "MAT2000 – Design, Collection, and Validation on a Mandarin 2000-speaker Telephone Speech Database," *ICSLP*, pp. 460-463, Beijing, China, 2000.

[13] http://www.aclclp.org.tw/doc/eat_brief.pdf.

強健性語音辨識中基於小波轉換之分頻統計補償技術的研究
A Study of Sub-band Feature Statistics Compensation Techniques Based on a Discrete Wavelet Transform for Robust Speech Recognition

范顯騰 Hao-teng Fan
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s96323516@ncnu.edu.tw

杜文祥 Wen-Hsiang Tu
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s96323905@ncnu.edu.tw

洪志偉 Jeih-weih Hung
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
jwhung@ncnu.edu.tw

摘要

本論文主要是發展語音特徵強健化技術，來改進雜訊環境下語音辨識的效能。我們改良原始全頻帶式的特徵序列統計正規化技術，使用著名的離散小波轉換來對語音特徵時間序列進行分頻帶的處理，進而發展出兩種新的特徵統計補償法，分別爲分頻式平均值與變異數正規化法與分頻式統計圖等化法。在這兩種新方法中，我們將經由離散小波轉換所得之分頻帶的序列，分別以平均值與變異數正規化法與統計圖等化法處理，再將處理後的各分頻帶之特徵序列，藉由反離散小波轉換組合成新的特徵序列。如此處理的特點爲，可以將特徵序列作不等切的調變頻帶切割，進而對語音辨識較重要的低調變頻帶作個別的強健性處理。從 Aurora-2 連續數字資料庫的實驗結果證實，我們提出的分頻式新方法在各種雜訊環境下都優於傳統全頻帶式之方法，與基礎實驗結果相比較，其相對錯誤降低率皆在 50%以上，顯示了我們所提出之新方法能十分有效地提昇語音特徵在雜訊環境下的強健性。

Abstract

The environmental mismatch caused by additive noise and/or channel distortion often degrades the performance of a speech recognition system seriously. Various robustness techniques have been proposed to reduce this mismatch, and one category of them aims to normalize the statistics of speech features in both training and testing conditions. In general, these statistics normalization methods deal with the speech feature sequences in a full-band manner, which somewhat ignores the fact that different modulation frequency components

have unequal importance for speech recognition.

With the above observations, in this paper we propose that the speech feature streams be processed in a sub-band manner. The processed temporal-domain feature sequence is first decomposed into non-uniform sub-bands using discrete wavelet transform (DWT), and then each sub-band stream is individually processed by the well-known normalization methods, like mean and variance normalization (MVN) and histogram equalization (HEQ). Finally, we reconstruct the feature stream with all the modified sub-band streams using inverse DWT. With this process, the components that correspond to more important modulation spectral bands in the feature sequence can be processed separately. For the Aurora-2 clean-condition training task, the new proposed sub-band MVN and HEQ provide relative error rate reductions of 20.32% and 16.39% over the conventional MVN and HEQ, respectively. These results reveal that the proposed methods significantly enhance the robustness of speech features in noise-corrupted environments.

關鍵詞：離散小波轉換、語音辨識、強健性語音特徵參數

keywords: speech recognition, discrete wavelet transform, robust speech features

一、緒論

近年來，語音處理之領域的學者持續地開發研究，使語音處理相關理論與技術不斷精進成熟，逐漸趨於實際應用的目的，就語音辨識(speech recognition)而言，其系統常因所在環境之雜訊干擾或是傳輸通道的效應，而使辨識效能受到明顯影響。針對這樣的問題，近年來的研究學者提出了一系列的環境強健性(environmental robustness)技術，藉此降低雜訊或通道干擾或凸顯語音的獨特成份，而達到明顯的改進效果，本論文的研究方向，即爲開發出新的降低雜訊與通道干擾之相關的語音強健性演算法。然而，跟過去相關之強健性技術較爲不同的是，我們採用了小波轉換(wavelet transform)，對於語音特徵之時間序列(temporal trajectory)加以處理，來改善語音特徵的強健性。

小波相關理論在訊號處理的範疇中雖已發展數十年，然而相對於其他許多理論而言，應用於在語音強健性處理之領域中仍偏少數，而其應用的方向大致上主要包含了：語音強化(speech enhancement)、語音端點偵測(voice activity detection, VAD)、強健性語音特徵(robust speech feature)與聽覺濾波器設計(auditory filter design)等。我們將它們簡述如下：

（一）語音強化(speech enhancement)

語音強化主要目的，通常是在一段訊號中，將雜訊抑制，並將語音訊號成份強調出來，常用的方式是假設雜訊在頻譜(spectrum)上具有較爲穩態(stationary)的特性，在頻域上將雜訊成份減低，例如設計一濾波器來過濾雜訊等。而以目前基於小波的信號強化方法，其中之一爲 Donoho[1]學者所提出使用小波收縮(wavelet shrinkage)的方式，其方法是由小波轉換所得之係數，經由門檻值的設定將雜訊適度地抑制。在其相關論文之實驗結果顯示了，透過小波轉換處理的語音強化效能比起之前所提出的傳統語音強化方法[2]要來的好。

（二）語音端點偵測(voice activity detection, VAD)

由於一段錄音(recording)裡可能包含有非語音的區段，如果一併辨識整段錄音，將會影響辨識處理的速度，並可能造成辨識精確度明顯下降。語音端點偵測(voice activity detection, endpoint detection)相關技術即是於決定出一段訊號中真正語音存在的位置。在傳統的作法上，以時域(time domain)而言，透過計算一段語音信號的能量(energy)或過零

率(zero-crossing rate)來決定含有語音成分的位置；在頻域(frequency domain)上，則通常是計算語音頻譜的熵(entropy)來獲得語音成分的資訊[3]。而小波在此方向上所提出的技術相對較多，譬如在文獻[4]中提到了使用小波轉換的係數能量比例判定語音及非語音(non-speech)成分，或是在另一[5]文獻裡提出計算小波係數之變異數，將其視爲一組隨機變數(random variable) 經由機率理論之結果判定，所得分類方法相較於之前方式能更精確判別出語音跟非語音之成份。

（三）強健性語音特徵擷取(robust speech feature extraction)

此類的語音處理技術方法目的是擷取不容易受到雜訊干擾的語音特徵參數，傳統的強健性語音特徵擷取技術大多數是在探討語音特徵的頻譜性質進而發展而得，換句話說，其所使用的轉換法爲有名的傅立葉轉換(Fourier transform)。然而小波處理也相繼應用於強健性語音特徵擷取技術上，例如，在[6]提出將原始梅爾倒頻譜特徵(mel-frequency cepstral co efficients, M FCC)中的離散餘弦轉換(discrete cosine trans form, DCT) 程序改變爲離散小波轉換(discrete wavelet transform, DWT)，其論文呈現的實驗結果顯示所得到的特徵比原始 MFCC 更具有雜訊環境之強健性。

（四）聽覺濾波器設計(auditory filter design)

一般而言，語音辨識中特徵參數求取程序裡所應用的語音聽覺濾波器組爲梅爾尺度(mel-scaled)的濾波器組，這些濾波器其分佈特性爲：1 kHz 頻率以下爲線性分佈，1 kHz 以上頻率爲非線性分佈，彼此相互部分重疊，其可近似模擬人耳聽覺效應。相對而言，小波處理之研究學者[7]也提出了利用小波包(wavelet packet) 的特性來仿效人耳聽覺效應，其適當透過一連串小波包轉換所切割的部份頻帶，選擇出能趨近於人耳聽覺的濾波器組效應，而由於小波處理所得之彼此頻帶間都假設爲不相關，即爲互不影響，因此所切割出來的各頻率範圍的語音信號都涵蓋了獨立的辨識資訊，其中的實驗結果驗證了以上的處理可以優於傳統的梅爾濾波器組處理，達到將語音辨識精確度提升的目的。

在本論文中，所發展出的新技術，並不同於上述所提的幾個傳統小波處理所應用的方向，而是著重於將小波處理其特殊的分頻技術適當地運用於語音特徵時間序列(temporal trajectory) 上，結合各種統計正規化的技術，來處理小波轉後各子頻帶的特徵時間序列，在之後的章節中我們將會逐步介紹此新技術，分析其主要觀念、作法與可能優於傳統技術的原因，並以一系列的實驗證實此新技術相對於傳統相近的技術而言，更能有效提昇語音辨識在雜訊干擾環境下的精確性。

本論文其餘的章節概要如下：在第二章裡，介紹目前常用之強健性特徵統計正規化法並探討傳統統計正規化法之可能缺失。在第三章，我們將簡要介紹離散小波轉換之分頻技術的實現，第四章爲本論文的重點，我們將在此章中介紹我們所提出的新方法，即兩種調變頻譜域的分頻統計特徵補償法：分頻帶平均值與變異數正規化法與分頻帶統計圖等化法，並對其初步效果加以介紹。在第五章，我們將執行一系列的語音辨識實驗，來驗證所提之新方法足以有效提昇語音特徵在雜訊環境下的強健性，最後，第六章則爲簡要結論，及未來可進一步研究的方向。

二、各種強健性技術介紹

在這裡我們首先目前常用之強健性特徵統計正規化法，之後探討傳統統計正規化法之可能缺失，並說明爲何使用小波轉換(discrete wavelet transform, DWT)改善這些問題。

由於語音辨識系統容易受到雜訊環境影響使得其辨識效能降低，因此語音處理相關研究的學者針對此雜訊干擾的問題，提出諸多的強健性技術，這些技術中有一大類是藉由正規化語音特徵的統計特性，來降低雜訊對語音特徵造成的失真。以下將介紹近年來在強健性語音辨識中常用的幾種語音特徵正規化技術。其中包含了：倒頻譜平均消去法

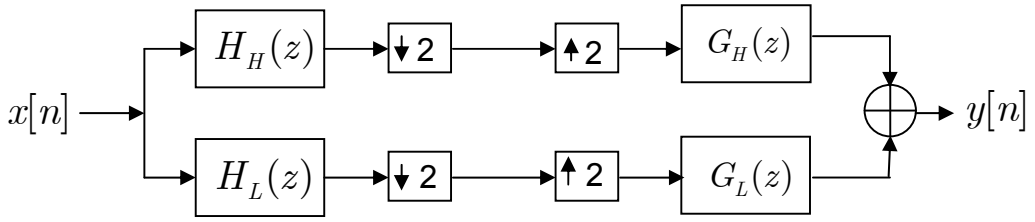
(cepstral mean subtraction, CMS)[8] 、倒頻譜平均值與變異數正規化法(cepstral mean and variance normalization, MVN)[9]、倒頻譜平均與變異數正規化法結合自動回歸動態平均濾波器法(cepstral m ean and variance norm alization p lus auto-regress ive m oving average filtering, MVA)[10]與統計圖正規化法(histogram equalization, HEQ)[11]等。

上述各種的正規化技術中，皆是把單一維特徵序列之所有特徵視爲同一個隨機變數的取樣(sample)，進而直接估測此隨機變數之統計參數，譬如期望值(mean)、變異數(variance)與機率分佈(probability distribution)等。雖然程序上易於實現，卻相對忽略了一段語句之中，其特徵隨時間變化的特性，例如調變頻譜的資訊。從另一觀點來看，這些作法等同於將全部調變頻率之成份一併做處理。然而根據過去許多的研究發現，不同的調變頻譜成份對於語音辨識擁有不同的重要性，更精確地說，在 N.Kanedera 學者[12]詳細指出大部分的語音辨識資訊分布在 1 Hz 和 16 Hz 的調變頻率之間，且主要集中在 4 Hz 附近。因此，許多知名且成功的時間序列濾波器(temporal filters)[13,14] ，都是特別強調出這些重要的調變頻率成分，進而顯示能有效改善雜訊環境下語音辨識的效能。

而前面介紹的各種特徵統計正規化演算法，可能缺失在於無法有效突顯不同調變頻率成份對於語音辨識的重要性，因此我們希望能把一特徵時間序列中的不同頻率成份分離出來，進而個別處理，初步的構想是能對於調變頻率較重要之低頻的部份較精細的處理，相對比較不重要之高頻的部份則使用較粗略的方式處理。基於此目的，我們發現小波轉換是個十分有用的工具，優點爲其能對一頻率區域作不等分的切割，即將訊號其較低頻率部分使用較窄的濾波器過濾出來，而高頻部分則用較寬的濾波器得之，之後對於每個子頻帶的特徵序列作統計正規化法。這樣的程序，相較於傳統的全頻帶式的特徵統計正規化法，理應可以進一步提昇處理後之特徵的強健性。之後一系列的章節，我們將逐步介紹小波轉換之分頻理論以及所提出的分頻特徵統計正規化法，最後以實驗結果證實此分頻式正規化法優於傳統之全頻式正規化方法。

三、小波轉換之分頻技術理論的概述

在這一章中，我們將專門討論小波轉換運用於離散時間訊號(discrete-time signal)的分頻(frequency division) 技術，此應算是小波轉換最常被用以處理訊號的方向。首先我們考慮一組典型雙通道的正交鏡像濾波器(quadrature-mirror filter bank, QMF)[15]，如圖三中所示：



圖三 雙通道 QMF 濾波器組

其中 $H_H(z)$ 與 $H_L(z)$ 表示爲分析(analysis)濾波器之高通與低通的轉換函數(transfer function)， $G_H(z)$ 與 $G_L(z)$ 則爲合成(synthesis)濾波器之高通與低通的轉換函數，且它們須符合以下的條件：

$$G_L(z) = H_H(z) \quad , \quad G_H(z) = -H_L(-z) \tag{3-1}$$

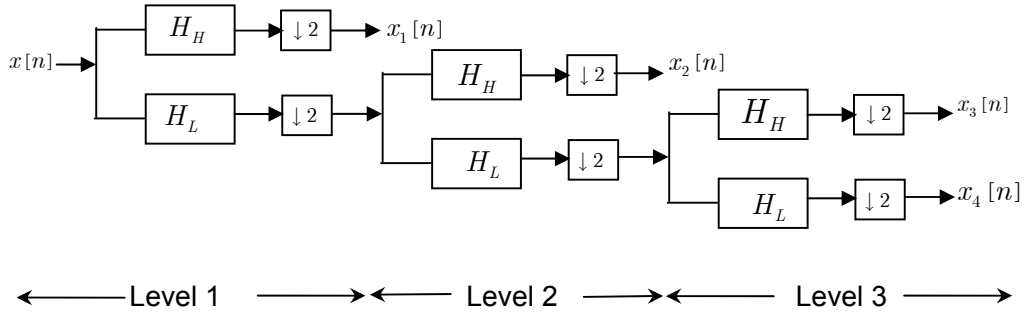
而在高通與低通分析濾波器之間存有以下關係：

$$H_H(z) = H_L(-z) \tag{式(3-2)}$$

意即其頻率特性爲 $H_H\left(e^{j(\omega-\pi)}\right) = H_L\left(e^{j\omega}\right)$ ，其意義在於高通與低通濾波器之頻率響應會

以$\omega = \frac{\pi}{2}$為中心形成左右對稱的圖形，在小波轉換中，即利用此形式的濾波器來對訊號作分頻處理。

圖四所表示了一訊號藉由上述之濾波器處理的分解程序(decomposition process) ，即離散小波轉換的分頻處理。其中，一連串的兩倍頻(octave-band)分析濾波組與之後的降低取樣(down-sampling)的組合通常被稱作二元樹(binary tree)結構，單一輸入序列經由分頻處理與降低取樣器(down-sampler)的轉換，輸出變為各子頻帶序列(sub-sequences)的集合。在圖四中，我們看到了一個三階(three-level)的二元樹分析濾波器組結構，其中高通($H_H(z)$)與低通($H_L(z)$)濾波器都具有完全重構(perfect reconstru ction)的雙通道(two-channel)特性，即訊號通過此兩濾波器之後，並未喪失任何資訊或引進未知的干擾訊號，而得以將分頻後的訊號完美重建回原始訊號。另外，如果輸入此濾波器組的訊號$x[n]$長度為N，在第一階高通分析濾波器之輸出$x_1[n]$即約為$N/2$，而再下一階高通分析濾波器輸出$x_2[n]$約為$N/4$，如此重複這程序，就可以得到所有階層之濾波器的輸出。表一列出了各層濾波器的其頻帶範圍及輸出訊號的長度。以上所述之兩倍頻(octave)完全重構 QMF 濾波器組對輸入訊號的處理程序，即為離散小波轉換(discrete wavelet transform, DWT)，由上述可知，如果所用之濾波器組的階層數為L（相當於L層的離散小波轉換），則輸出的總分頻訊號的數目則為$L+1$個。



圖四　離散小波轉換的分解程序圖（階層數為 3）

表一、三層離散小波轉換(DWT)每一階層的輸出訊號點數及相對應的頻率範圍
（$x[n]$取樣頻率為F_s Hz ）

訊號	總點數	頻率範圍
$x[n]$	N	$\left[0, F_s/2 \text{ Hz}\right]$
$x_1[n]$	$N/2$	$\left[F_s/4 \text{ Hz}, F_s/2 \text{ Hz}\right]$
$x_2[n]$	$N/4$	$\left[F_s/8 \text{ Hz}, F_s/4 \text{ Hz}\right]$
$x_3[n]$	$N/8$	$\left[F_s/16 \text{ Hz}, F_s/8 \text{ Hz}\right]$
$x_4[n]$	$N/8$	$\left[0, F_s/16 \text{ Hz}\right]$

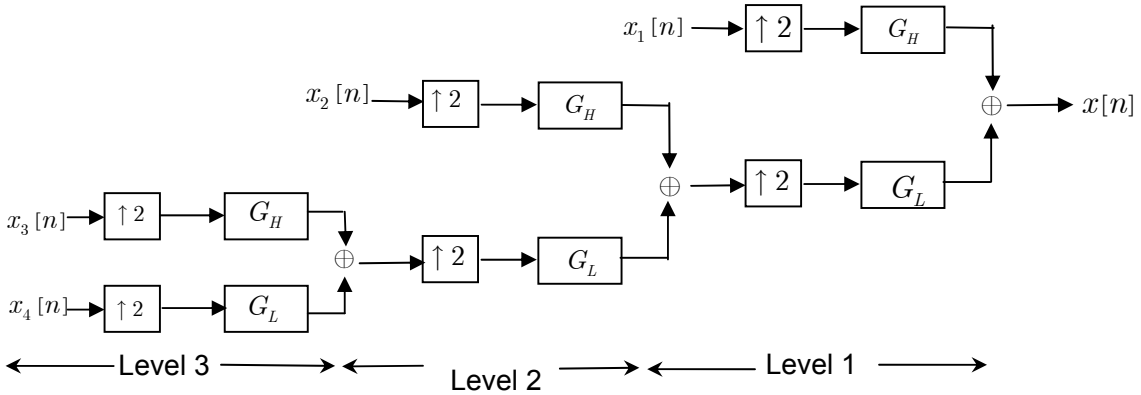
從上表一可知，如果序列$x[n]$涵蓋的頻率範圍為[0, $F_s/2$ Hz] ，其中F_s為$x[n]$的取樣頻率，則經由第一階正交鏡像濾波器組之高頻輸出$x_1[n]$，頻率範圍為$\left[F_s/4 \text{ Hz}, F_s/2 \text{ Hz}\right]$，依此類推，逐步往低頻率部份做不等分切割，隨著頻率越高，其

頻寬則越大。由上所述，離散小波轉換的第 k 個輸出 $x_k[n]$ ，相當於是原始序列 $x[n]$ 與第 k 個帶通濾波器之脈衝響應(impulse response)相互摺積(convolution)的結果，如式(3-3)所示：

$$x_k[n]=\begin{cases}\sum_{m=-\infty}^{\infty} h_{k,1}\left[2^{k+1}n-m\right]x[m], & 0\le k\le L\text{-}1,\\ \sum_{m=-\infty}^{\infty} h_k\left[2^k n-m\right]x[m], & k=L.\end{cases} \qquad \text{式(3-3)}$$

其中 $h_{k,1}\left[2^{k+1}n\right]$ 與 $h_k\left[2^k n\right]$ 爲原始脈衝響應 $h_{k,1}[n]$ 與 $h_k[n]$ 降低取樣而得，而高通濾波器之輸出，稱爲細節(detail)係數；低通濾波器之輸出則稱爲近似(approximation)係數。

若要藉由所有子頻帶訊號的集合得到原始序列 $x[n]$ ，其過程稱爲重建程序(reconstruction process)，此恰爲前述之分解程序的反程序(inverse process)，即使用所得之 $\{x_k[n]\}$ 經 L 階兩倍頻完全重構 QMF 合成濾波器組逐層處理，此過程即爲反離散小波轉換(inverse discrete wavelet transform, IDWT)，如下圖五所示：



圖五　反離散小波轉換的重建程序圖（階層數爲 3）

還原程序其數學式如式(3-4)：

$$x[n]=\sum_{k=0}^{L-1}\sum_{m=-\infty}^{\infty} g_{k,1}\left[n-2^{k+1}m\right]x_k[m]+\sum_{m=-\infty}^{\infty} g_k\left[n-2^L m\right]x_L[m], \qquad (3\text{-}4)$$

其中 $g_{k,1}\left[2^{k+1}n\right]$ 與 $g_k\left[2^k n\right]$ 分別爲原始脈衝響應 $g_{k,1}[n]$ 與 $g_k[n]$ 提升取樣而得。圖五之還原程序，即是將各子頻帶的訊號以提升取樣(up-sampling)的方式增加序列點數，再經過高通($G_H(z)=H_H(z)$)與低通($G_L(z)=H_L(z)$)之合成濾波器處理，如果第三階輸入訊號點數爲 $N/8$ ，則在第三階輸出訊號點數約爲 $N/4$ ，而第二階輸出訊號點數約爲 $N/2$ ，如此重覆此程序，則最後所得之訊號爲原始 N 點之訊號 $x[n]$:

以上所述爲小波轉換之分析(analysis)與合成(synthesis)程序，經由此轉換後，訊號被分解成各個子頻帶之訊號，如表一所示，低頻部分的子頻帶頻寬較小，而高頻部分的子頻帶頻寬較大。藉由以上所述的離散小波轉換程序，我們可以將語音特徵時間序列作分頻的處理，進而針對不同調變頻帶成分的語音特徵序列分別作處理，在下一章裡，我們將介紹其對應的的分頻式特徵統計補償法。

四、分頻帶特徵統計正規化法

在這一章中，我們首先在第一節介紹所新提出之分頻帶特徵統計補償法的步驟及特性，接著在第二節中，我們將以一語句爲例，驗證所提之新方法足以有效降低雜訊對語音調變頻譜之干擾。

(一) 分頻帶特徵統計正規化法的步驟說明

假設一段語句(utterance)的某一維梅爾倒頻譜語音特徵以下式(4-1)表示:

$$\left\{x^{(m)}[n];1<n\le N\right\},\ 0\le m\le M-1\ , \tag{4-1}$$

其中 N 爲此特徵序列的總音框數，M 表示每一音框中的特徵總數。此特徵序列相當於涵蓋了全調變頻帶(full-band)的語音資訊，然而，由前面章節所述，不同的頻帶成份，對於語音辨認的重要性有所不同，基於此項理由，這裡我們使用分頻的技術，將此特徵序列分解成各不同頻率的成分，如以下步驟（爲了簡易說明起見，我們在之後的討論中，將省略式(4-1)中代表不同維特徵的上標"(m)"，因爲我們是對每一個不同維的特徵序列皆作同樣處理)：

首先，我們將原始特徵序列 $\{x[n]\}$ 切割成 L 個分頻帶且假設每一分頻帶都爲各自獨立，而每一頻帶中的序列表示爲 $\left\{x_\ell[n]\right\},1\le\ell\le L$ ，此切割頻帶的方法是將原始特徵通過一倍頻(octave-band)帶通濾波器組，每一子頻帶訊號再作降低取樣(down-sampling)處理，此步驟等效於執行 $(L-1)$ 階的離散小波轉換(discrete wavelet transform, DWT) 於特徵序列 $x[n]$ 上。另外，假設特徵序列 $\{x[n]\}$ 音框取樣率爲 F_s (Hz) ，則其調變頻譜頻率範圍爲 $\left[0,F_s/2\right]$ ，因此，第ℓ個分頻帶序列的頻率範圍，可被近似表示成式(4-2)：

$$\begin{cases}\left[0,\dfrac{1}{2^{L-1}}\left(F_s/2\right)\right] & \text{if } \ell=1\\ \left[\dfrac{2^{\ell-2}}{2^{L-1}}\left(F_s/2\right),\dfrac{2^{\ell-1}}{2^{L-1}}\left(F_s/2\right)\right] & \text{if } \ell=2,3,\cdots,L\end{cases} \tag{4-2}$$

在 DWT 程序中，其方式是將一主頻帶依頻寬先等切爲兩個副頻帶，然後保持高頻帶不動，將低頻帶再等切成兩個副頻帶，如此反覆進行，因此相當於低頻部份會使用較多個頻寬較小的濾波器，而高頻部份則用較少個頻寬較大的濾波器，而因爲 DWT 程序中的降低取樣(down-sampling)的運算，所以每一分頻帶的序列 $\left\{x_\ell[n]\right\}$ 長度約正比於頻寬的大小。

接著，將上步驟所得的分頻帶序列 $\left\{x_\ell[n]\right\}$ 做特徵統計正規化，得到新的分頻帶序列，表示爲 $\left\{\tilde{x}_\ell[n]\right\}$ ，其特徵統計正規化的方式是將每一語句之子頻帶特徵 $\left\{x_\ell[n]\right\}$ 的統計量，譬如平均值(mean)、變異數(variance)或是更高階的動差(moments)作處理，使新的特徵參數 $\left\{\tilde{x}_\ell[n]\right\}$ 的統計量等同或逼近一目標(target)統計量，而此目標統計量是由乾淨訓練語料庫中，所有語句之子頻帶特徵 $\left\{x_\ell[n]\right\}$ 估測計算而得。在這裡我們使用的特徵統計正規化法有兩種，分別爲倒頻譜平均值與變異數正規化法(MVN)與統計圖等化法(HEQ)，以 MVN 法而言，所得新的分頻帶序列 $\left\{\tilde{c}_\ell[n]\right\}$ 可表示爲下式(4-3)：

$$\tilde{x}_\ell[n]=\left(\frac{x_\ell[n]-\mu_{\ell,s}}{\sigma_{\ell,s}}\right)\times\sigma_{\ell,t}+\mu_{\ell,t} \tag{4-3}$$

其中 $\mu_{\ell,s}$ 與 $\sigma_{\ell,s}^2$ 分別爲目前處理的單一(single)分頻帶序列 $\left\{x_\ell[n]\right\}$ 的平均值與變異數，而 $\mu_{\ell,t}$ 與 $\sigma_{\ell,t}^2$ 爲目標(target)平均值與變異數，此目標平均值與變異數是由原始乾淨訓練語料庫中所有分頻帶特徵序列 $\left\{x_\ell[n]\right\}$ 估測而得。同樣地，如以 HEQ 作爲統計補償法，則 $\left\{\tilde{x}_\ell[n]\right\}$ 與 $\left\{x_\ell[n]\right\}$ 彼此關係爲下式(4-4)：

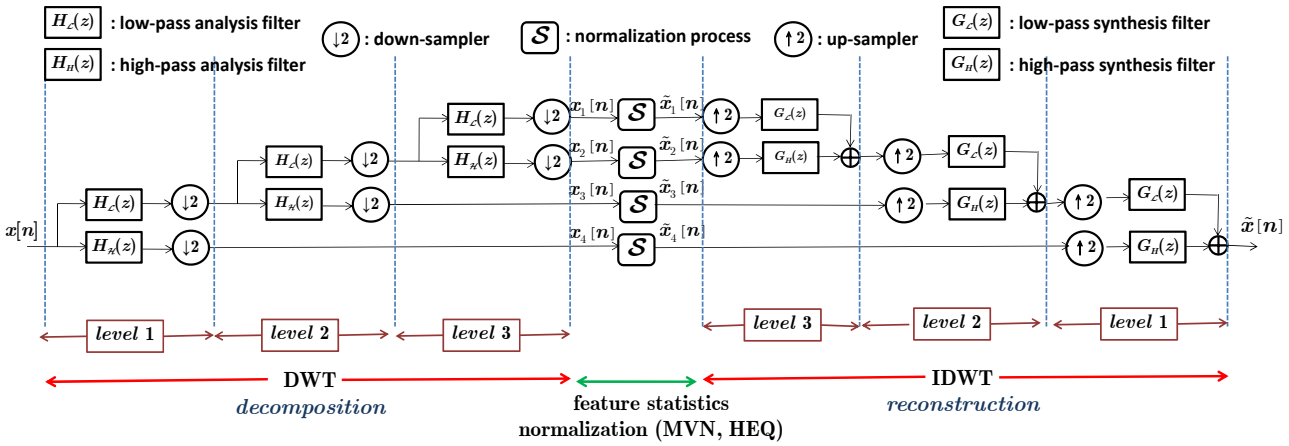
$$\tilde{x}_\ell[n]=F_{X,t}^{-1}\left(F_{X,s}\left(x_\ell[n]\right)\right) \tag{4-4}$$

其中 $F_{X,s}(.)$ 爲目前處理的單一分頻帶序列 $\left\{x_\ell[n]\right\}$ 所估測的機率分佈函數(probability distribution function)，而 $F_{X,t}(.)$ 是由原始乾淨訓練語庫中所有分頻帶特徵序列 $\left\{x_\ell[n]\right\}$ 所

估測而得的機率分佈函數。

最後，將所有的分頻帶序列$\{\tilde{x}_\ell[n]\}$（包含了更新過後與未更新的分頻帶序列）透過$(L-1)$階反離散小波轉換(inverse discrete wavelet transform, IDWT)，重建爲新的特徵時間序列，此即爲我們最後使用之語音特徵序列$\{\tilde{x}[n]\}$。

上述分頻帶統計正規化法的流程圖繪於下圖六：



圖六 分頻帶特徵統計正規化法的運作程序圖

爲了在之後的討論中，有效區隔傳統方法與所提出的新方法，對傳統全頻帶(full-band)的特徵統計正規化法 MVN 與 HEQ，我們分別稱之爲 FB-MVN 與 FB-HEQ，而如式(4-3)與(4-4)中分頻(sub-band)處理的特徵統計正規化法，我們則分別稱爲 SB-MVN 和 SB-HEQ。相較於傳統的全頻帶統計補償法，我們所提出之分頻帶統計補償法有以下幾點相異之處：

1. 傳統的全頻帶 MVN(FB-MVN)法中，任一特徵序列的平均值與變異數通常分別被正規化爲 0 與 1，但對於 SB-MVN 而言，不同分頻帶的特徵序列並不擁有相同的目標平均值與變異數，因此不同分頻帶特徵序列即使在正規化後，仍保有彼此統計特性的差異。相同地，SB-HEQ 也是具有此特性，不同的分頻帶特徵序列對應至不同的目標機率分佈函數。
2. 在 SB-MVN 與 SB-HEQ 中，可任意選擇某些分頻帶序列來作正規化。一般而言，對於語音辨識來說，低(調變)頻率的成分，包含的語音鑑別資訊較多，因此我們通常優先選擇低頻率的分頻帶特徵加以正規化。但是，如果有些非穩態雜訊(non-stationary noise)存在於高調變頻率的區域，爲了降低此類雜訊干擾，就須將高頻的分頻帶考慮進去一同處理。
3. 由於 DWT 程序中的降低取樣(down-sampling)步驟，我們所需處理之所有分頻帶序列的特徵總數近似等同於原始序列的特徵總數，因此處理上並不會因爲增加分頻帶的數目而使計算複雜度大幅提升。但若以傳統的分頻濾波器組(filter-bank)之方法，所需處理的總特徵數會明顯隨分頻帶的個數而增加，相對而言，其運算的複雜度會因此大幅提高。

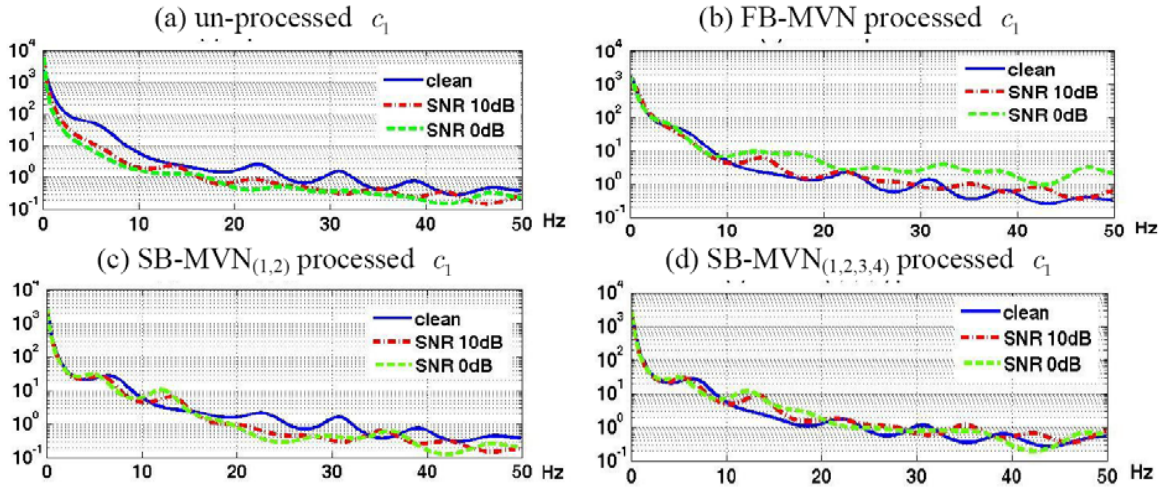
（二）分頻帶特徵統計正規化法的初步效能討論

在這裡，我們將所提出的分頻帶統計正規化法跟原始之全頻帶統計正規化法作初步的效能比較，根據這些方法在一語音特徵序列之調變頻譜的失真改善程度，來評估這些方法的效能。我們使用 AURORA-2 資料庫[20]裡的 MAH_27O6571A 語音檔，然後加入不同訊雜比(SNR)的地下鐵(subway)雜訊，繼而加以處理。

在我們所提出之方法中，初步使用了三階的 DWT 轉換，將整個調變頻帶[0, 50 Hz]切割出四種分頻帶範圍，分別是[0, 6.25 Hz]、[6.25 Hz, 12.5 Hz]、[12.5 Hz, 25 Hz]和[25 Hz, 50 Hz]，（由於特徵音框取樣率爲 100 Hz，因此特徵序列涵蓋之頻率範圍爲[0, 50 Hz]）。在之後討論的每個頻帶之正規法中，我們在方法名稱右下方使用下標數字來表示被正規化的頻帶，例如 SB-MVN$_{(1,2)}$與 SB-HEQ$_{(1,2)}$表示了第一個分頻帶（[0, 6.25 Hz] ）與第二個分頻帶([6.25 Hz, 12.5 Hz])使用了 MVN 或 HEQ 處理，剩餘的兩個高頻帶([12.5 Hz, 25 Hz]和[25 Hz, 50 Hz])則維持不動，而 SB-MVN$_{(1,2,3,4)}$與 SB-HEQ$_{(1,2,3,4)}$ 表示了全部四個分頻帶皆個別以 MVN 或 HEQ 處理。

首先，我們對於全頻帶與各種分頻帶之 MVN 法的處理結果加以討論。圖七(a)(b)(c)(d)分別表示爲原始未處理之第一維 MFCC(c_1)特徵序列、FB-MVN、SB-MVN$_{(1,2)}$與 SB-MVN$_{(1,2,3,4)}$處理後之 c_1 序列之功率頻譜密度(power spectral density, PSD)曲線。

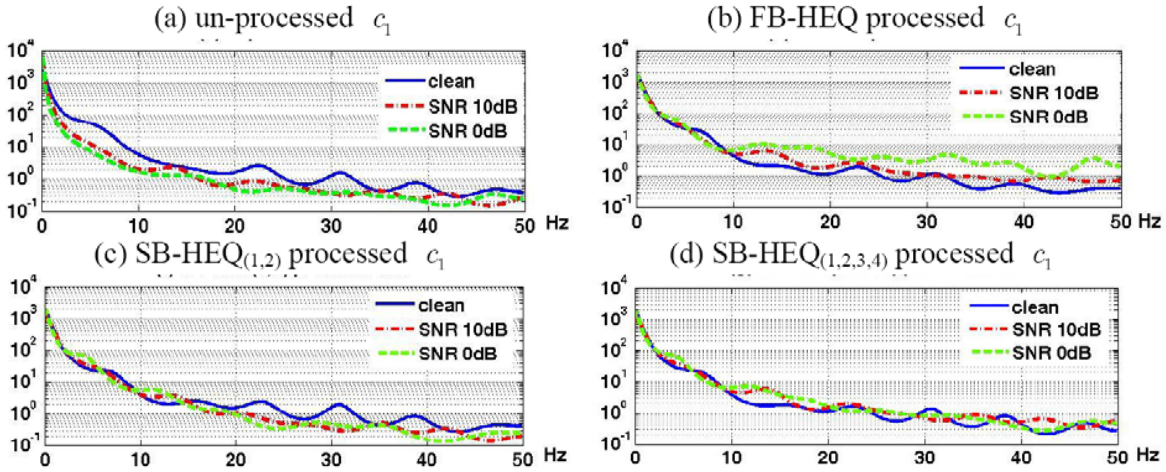
在圖七(a)中，可看出不同 SNR 值下(clean, 10 dB 與 0dB)之未處理過的 c_1 序列，其 PSD 曲線，受到加成性雜訊(additive noise) 的影響，存在嚴重的失真情形。而經由圖七(b)可看出，FB-MVN 處理後之 c_1 序列，在較低的調變頻率[0, 10 Hz]之間，其 PSD 失真的情況很明顯降低，但在高調變頻率範圍[10Hz, 50 Hz]，PSD 失真的情形並沒有太大的改善。圖七(c)爲 SB-MVN$_{(1,2)}$所得之特徵序列之 PSD 圖，其所處理的頻帶分別爲[0, 6.25 Hz]和[6.25 Hz, 12.5 Hz] ，從此圖可以發現，約在調變頻率 20 Hz 以下，其 PSD 失真情形相對減低，但在未處理的調變頻率範圍[12.5 Hz, 50 Hz]，同樣存有明顯的失真情況。圖七(d)爲 SB-MVN$_{(1,2,3,4)}$所得之特徵序列之 PSD 圖，其所處理的頻帶分別爲[0, 6.25 Hz]、[6.25 Hz, 12.5 Hz]、[12.5 Hz, 25 Hz]與[25 Hz, 50 Hz]，很明顯可看出在全部的調變頻率範圍，其 PSD 失真的情況皆有效降低。



圖七 (a) 原始 c_1 特徵序列及(b)FB-MVN、(c)SB-MVN$_{(1,2)}$與(d)SB-MVN$_{(1,2,3,4)}$ 作用在不同訊雜比下之 c_1 特徵序列之功率頻譜密度曲線圖

接下來，圖八(a)(b)(c)(d)分別表示爲原始未處理之第一維 MFCC(c_1)特徵序列、FB-HEQ、SB-HEQ$_{(1,2)}$與 SB-HEQ$_{(1,2,3,4)}$處理後之 c_1 序列之 PSD 曲線，其中括弧中的數字表示所處理的頻帶。比較圖八(a)與圖八(b)可知，對於較低的調變頻率範圍[0, 10 Hz] ，FB-HEQ 可有效降低 PSD 之失真，但對於其他調變頻率範圍[10 Hz, 50 Hz] ，PSD 失真的情形並沒有獲得太大的改善。圖八(c)爲 SB-HEQ$_{(1,2)}$所得之特徵序列之 PSD 圖，其所處理的頻帶分別爲[0, 6.25 Hz]與[6.25 Hz, 12.5 Hz] ，在此圖中，可以發現約在調變頻率

20 Hz 以下之 PSD 失真現象相對被減低，但在其他調變頻率範圍，仍有明顯的失真情況。跟之前圖七(c)SB-MVN$_{(1,2)}$的效果比較，可看出 SB-HEQ$_{(1,2)}$優於 SB-MVN$_{(1,2)}$ ，更有效降低約在頻率 20 Hz 以下的 PSD 失真度。圖八(d)爲 SB-HEQ$_{(1,2,3,4)}$所得之特徵序列之 PSD 圖，其所處理的頻帶個別爲[0, 6.25 Hz]、[6.25 Hz, 12.5 Hz]、[12.5 Hz, 25 Hz]與[25 Hz, 50 Hz] ，從此圖很明顯可看出全部的調變頻率範圍之 PSD 曲線，其失真的情況皆已有效降低。類似之前的狀況，當比較圖八(d)與圖七(d)時，可看出 SB-HEQ$_{(1,2,3,4)}$在降低 PSD 失真的性能上優於 SB-MVN$_{(1,2,3,4)}$。



圖八 (a) 原始c_1特徵序列、(b)FB-HEQ、(c)SB-HEQ$_{(1,2)}$與(d)SB-HEQ$_{(1,2,3,4)}$ 作用在不同訊雜比下之c_1特徵序列之功率頻譜密度曲線圖

五、調變頻譜分頻帶正規化法的辨識實驗結果與討論

本章主要內容爲呈現並分析一系列的強健性特徵技術所得之語音辨識的效果，這些技術包括了傳統的全頻式特徵統計正規化法、我們所新提出的分頻式 MVN(SB-MVN)法與分頻式 HEQ(SB-HEQ)法。

（一）實驗環境與架構設定

本辨識實驗所採用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute, ETSI) 所發行的語料庫 AURORA-2[16]，內容是以美國成年男女所錄製的一系列連續的英文數字字串，測試語音本身加上各種加成性雜訊或通道效應的干擾。加成性雜訊共有八種，分別是地下鐵(subway)、人聲(babble)、汽車(car)、展覽會館(exhibition)、餐廳(restaurant)、街道(street)、飛機場(airport)和火車站(train station)雜訊等；而通道效應有兩種，分別爲 G712 和 MIRS。雜訊比例的大小包含了乾淨無雜訊的狀態(clean)，以及六種不同雜訊比(signal to noise ratio, SNR)，分別是 20 dB、15 dB 、10 dB 、5 dB 、0 dB 與-5 dB ，因此我們可以觀察分析不同雜訊環境下對於語音辨識的影響。由於雜訊的不同，測試環境可分爲 Set A、Set B 與 Set C 三組。

在辨識中所使用的聲學模型是由隱藏式馬可夫模型工具(Hidden Markov Model Tool Kit, HTK)[17] 訓練而得，包括了 11 個數字模型(zero, one, two,…, nine 及 oh)以及靜音(silence)模型，每個數字模型則有 16 個狀態，各狀態包含 20 個高斯密度混合。

（二）全頻帶補償法與各種分頻帶正規化法之實驗結果

本章節實驗採用梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)

特徵參數13維(c0~c12)，加上一階與二階差量，總共爲39維特徵參數。在表三中，我們呈現了基礎實驗(baseline)、各種分頻式 SB-MVN 與 SB-HEQ、全頻式 FB-MVN 和 FB-HEQ 作用在原始 MFCC 特徵上所得的平均辨識結果（不同種辨識環境的平均辨識率及相對改善率），其中 RR1和 RR2分別爲相較於基礎實驗和全頻帶法之相對錯誤降低率(relative error rate reductions)。表四列出在各種不同的 SNR 值下的各種方法的平均辨識率，而圖九簡要畫出各方法平均辨識率的比較圖。

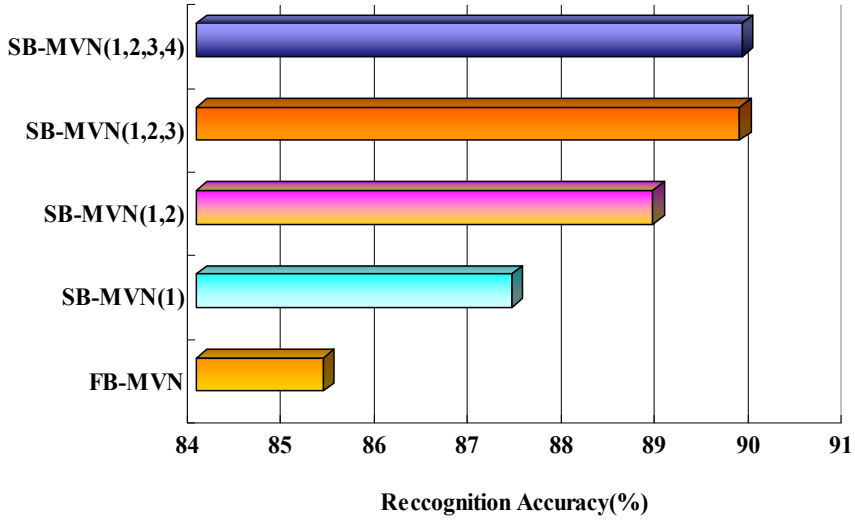
表三、各分頻帶方法與全頻帶方法的平均辨識率(%)與相對錯誤降低率(%)

Method	Set A	Set B	Set C	Avg.	RR1	RR2
Baseline	71.92	68.22	77.61	71.58	—	—
FB-MVN	85.03	85.56	85.60	85.36	48.49	—
SB-MVN$_{(1)}$	86.87	87.90	87.37	87.38	55.59	13.80
SB-MVN$_{(1,2)}$	87.28	90.23	89.44	88.89	60.91	24.11
SB-MVN$_{(1,2,3)}$	89.44	90.31	89.61	89.82	64.18	30.46
SB-MVN$_{(1,2,3,4)}$	89.47	90.31	89.62	**89.84**	**64.25**	**30.60**
FB-HEQ	87.59	88.84	87.64	88.10	58.13	—
SB-HEQ $_{(1)}$	87.70	89.31	87.81	88.37	59.08	2.27
SB-HEQ $_{(1,2)}$	89.22	90.55	90.23	89.95	64.64	15.55
SB-HEQ$_{(1,2,3)}$	89.51	90.75	89.54	90.01	64.85	16.05
SB-HEQ $_{(1,2,3,4)}$	89.51	90.83	89.57	**90.05**	**64.99**	**16.39**

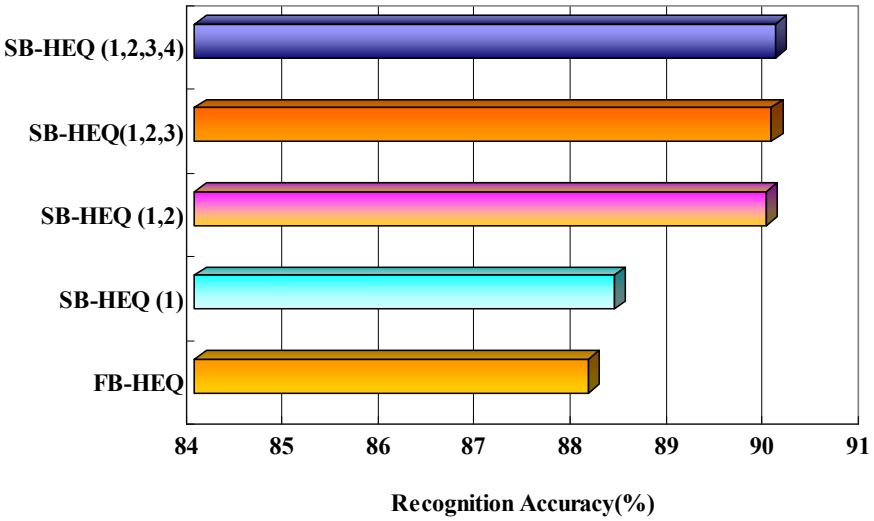
表四、所有不同 SNR 值雜訊環境下的平均辨識率(%)

Method	clean	20dB	15dB	10dB	5dB	0dB	-5dB
Baseline	99.79	95.80	88.15	73.81	56.32	43.82	40.13
FB-MVN	99.82	98.73	96.83	91.88	79.52	59.80	46.70
SB-MVN$_{(1)}$	99.79	98.67	96.98	92.24	82.42	66.60	51.95
SB-MVN$_{(1,2)}$	99.80	98.97	97.76	94.33	86.40	70.99	53.80
SB-MVN$_{(1,2,3)}$	99.78	98.99	97.75	94.57	86.59	71.19	53.69
SB-MVN$_{(1,2,3,4)}$	99.81	98.97	97.75	94.51	86.60	71.34	53.72
FB-HEQ	99.77	99.01	97.76	94.22	84.30	65.21	48.96
SB-HEQ $_{(1)}$	99.72	98.72	97.31	93.34	83.98	68.49	52.70
SB-HEQ $_{(1,2)}$	99.64	98.84	97.64	94.50	87.52	71.28	53.65
SB-HEQ$_{(1,2,3)}$	99.66	98.84	97.70	94.68	87.09	71.74	53.78
SB-HEQ $_{(1,2,3,4)}$	99.64	98.85	97.69	94.74	87.15	71.83	54.01

(a) 不同形式之平均值與變異數正規化法的辨識率比較



(b) 不同形式之統計圖等化法的辨識率比較



圖九 各分頻帶方法與全頻帶方法的平均辨識率(%)之綜合比較圖

從表三、表四和圖九可發現，我們所新提出的分頻帶正規化法，確實能有效提昇其雜訊環境下的強健性，其詳細現象如以下幾點：

1. 無論全頻帶與分頻帶正規化方法，相較於基本實驗而言，都有良好的改善效能，相對錯誤降低率都在 48%以上，除此之外，每一種 HEQ 的效果都比其相同形式的 MVN 來的好。相較於 MVN，HEQ 額外對於特徵高階動差做補償處理，所以整體來說，HEQ 更有助於改善雜訊環境所造成的特徵失真。
2. SB-MVN 的四種分頻模式效能都優於原始全頻式的 FB-MVN，此情況在 SB-HEQ 與 FB-HEQ 之間的比較也是如此。而 SB-MVN 和 SB-HEQ 相較於原始 FB-MVN 和 FB-HEQ 的相對錯誤降低率分別高達 30.60%與 16.39%，此結果顯示所提出的新分頻處理技術優於傳統全頻帶的處理，因此我們成功的驗證了之前章節的推論，即不同的調變頻譜成份對於語音辨識有不同的重要性，對不同頻帶分別作補償可帶來更好的效能。
3. 從表四中觀察在不同 SNR 值情況下的平均辨識率，我們可知在不受任何雜訊干擾

之匹配情況下，所有方法都有很高的辨識率，也就是說這些方法並不會降低與原始 MFCC 高鑑別度的特性。但在受到不同雜訊干擾之不匹配情況下，從表中可看出所有方法都能有效改善辨識效果，即增加原始 MFCC 特徵的強健性，在 SNR 值爲 20 dB 和 15 dB 時，分頻帶與全頻帶方法其效能差異並不顯著，如果訊雜比繼續下降時，可以發現到分頻式的 SB-MVN 與 SB-HEQ 平均辨識率明顯優於全頻式的 FB-MVN 和 FB-HEQ。

4. 對於不同分頻式的 SB-MVN 與 SB-HEQ，若只正規化最低的頻帶[0, 6.25 Hz](即 $\text{SB-MVN}_{(1)}$和 $\text{SB-HEQ}_{(1)}$)，其相對於基礎實驗就有顯著改善效果。當我們增加正規化的頻帶數目時，相對改善率都有明顯的成長，特別是在處理兩低頻帶([0, 6.25 Hz] 與[6.25 Hz, 12.5 Hz]) 後，也就是 $\text{SB-MVN}_{(1,2)}$和 $\text{SB-HEQ}_{(1,2)}$時，幾乎其效能已是最佳，如進一步再處理較高的分頻帶，如 $\text{SB-MVN}_{(1,2,3,4)}$或 $\text{SB-HEQ}_{(1,2,3,4)}$，所改善的效果就有限。此結果顯示與過去文獻互相吻合，即在 1 到 16Hz 之間的調變頻率成分，對於語音辨識而言是相對重要的。

六、結論與未來展望

我們提出了兩種分頻式特徵統計正規化技術，分別爲分頻式平均與變異數正規化法 (sub-band cepstral m ean and variance norm alization, SB-MVN) 與分頻式統計圖等化法 (sub-band histogram equalization, SB-HEQ) ，我們使用了著名的離散小波轉換(discrete wavelet tra nsform, DW T)來對語音的特徵時間序列作分頻處理，其特點在於利用 DWT 可以將對語音辨識較有幫助之調變頻譜低頻成份做較細緻的切割，高頻部份則相對應的切割區間數較少。之後，對每個子頻帶的特徵序列個別作統計正規化處理，再將所有子頻帶的特徵序列以反離散小波轉換，組合成新的特徵序列。經由連續數字之語音辨識實驗，顯示了上述之分頻式的新方法相對於傳統全頻式的方法，更能提昇雜訊干擾環境下語音辨識的精確度，相當於這些新方法能更有助於增加語音特徵的強健性。除了此優點外，此分頻式的新方法並未增加所須處理之語音特徵的個數，因此並不會因所分頻帶數目的增加，而大幅增加執行的複雜度。

在未來研究中，我們期望相關的實驗不只限制在連續數字辨識中，而是進一步應用在較大字彙之語音資料庫，探究其效能爲何。其次，我們亦將執行其他種類的特徵統計正規化於此所述的分頻帶特徵序列上，譬如倒頻譜增益正規化法(CGN)[18]、高階倒頻譜動差正規化法(HOCMN)[19]與倒頻譜形狀正規化法(CSN)[20]等，進一步驗證在這些方法中，分頻處理上是否能得到更好的效能。另外，我們也會嘗試使用各種不同的小波函數，分析每一種小波函數的特性，並探討這些函數對於所述之分頻式技術的影響，或是使用有別於小波轉換的小波包(wavelet pack et)[21]，研究這兩者對於所提之新方法上效能的差異。希望在未來，小波轉換能更成熟地應用於語音相關的分析研究上，使語音強健性技術能趨於成熟與多樣化，使語音處理更具理論性與實用性。

參考文獻

[1] D.L. Donoho, "De-noising by soft-thresholding", *IEEE Trans. on Information Theory*, vol. 41, no. 3, pp. 613-627, May 1995.

[2] S. F . Boll, "Suppression of aco ustic noise in speech u sing spectral subtraction ", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979

[3] B.-F . W u, K.-C. W ang, "Noise spectrum estim ation with entropy-based V AD i n non-stationary environm ents", *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E89-A, no. 2, Feb 2006

[4] J.K. Lee , C.D. Y oo, "Wavelet speech e nhancement based on voiced /unvoiced decision",

32nd Inter-Noise, pp. 4149-4156, Aug 2003
[5] X. Zhang, Z. Zhao and G. Zhao, "A speech endpoint detection method based on wavelet coefficient variance and sub-band amplitude variance", *International Conference on Innovative Computing, Information and Control*, vol. 3, pp. 83-86, 2006
[6] J.N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), vol. 3, pp. 1351-1354, 2000
[7] M. Siafarikas, T. Ganchev and N. Fakotakis, "Objective wavelet packets features for speaker verification" , *in Proc. International Conference on Spoken Language Processing* (ICSLP), pp. 2365-2368, 2002
[8] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, Apr 1981
[9] C.-P. Chen, K. Filaliy and J. A. Bilmes, "Frontend post-processing and backend model enhancement on The Aurora 2.0/3.0 databases' ', *in Proc. International Conference on Spoken Language Processing* (*ICSLP*), 2002
[10] C.-P. Chen and J. -A. Bilmes, "MVA processing of speech features", *IEEE Trans. on Audio, Speech, and Language Processing*, vol.15, no. 1, pp.257-270, Jan 2006.
[11] F. Hilger and H. Ney , "Quantile based histogram equalization for noise robust speech recognition", *in European Conference on Speech Communication and Technology* (*Eurospeech*), 2001
[12] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition ", *in European Conference on Speech Communication and Technology* (*Eurospeech*), 1997
[13] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, vol.2, no. 4, Oct. 1994.
[14] J- W. Hung and L-S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition", *IEEE Trans. on Audio, Speech and Language Processing*, vol.4, no. 3, May 2006
[15] D. Esteban and C. Galand. "Application of quadrature mirror filters to spilt-band voice coding schemes", *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), vol. 2, pp. 191-195, May 1977.
[16] H. G . Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions ", *in Proc. of ISCA IIWR ASR2000*, Paris, France, 2000
[17] http://htk.eng.cam.ac.uk/
[18] S. Yoshizawa et al., "Cepstral gain normalization for noise robust speech recognition," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), vol. 1, pp. I-209-12, May 2004
[19] C.-W. Hsu and L.-S. Lee, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition ", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), vol. 17, no. 2, pp. 205-220, Feb 2004
[20] J. Du and R.-H Wang, "Cepstral shape normalization (CSN) for robust speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4389-4392, April 2008
[21] R. R. Coifman and M. V . Wickerhauser. "Entropy-based algorithms for best basis selection", *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 713-718, March 1992

併合式倒頻譜統計正規化技術於強健性語音辨識之研究
A Study of Hybrid-based Cepstral Statistics Normalization Techniques for Robust Speech Recognition

何冠旻 Guan-min He
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s96323528@ncnu.edu.tw

杜文祥 Wen-Hsiang Tu
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
aero3016@ms45.hinet.net

洪志偉 Jeih-weih Hung
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
jwhung@ncnu.edu.tw

摘要

一語音辨識系統，在雜訊干擾的環境下，其辨識效能通常會明顯下降，如何改善此問題，是歷年來許多語音處理領域之學者所研究的重點。本論文也是針對此問題，提出了幾種新的語音強健性技術，來降低雜訊的干擾，以提升語音辨識的效能。

在本論文中，我們提出了新的語音特徵統計估測資訊演算法，藉此改進五種有名的強健性語音特徵正規化技術的效能，這些正規化技術包括了倒頻譜平均消去法(CMS)、倒頻譜平均值與變異數正規化法(CMVN)、高階倒頻譜動差正規化法(HOCMN)、倒頻譜增益正規化法(CGN)以及倒頻譜統計圖等化法(HEQ)等，這些技術皆被證明有效提升語音特徵之強健性。這些方法中的關鍵步驟之一，爲特徵統計資訊的估測。在傳統上，有三種統計估測的演算法，分別爲整句式、分段式與碼簿式演算法。在此論文中，我們討論這三種估測方式可能的優缺點，進而提出新的估測方式，稱作併合式統計估測演算法，其適當地組合碼簿式與整句式(或分段式)統計值估測法所求得的特徵統計資訊。在一系列之雜訊環境下的語音辨識實驗中，我們驗證了新提出的併合式統計估測法相對於傳統三種估測法而言，能夠更有效地改進上述五種語音特徵正規化技術的效能，而能得到更明顯的辨識精確率提昇。此外，我們所提出的併合碼簿與分段式的統計估測法具有近似線上運算的功能，因此更具有實際應用之價值。

Abstract

. Cepstral statistics normalization techniques have been shown to be very successful at improving the noise robustness of speech features. In this paper, we propose a hybrid-based scheme to achieve a more accurate estimate of the statistical information of features in these techniques. By properly integrating codebook and utterance/segment knowledge, the

resulting hybrid-based normalization methods significantly outperform conventional utterance-based, segment-based and codebook-based ones in recognition accuracy.

For the Aurora-2 clean-condition training task, the proposed hybrid codebook/segment-based histogram equalization (CS-HEQ) achieves an average recognition accuracy of 90.66%, which is better than utterance-based HEQ (87.62%), segment-based HEQ (85.92%) and codebook-based HEQ (85.29%). Furthermore, the high-performance CS-HEQ can be implemented with a short delay and can thus be applied in real-time online systems. A similar performance promotion can be also found in the methods of hybrid-based cepstral mean subtraction (CMS), cepstral mean and variance normalization (CMVN), cepstral gain normalization (CGN) and higher-order cepstral moment normalization (HOCMN).

關鍵詞：語音辨識、碼簿、特徵統計值估測法、強健性語音特徵參數

Keywords: speech recognition, codebook, feature statistics estimate, robust speech features

一、緒論

一語音辨識系統，當其應用於真實環境時，常因環境中諸多無法預期的變異性(variation)，而使其辨識效能受到明顯影響，爲了降低諸多的變異性所發展的各種技術，一般而言統稱爲強健性技術(robustness techniques)，而本論文中，我們則是主要著重於發展降低環境之雜訊干擾或通道效應的強健性技術。

在諸多降低錄音環境之雜訊干擾的強健性演算法中，有一大類的方法是將訓練與測試環境下的語音特徵其時間序列統計特性加以正規化(normalization)，以降低訓練與測試環境之間的不匹配，達到提昇辨識率的目的。在這些演算法中，首要步驟通常是估測語音特徵的統計值相關資訊，例如在 CMVN 法中所需估測的統計值爲平均值(mean)與變異數(variance)，而在 HEQ 法中必需估測出特徵時間序列的機率分佈(probability distribution)。這些統計估測值的精確度，直接影響到其對應之正規化演算法的效能。

在過去關於上述特徵統計正規化法的文獻中，根據不同的樣本來源，大致上有三種統計值估測法，分別爲整句式、片段式與碼簿式的估測法，顧名思義，第一種直接使用了整句的語音特徵來估測統計值，第二種則使用了部分(片段)的語音特徵，而第三種則間接透過語音特徵建立的碼簿[7]來作統計值之估測。我們發現這三種方法各有其優缺點，因此在本論文中，我們所提出的新統計估測技術，適當地併合碼簿與整句或片段的特徵資訊，希望得到更精準的語音特徵統計值，進而使各種特徵統計正規化法，在受雜訊干擾的環境中能夠更有效地提昇語音特徵的強健性，以改善辨認精確度。

本論文其餘的章節概要如下：在第二章，我們將簡要介紹過去三種特徵統計值估測法之步驟及其可能的優缺點。第三章則介紹我們新提出的兩種併合式(hybrid-based)的統計值估測法，及其如何運用於各種特徵統計正規化法中。在第四章中，我們介紹語音辨識實驗之語音資料庫、及新提出的兩種統計估測法在各種特徵統計正規化法的語音辨識結果及其相關討論。最後，第五章爲一簡要結論及未來研究之展望。

二、整句式、分段式與碼簿式特徵統計值估測法

我們在本論文中所討論的五種著名強健性語音特徵正規化技術，分別爲倒頻譜平均消去法(CMS)[1]、倒頻譜平均值與變異數正規化法(CMVN)[2,3]、高階倒頻譜動差正規

化法(HOCMN)[4]、倒頻譜增益正規化法(CGN)[5]以及倒頻譜統計圖等化法(HEQ)[6]等，這些技術所需使用的特徵統計相關資訊，例如：平均值、變異數、高階動差或是機率分佈等，可由不同的方法估測，而有不同的效果。在本章中，我們將介紹過去學者所提之主要三種特徵統計值估測法，包括了整句式(utterance-based)[8]、分段式(segment-based)[8]與碼簿式(codebook- based)[9]三類方法，及它們可能的優點與缺點。

(一) 整句式特徵統計值估測法

假設某單一語句之某一維特徵序列表示爲

$$\{x[n];1 \le n \le N\} \qquad \text{(式 2-1)}$$

其中N爲特徵序列之特徵總個數(即音框總數)。在整句式特徵統計值估測法裡，我們利用(式 2-1)所列之單句所有特徵，共同估測第m項特徵$x[m]$的統計值。換言之，我們假設$x[m]$對應至一隨機變數$X[m]$，進而假設整句特徵序列$\{x[n];1 \le n \le N\}$爲此隨機變數之樣本(sample)，根據這些樣本，我們可估測出$X[m]$此隨機變數的各種統計值，例如：

1. $X[m]$的期望值(平均值)爲

$$\mu_{X[m],(u)}[m] = \frac{1}{N}\sum_{n=1}^{N} x[n], \qquad \text{(式 2-2)}$$

2. $X[m]$的變異數(variance)爲

$$\sigma^2_{X[m],(u)}[m] = \frac{1}{N}\sum_{n=1}^{N}\left(x[n] - \mu_{X[m],(u)}[n]\right)^2, \qquad \text{(式 2-3)}$$

3. $X[m]$的第J階中央動差(central moment)爲

$$\xi^{(J)}_{X[m],(u)}[m] = \frac{1}{N}\sum_{n=1}^{N}\left(x[n] - \mu_{X[m],(u)}[n]\right)^J, \text{ 其中}J\text{爲任意之正偶數} \qquad \text{(式 2-4)}$$

4. $X[m]$的動態範圍(dynamic range)爲

$$d_{X[m],(u)}[m] = \max_{1\le n\le N}\{x[n]\} - \min_{1\le n\le N}\{x[n]\}, \qquad \text{(式 2-5)}$$

5、$X[m]$的機率分佈函數(probability distribution function)爲

$$F_{X[m],(u)}(z) = \frac{1}{N}\sum_{n=1}^{N} u(z - x[n]). \qquad \text{(式 2-6)}$$

其中，$u(\bullet)$爲單位步階函數(unit step function)，定義爲：

$$u(z) = \begin{cases} 1, & \text{if } z \ge 0 \\ 0, & \text{if } z < 0 \end{cases}$$

在以上五個式子中的各種統計值的代號中，我們以下標$_{(u)}$來代表這些統計值是由整句(utterance)的特徵估測而得，值得注意的是，以上所算之針對某一項特徵$x[m]$所得的各種統計值，事實上跟$x[m]$於序列之順序m無關，意即在整句式估測法而言，我們只須計算一次統計值，就可將此統計值供整句裡每項特徵$x[m]$作正規化使用。換言之，不同項特徵$x[m]$共用同一組統計值。接下來，我們將討論整句式特徵統計值估測法運用在語音特徵正規化技術之可能的優缺點。

- **整句式特徵統計值估測法運用在語音特徵正規化技術之優缺點**

在以前的文獻裡，大多數強健性語音特徵正規化技術所用的統計值，皆是藉由前述的整段語句之語音特徵所求得，雖然執行上簡單有效率，而且確實對語音特徵有明顯提升強健性的效果，但還是有一些潛在的缺點，例如，整句式語音特徵正規化技術無法達到即時處理(real-time processing)的要求，因爲對一連串的語音特徵序列而言，必須等到最後一個語音特徵得到之後，才能求取統計值。除此之外，隨著語句的不同，而產生的

語音特徵序列的長度(音框數)也不一樣，不同語句所包含的音素數目或種類，及其長度的變化可能相差很大，導致影響到所估測之統計資訊的準確性。

(二) 分段式特徵統計值估測法

首先，假設某單一語句特徵$\{x[n];1 \le n \le N\}$其中某一片段特徵序列表示爲

$$\{x[k];m-L \le k \le m+L\} \qquad \text{(式 2-7)}$$

其中$2L+1$爲片段特徵序列之特徵總數，(式 2-7)所表示的即爲以單項特徵$x[m]$爲中心點，前後各延展L項特徵所得的動態特徵片段。在分段式特徵統計值估測法裡，第m項特徵$x[m]$的統計值，是藉由(式 2-7)所列之片段語句所有特徵中求得。換言之，我們假設$x[m]$對應至一隨機變數$X[m]$，進而假設片段特徵序列$\{x[k];m-L \le k \le m+L\}$爲此隨機變數之樣本，然後根據此樣本，我們可估測出隨機變數$X[m]$的各種統計值，如：

1. $X[m]$的期望值(mean)爲

$$\mu_{X[m],(s)}[m] = \frac{1}{2L+1}\sum_{k=m-L}^{m+L} x[k], \qquad \text{(式 2-8)}$$

2. $X[m]$的變異數(variance)爲

$$\sigma^2_{X[m],(s)}[m] = \frac{1}{2L+1}\sum_{k=m-L}^{m+L}\left(x[k]-\mu_{X[m],(s)}[m]\right)^2, \qquad \text{(式 2-9)}$$

3. $X[m]$的第J階中央動差(central moment)爲

$$\xi^{(J)}_{X[m],(s)}[m] = \frac{1}{2L+1}\sum_{k=m-L}^{m+L}\left(x[k]-\mu_{X[m],(s)}[m]\right)^J, J\text{爲任意之正偶數} \qquad \text{(式 2-10)}$$

4. $X[m]$的動態範圍(dynamic range)爲

$$d_{X[m],(s)}[m] = \max_{m-L\le k\le m+L}\{x[k]\} - \min_{m-L\le k\le m+L}\{x[k]\}, \qquad \text{(式 2-11)}$$

5. $X[m]$的機率分佈函數(probability distribution function)爲

$$F_{X[m],(s)}(z) = \frac{1}{2L+1}\sum_{k=m-L}^{m+L} u(z-x[k]). \qquad \text{(式 2-12)}$$

在以上五個式子中的各種統計值的代號中，我們以下標$_{(s)}$來代表這些統計值是由片段(segment)的特徵估測而得，從以上數式之各種統計值求取法得知，本節所用的估測法不同於上一節的整句式統計值估測法，它所針對某一項特徵$x[m]$所得的各種統計值，事實上與$x[m]$中的序列順序m有關，也就是說在分段式統計值估測法中，我們必須要個別計算整段語句中每一項特徵$x[m]$的統計值，接著將每項特徵$x[m]$的統計值供當下的特徵$x[m]$作正規化處理。換言之，不同項特徵$x[m]$所用的統計值會不一樣。以下，我們將討論分段式特徵統計值估測法運用在語音特徵正規化技術之可能的優缺點。

- **分段式特徵統計值估測法運用在語音特徵正規化技術之優缺點**

分段式語音特徵正規化技術可以彌補整段式技術的缺點，使其可以達到近似即時處理的效果。假如片段(即動態的視窗)長度越短，即時處理的優點越明顯，且因片段長度設爲固定常數，其包含的音素數目相對而言較少，不同片段所包含的音素數目較爲一致，所估測之統計值的準確性受到一片段特徵中的音素數目影響較小，因此降低了相同音素在不同語句之間的變異性。然而其缺點爲，若片段長度不夠長，代表能用以估測的樣本數較少，則估測到的統計值可能會較不精確，導致特徵統計正規化的效果變差，這意味著在這分段式技術中，可能無法同時達成即時處理的效果與大幅正規化的辨識準確性，因此通常必須在即時處理與強健性效能這兩者優點之間作取捨(trade-off)。

(三) 碼簿式特徵統計值估測法

在這一節中，我們簡要介紹如何利用碼簿資訊來估測特徵的各種統計值，而碼簿構成的詳細程序請參照文獻[7,9]。首先，假設某單一語句之某一維特徵序列$\{x[n];1 \le n \le N\}$所構成的一組碼簿，表示爲

$$\{y[r], w_r; 1 \le r \le M\} \quad \text{(式 2-13)}$$

其中w_r爲每一碼字$y[r]$所對應的權重值，而M爲碼字總數。在碼簿式特徵統計值估測法方面，我們利用(式 2-13)所示之整組碼字，去求得每一項特徵$x[m]$所對應之隨機變數$X[m]$其統計值如下：

1. $X[m]$的期望值(mean)爲

$$\mu_{X[m],(c)}[m] = \sum_{r=1}^{M} w_r y[r], \quad \text{(式 2-14)}$$

2. $X[m]$的變異數(variance)爲

$$\sigma^2_{X[m],(c)}[m] = \sum_{r=1}^{M} w_r \left(y[r] - \mu_{X[m],(c)}[m]\right)^2, \quad \text{(式 2-15)}$$

3. $X[m]$的第J階中央動差(central moment)爲

$$\xi^{(J)}_{X[m],(c)}[m] = \sum_{r=1}^{M} w_r \left(y[r] - \mu_{X[m],(c)}[m]\right)^J, J \text{爲任意之正偶數} \quad \text{(式 2-16)}$$

4. $X[m]$的動態範圍(dynamic range)爲

$$d_{X[m],(c)}[m] = \max_{1 \le r \le M} \{y[r]\} - \min_{1 \le r \le M} \{y[r]\}, \quad \text{(式 2-17)}$$

5. $X[m]$的機率分佈函數(probability distribution function)爲

$$F_{X[m],(c)}(z) = \sum_{r=1}^{M} w_r u(z - y[r]). \quad \text{(式 2-18)}$$

從上述各式所示，我們得知某一項特徵$x[m]$所對應的各種統計值，其實與$x[m]$中的序列順序m無關，也就是說在碼簿式統計值估測法方面，我們從一組碼字$\{y[r], w_r; 1 \le r \le M\}$中，只計算一次統計值，就可供整段語句中的每項特徵$x[m]$作正規化處理。換言之，不同項特徵$x[m]$將共用同一組統計值，所以本節估測法類似於整句式統計值估測法，然而主要的差別在於，整句式統計值估測法所求得的統計值是從整段語句之特徵序列求得的；碼簿式統計值估測法所求得的統計值是間接從一組碼字求得的，而不是直接從自身語句之特徵序列求得。以下，我們將討論碼簿式特徵統計值估測法運用在語音特徵正規化技術之優缺點。

- **碼簿式特徵統計值估測法運用在語音特徵正規化技術之優缺點**

碼簿式語音特徵統計估測法不同於前兩小節所提的整段式與分段式統計估測法，是由碼簿來幫助我們估算出代表訓練語音特徵與測試語音特徵的統計值，藉此有效執行各種語音特徵正規化演算法，而且也有近似即時處理的優點。在過去本實驗室的研究中[8]，我們提出了兩種碼簿式特徵統計正規化法，包括了碼簿式倒頻譜平均消去法(C-CMS)與碼簿式倒頻譜平均值與變異數正規化法(C-CMVN)，其發現 C-CMS 與 C-CMVN 的辨識結果都比前一類之整段式或分段式的方法來的好，但將其延伸至其他種類的特徵正規化法（如 HOCMN、 CGN 與 HEQ 等）時，發現其效果並沒有比整段式或分段式方法來的好，因此碼簿的統計值估測法可能不是都適用於每一種特徵正規化技術。此可能肇因於碼簿式估測法的一些缺點，例如，在雜訊語音碼簿求取的過程中，

只利用每句語音前幾個音框作爲純雜訊的代表，這會造成雜訊語音碼簿的估測不精確，而且當雜訊環境爲非穩定性(non-stationary)時，所得到的雜訊語音碼簿可能更不精準，因此其改善辨識率結果就可能較不理想。

三、併合式倒頻譜統計正規化技術

在本章中，我們參照上一章所述的整句式、分段式與碼簿式三種特徵統計估測法，提出兩種新的特徵統計估測法，稱之爲併合式(hybrid-based)統計估測法，第一種併合式統計估測法是整合了語音特徵碼簿與**整句**語音特徵的統計資訊，第二種併合式統計估測法則是整合了語音特徵碼簿與**片段**語音特徵的統計資訊。在以下各節中，我們將介紹如何將它們運用於第二章所提到之五種著名的特徵參數統計正規化技術(CMS, CMVN, HOCMN, CGN與HEQ)中，以期得到更準確的特徵強健化結果。

(一) 併合式倒頻譜平均消去法與併合式倒頻譜平均值與變異數正規化法

在這裡我們將一同介紹併合式倒頻譜平均消去法(hybrid-based CMS)與併合式倒頻譜平均值與變異數正規化法(hybrid-based CMVN)。假設某一維原始之輸入特徵序列爲 $\{x[n];1 \le n \le N\}$，則經過 CMS 處理後的輸出特徵參數表示式如下：

$$\tilde{x}[n] = x[n] - \mu[n], \qquad 1 \le n \le N, \tag{式 3-1}$$

而經過 CMVN 處理後的特徵參數表示式如下：

$$\tilde{x}[n] = \left(x[n] - \mu[n]\right) / \sigma[n], \qquad 1 \le n \le N, \tag{式 3-2}$$

其中 N 爲整段序列之特徵總數，而 $\mu[n]$ 與 $\sigma[n]$ 分別爲特徵 $x[n]$ 的平均值與標準差。

在第一種併合式特徵統計估測法中，$\mu[n]$ 與 $\sigma[n]$ 可由下列兩公式估測而得：

CU-CMS/CU-CMVN：

$$\mu_{(c,u)}[n] = \alpha\mu_{(c)}[n] + (1-\alpha)\,\mu_{(u)}[n], \tag{式 3-3}$$

$$\sigma^2_{(c,u)}[n] = \alpha\left[\sigma^2_{(c)}[n] + \mu^2_{(c)}[n]\right] + (1-\alpha)\left[\sigma^2_{(u)}[n] + \mu^2_{(u)}[n]\right] - \mu^2_{(c,u)}[n], \tag{式 3-4}$$

其中下標"(c)"、"(u)"與"(c,u)"分別代表使用碼簿式、整句式與併合碼簿/整句式統計值估測法，而 $\mu_{(c)}[n]$、$\mu_{(u)}[n]$、$\sigma^2_{(c)}[n]$ 與 $\sigma^2_{(u)}[n]$ 分別定義於前一章的(式 2-14)、(式 2-2)、(式 2-15)與(式 2-3)，α 爲權重值，介於 0 到 1 之間，被用來調整碼簿式統計資訊與整段式統計資訊之間的比例。藉由(式 3-3)與(式 3-4)所估測之平均值與變異數而成的 CMS 與 CMVN，我們分別稱爲併合碼簿/整句式 CMS(hybrid codebook/utterance-based CMS, CU-CMS) 與併合碼簿/整句式 CMVN(hybrid codebook/utterance-based CMVN, CU-CMVN)，由(式 3-3)與(式 3-4)可明顯看出，CU-CMS 與 CU-CMVN 所使用的平均值與變異數是將前一章所述之語音特徵碼簿與整段語音特徵的平均值與變異數作一線性的組合。如果權重值 $\alpha = 1$ 時，CU-CMS 和 CU-CMVN 將分別等同於碼簿式 CMS (C-CMS) 和碼簿式 CMVN (C-CMVN)，另一方面，如果 $\alpha = 0$ 時，CU-CMS 和 CU-CMVN 將分別等同於整句式 CMS (U-CMS)和整句式 CMVN (U-CMVN)。

在第二種併合式特徵統計估測法中，$\mu[n]$ 與 $\sigma[n]$ 可由下列兩公式估測而得：

CS-CMS/CS-CMVN：

$$\mu_{(c,s)}[n] = \alpha\mu_{(c)}[n] + (1-\alpha)\,\mu_{(s)}[n], \tag{式 3-5}$$

$$\sigma^2_{(c,s)}[n] = \alpha\left[\sigma^2_{(c)}[n] + \mu^2_{(c)}[n]\right] + (1-\alpha)\left[\sigma^2_{(s)}[n] + \mu^2_{(s)}[n]\right] - \mu^2_{(c,s)}[n] \tag{式 3-6}$$

其中下標"(c)"、"(s)"與"(c,s)"分別代表使用碼簿式、分段式與併合碼簿/分段式統計值

估測法，而$\mu_{(c)}[n]$、$\mu_{(s)}[n]$、$\sigma^2_{(c)}[n]$與$\sigma^2_{(s)}[n]$分別定義於前一章的(式 2-14)、(式 2-8)、 (式 2-15)與(式 2-9)，α爲權重值，介於 0 到 1 之間，被用來調整碼簿式統計資訊與分段式統計資訊之間的比例。藉由(式 3-5)與(式 3-6)所估測之平均值與變異數而成的 CMS 與 CMVN 法，我們分別稱爲併合碼簿/分段式 CMS(hybrid codebook/segment-based CMS, CS-CMS) 與併合碼簿/分段式 CMVN(hybrid codebook/segment-based CMVN, CS-CMVN)，類似前面所提之 CU-CMS 與 CU-CMVN，從(式 3-5)與(式 3-6)可看出，CS-CMS 與 CS-CMVN 所使用的平均值與變異數是將前一章所述之語音特徵碼簿與片段語音特徵的平均值與變異數作一線性的組合。如果權重$\alpha=1$時，CS-CMS 和 CS-CMVN 將分別等同於碼簿式 CMS (C-CMS)和碼簿式 CMVN (C-CMVN)，然而，若$\alpha=0$時，CS-CMS 和 CS-CMVN 將分別等同於分段式 CMS (S-CMS)和分段式 CMVN (S-CMVN)。

(二) 併合式高階倒頻譜動差正規化法

對一特徵時間序列$\{x[n];1\le n\le N\}$而言，經高階倒頻譜動差正規化法(HOCMN)處理後所得的新特徵時間序列如下式：

$$\tilde{x}[n]=\big(x[n]-\mu[n]\big)/\Big(\xi^{(J)}[n]\Big)^{1/J},\qquad 1\le n\le N \tag{式 3-7}$$

其中N爲整段序列之特徵數，$\mu[n]$與$\xi^{(J)}[n]$分爲特徵$x[n]$的平均值與第J階中央動差。

類似上一節所述，這裡我們有兩種方式來估測$\mu[n]$與$\xi^{(J)}[n]$，所對應的 HOCMN 法我們分別稱爲併合碼簿/整句式 HOCMN(CU-HOCMN) 與併合碼簿/分段式 HOCMN(CS-HOCMN)，它們在$\mu[n]$與$\xi^{(J)}[n]$的估測運算如下列數式：

CU-HOCMN (hybrid codebook/utterance-based HOCMN)：

$$\mu_{(c,u)}[n]=\alpha\mu_{(c)}[n]+(1-\alpha)\,\mu_{(u)}[n], \tag{式 3-8}$$

$$\xi^{(J)}_{(c,u)}[n]=\alpha\left(\sum_{r=1}^{M}w_r\Big(y[r]-\mu_{(c,u)}[n]\Big)^J\right)+(1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N}\Big(x[n]-\mu_{(c,u)}[n]\Big)^J\right). \tag{式 3-9}$$

CS-HOCMN (hybrid codebook/segment-based HOCMN)：

$$\mu_{i,(c,s)}[n]=\alpha\mu_{i,(c)}[n]+(1-\alpha)\,\mu_{i,(s)}[n], \tag{式 3-10}$$

$$\xi^{(J)}_{(c,s)}[n]=\alpha\left(\sum_{r=1}^{M}w_r\Big(y[r]-\mu_{(c,s)}[n]\Big)^J\right)+(1-\alpha)\left(\frac{1}{2L+1}\sum_{k=n-L}^{n+L}\Big(x[k]-\mu_{(c,s)}[n]\Big)^J\right). \tag{式 3-11}$$

其中$\mu_{(c)}[n]$、$\mu_{(u)}[n]$、$\mu_{(s)}[n]$分別定義於前一章的(式 2-14)、(式 2-2)與(式 2-8)，α爲權重值，介於 0 到 1 之間，用來調整碼簿統計資訊與整段或片段特徵統計資訊之間的比例。

(三) 併合式倒頻譜增益正規化法

對一特徵時間序列$\{x[n];1\le n\le N\}$而言，經倒頻譜增益正規法(CGN)處理後所得的新特徵時間序列如下式：

$$\tilde{x}[n]=\big(x[n]-\mu[n]\big)/\,d[n],\qquad 1\le n\le N, \tag{式 3-12}$$

其中N爲整段序列之特徵總數，$\mu[n]$與$d[n]$分別爲特徵$x[n]$的平均值與動態範圍。

類似上兩節的方法，這裡我們有兩種方式來估測$\mu[n]$與$d[n]$，所對應的 CGN 法我們分別稱爲併合碼簿/整句式 CGN(CU-CGN) 與併合碼簿/分段式 CGN(CS-CGN)，它們對$\mu[n]$與$d[n]$的估測運算分別如下數式：

CU-CGN(hybrid codebook/utterance-based CGN)：

$$\mu_{(c,u)}[n] = \alpha\mu_{(c)}[n] + (1-\alpha)\,\mu_{(u)}[n], \qquad \text{(式 3-13)}$$

$$d_{(c,u)}[n] = \max\left\{Y_{(c)} \cup X_{(u)}\right\} - \min\left\{Y_{(c)} \cup X_{(u)}\right\} \qquad \text{(式 3-14)}$$

CS-CGN(hybrid codebook/segment-based CGN)：

$$\mu_{(c,s)}[n] = \alpha\mu_{(c)}[n] + (1-\alpha)\,\mu_{(s)}[n], \qquad \text{(式 3-15)}$$

$$d_{(c,s)}[n] = \max\left\{Y_{(c)} \cup X_{(s)}\right\} - \min\left\{Y_{(c)} \cup X_{(s)}\right\} \qquad \text{(式 3-16)}$$

其中$Y_{(c)}$、$X_{(u)}$與$X_{(s)}$代表了(式 2-13)、(式 2-1)與(式 2-7)。$\mu_{(c)}[n]$、$\mu_{(u)}[n]$與$\mu_{(s)}[n]$分別定義於前一章的(式 2-14)、(式 2-2)與(式 2-8)，其中α爲一個介於 0 到 1 之間的權重值，代表了碼簿式統計資訊與整句式或分段式統計資訊這兩者之間所使用的比例，$\max(\cdot)$與$\min(\cdot)$分別爲取最大值與最小值的函數，而『∪』爲聯集符號，意指將碼簿與整句(或片段)語句的特徵串在一起。

(四) 併合式倒頻譜統計圖等化法

對一特徵時間序列$\{x[n];1 \le n \le N\}$而言，經倒頻譜統計圖等化法(HEQ)處理後所得的新特徵時間序列如下式：

$$\tilde{x}[n] = F_{ref}^{-1}\left(F_X(x[n])\right), \qquad 1 \le n \le N, \qquad \text{(式 3-17)}$$

其中N爲整段序列之特徵總數，$F_{ref}(\cdot)$爲預先定義的參考機率分布函數，而$F_X(\cdot)$則爲特徵$x[n]$的機率分布函數。

類似前面幾節所述，這裡我們有兩種方式來估測機率分布函數$F_X(\cdot)$，分別使用在 HEQ 上，因此我們分別稱爲併合碼簿/整句式 HEQ(CU-HEQ)與併合碼簿/分段式 HEQ(CS-HEQ)，它們對$F_X(\cdot)$的估測表示式如下所示：

CU-HEQ(hybrid codebook/utterance-based HEQ)：

$$F_{X,(c,u)}(z) = \alpha F_{X,(c)}(z) + (1-\alpha)\,F_{X,(u)}(z), \qquad \text{(式 3-18)}$$

CS-HEQ(hybrid codebook/segment-based HEQ)：

$$F_{X,(c,s)}(z) = \alpha F_{X,(c)}(z) + (1-\alpha)\,F_{X,(s)}(z), \qquad \text{(式 3-19)}$$

其中α爲一個介於 0 到 1 之間的權重值，代表了碼簿式統計資訊與整段式或分段式統計資訊這兩者之間所使用的比例，而$F_{X,(c)}(\cdot)$、$F_{X,(u)}(\cdot)$與$F_{X,(s)}(\cdot)$分別定義於前一章的(式 2-18)、(式 2-6) 與(式 2-12)。

四、實驗環境設定與各種強健性語音特徵正規化技術之實驗結果與討論

(一) 實驗環境設定

本論文採用歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行的 AURORA2 語音資料庫[10]，其內容是由連續的英文數字字串所構成。此語音資料庫有兩種不同的訓練環境：乾淨環境(clean-condition)與多重環境(multi-condition)以及三種不同的測試集合：A 組(地下鐵、人聲、汽車和展覽館雜訊)、B 組(餐廳、街道、機場和火車站雜訊)與 C 組(地下鐵、街道雜訊外加 MIRS 通道效應)雜訊語音集合。乾淨環境代表沒有任何雜訊的語音環境，而多重環境則代表適當加入各種附加雜訊的語音環境。本論文的實驗只採用乾淨環境的語音特徵作聲學模型的訓練，並對三組雜訊語音集合加以辨識。

在這裡，基礎實驗(baseline experiment)將採用未處理的梅爾倒頻譜特徵係數(MFCC)作爲訓練跟測試，所使用的MFCC特徵參數爲13維(c_0~c_{12})，再加上其一階和二階差量，總共有39維特徵參數作爲最終使用之特徵參數向量。

聲學模型爲由左向右(left-to-right)之隱藏式馬可夫模型(hidden Markov model, HMM)

的形式，是使用隱藏式馬可夫模型訓練軟體 HTK[11]訓練所得，其中包含 11 個數字模型(zero, one, two, …, nine 及 oh)以及靜音(silence)模型，每個數字模型包含 16 個狀態，而每個狀態則包含 20 個高斯密度混合。

(二) 各種強健性語音特徵正規化技術之實驗結果與討論

本章將介紹我們所提的五種強健性語音特徵正規化技術之辨識實驗結果(20dB、15dB、10dB、5dB 與 0dB 五種訊雜比下的辨識率平均)，分別爲 CMS、CMVN、HOCMN、CGN 以及 HEQ，而這些方法的特徵統計相關資訊，分別使用三種特徵統計值估測法(整句式、分段式與碼簿式)以及兩種併合式(hybrid-based)特徵統計值估測法(碼簿/整句式與碼簿/分段式)來求得，然後進一步比較、討論與分析。

在本論文中，所有片段式統計估測法實驗中的片段長度 $2L+1$ 都設爲 101(除了 HOCMN 之外，它的片段長度 $2L+1$ 設爲 87)；所有碼簿式與併合式實驗中的碼字數目 R 統一設定爲 16 或 256，而併合式實驗中的 α，我們固定設爲 0.5，使併合之雙方統計資訊所佔的比例相等。

1、各種倒頻譜平均消去法之實驗結果

從表一中，我們得知各種 CMS 的總平均辨識率情形。我們先探討三種傳統的特徵統計值估測法作用在 CMS 上的效果，發現 U-CMS、S-CMS 與 C-CMS 中，以 C-CMS (R=256)所得的整體平均辨識率最大，比基礎實驗結果提升了 10.75%，而相對錯誤率衰減率(RR)爲 37.83%。

接著，我們可看出本論文新提出的兩種併合式特徵統計值估測法運用在 CMS 上之效果，如以下幾點所述：

(1) CU-CMS 與 CS-CMS 皆明顯優於傳統之 U-CMS、S-CMS 與 C-CMS。

(2) CS-CMS(R=16)表現優於 CU-CMS(R=16)，相對於基礎實驗結果，在辨識率上提升了 14.61%，而相對錯誤率衰減率高達 51.41%。

(3)在各種不同型態的 CMS 中，以 CS-CMS (R=16)的總平均辨識結果最佳，而其它 CMS 的總平均辨識結果的優劣順序，依序爲：CS-CMS (R=256)、CU-CMS (R=16)、CU-CMS (R=256)、C-CMS (R=256)、C-CMS(R=16)、U-CMS 以及 S-CMS。因此，我們驗證了所新提出之兩種併合式 CMS 在提昇語音特徵強健性上，比 U-CMS、S-CMS 與 C-CMS 還要來的優越。

method	Set A	Set B	Set C	Average	RR
baseline	71.92	68.22	77.61	71.58	—
C-CMS (R=16)	80.83	79.29	86.13	81.27	34.10
C-CMS (R=256)	81.62	81.58	85.25	82.33	37.83
U-CMS	79.35	82.46	79.91	80.71	32.13
CU-CMS (R=16)	83.28	84.92	84.28	**84.14**	44.19
CU-CMS (R=256)	82.13	84.29	83.06	83.18	40.82
S-CMS	77.28	80.66	77.63	78.70	25.05
CS-CMS (R=16)	85.38	87.43	85.31	**86.19**	51.41
CS-CMS (R=256)	84.63	86.98	84.42	85.53	49.09

表一、各種 CMS 的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

2、各種倒頻譜平均值與變異數正規化法之實驗結果

表二呈現了各種倒頻譜平均值與變異數正規化法(CMVN)的辨識率，由表二中，首先，我們可以看出三種傳統的特徵統計值估測法運用於 CMVN 時，其辨識結果相較於基礎實驗而言，在 U-CMVN、S-CMVN 與 C-CMVN 中，以 C-CMVN (R=256)所得的總平均辨識率最高，比基礎實驗結果提升了 15.14%，而相對錯誤率降低率達到了 53.27%。

接下來，本論文提出兩種併合式特徵統計值估測法運用在 CMVN 上，同樣其辨識結果相較於基礎實驗也有明顯的進步。由表二所示，在 CU-CMVN 與 CS-CMVN 中，以 CS-CMVN (R=16)的總平均辨識率最高，比基礎實驗結果提升了 17.38%，而相對錯誤率降低率高達 61.15%。因此兩種併合式 CMVN 在語音特徵強健性方面，跟前一節的併合式 CMS 一樣，皆優於 U-CMVN、S-CMVN 與 C-CMVN。

Method	Set A	Set B	Set C	Average	RR
Baseline	71.92	68.22	77.61	71.58	—
C-CMVN (R=16)	85.75	85.5	83.78	85.26	48.14
C-CMVN (R=256)	87.16	87.44	84.39	86.72	53.27
U-CMVN	85.03	85.56	85.61	85.36	48.49
CU-CMVN (R=16)	87.87	88.67	86.14	**87.84**	57.21
CU-CMVN (R=256)	87.25	88.06	85.78	87.28	55.24
S-CMVN	83.99	84.85	84.78	84.49	45.43
CS-CMVN (R=16)	88.98	89.82	87.19	**88.96**	61.15
CS-CMVN (R=256)	88.18	89.09	86.73	88.25	58.66

表二、各種 CMVN 的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

3、各種高階倒頻譜動差正規化法之實驗結果

表三呈現了各種高階倒頻譜動差正規化法(HOCMN)的辨識率，在這裡，中央動差的階數 J 皆設爲 100。從表三可看出以下幾點現象：

(1) 在傳統之 U-HOCMN、S-HOCMN 與 C-HOCMN 中，以 U-HOCMN 所得的總平均辨識率最大，與基礎實驗相比，提升了 16.31%，而相對錯誤降低率達到了 57.39%。此現象跟前兩節所呈現的結果並不相同，因爲碼簿式的 HOCMN(C-HOCMN)表現並不理想，其可能原因爲，碼簿對於較低階的動差值（例如平均值與變異數）之估測較爲準確，但無法有效估測較高階的動差值。

(2) 本論文提出兩種併合式特徵統計值估測法，當其運用在 HOCMN 上時，其辨識結果相較於基礎實驗結果仍有明顯改善：CU-HOCMN(R=16)與 CS-HOCMN(R=16)分別比基礎實驗結果提升了 16.09%與 17.78%，相對錯誤降低率衰減分別高達 56.62%與 62.56%。

(3) 在各種型態的 HOCMN 中，以 CS-HOCMN (R=16)的總平均辨識率最佳，其它併合式 HOCMN 皆低於 U-HOCMN，不同於前面幾節所述的併合式 CMS 與 CMVN 所呈現的結果。更進一步觀察，可看出 CU-HOCMN 在 Set C 的辨識率相對較低，造成總平均辨識率不及 U-HOCMN，此現象可歸因於 Set C 中的語音包含了摺積性雜訊(convolutional noise)，而碼簿(codebook)中只考慮到加成性雜訊(additive noise)，所以

造成 Set C 之辨識結果不盡理想。

Method	Set A	Set B	Set C	Average	RR
Baseline	71.92	68.22	77.61	71.58	—
C-HOCMN (R=16)	84.86	83.40	85.68	84.44	45.25
C-HOCMN (R=256)	86.30	86.34	83.53	85.76	49.89
U-HOCMN	87.43	88.54	87.52	87.89	57.39
CU-HOCMN (R=16)	87.55	88.85	85.57	**87.67***	56.62
CU-HOCMN (R=256)	86.66	88.15	85.12	86.95	54.08
S-HOCMN	85.60	86.63	86.16	86.12	51.16
CS-HOCMN (R=16)	89.17	90.26	87.96	**89.36**	62.56
CS-HOCMN (R=256)	87.44	88.78	86.22	87.73	56.83

表三、各種HOCMN的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

4、各種倒頻譜增益正規化法之實驗結果

表四列出了各種倒頻譜增益正規化法(CGN)的辨識率，若與表三相比較，我們發現它們的結果十分類似，三種傳統估測法所對應的U-CGN、S-CGN與C-CGN中，以U-CGN所得的總平均辨識率最高，與基礎實驗結果相比，提升了 16.33%，而相對錯誤降低率爲 57.46%，相對而言，C-CGN 表現較差，此可能原因跟上一節所述類似，即碼簿可能無法較精確地估測 CGN 所需用到的動態範圍值，以及其未考慮到 Set C 的摺積性雜訊干擾。

然而，當兩種併合式特徵統計值估測法，分別使用在 CGN 時，其辨識結果都能有十分顯著的提昇，其中以 CS-CGN(R=16)的總平均辨識率最高，比基礎實驗提升了 17.88%，而相對錯誤降低率高達了 62.91%，我們證實了在併合式 CGN 中，除了 CU-CGN (R=256)略低於 U-CGN 之外，皆優於 U-CGN、S-CGN 與 C-CGN。

method	Set A	Set B	Set C	Average	RR
Baseline	71.92	68.22	77.61	71.58	—
C-CGN (R=16)	85.60	84.29	84.82	84.92	46.94
C-CGN (R=256)	86.65	87.07	84.41	86.37	52.04
U-CGN	87.62	88.49	87.32	87.91	57.46
CU-CGN (R=16)	88.11	89.14	86.07	**88.11**	58.16
CU-CGN (R=256)	86.97	88.46	85.30	87.23	55.07
S-CGN	86.36	87.31	86.99	86.86	53.76
CS-CGN (R=16)	89.31	90.40	87.89	**89.46**	62.91
CS-CGN (R=256)	88.42	89.73	86.92	88.64	60.03

表四、各種 CGN 的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

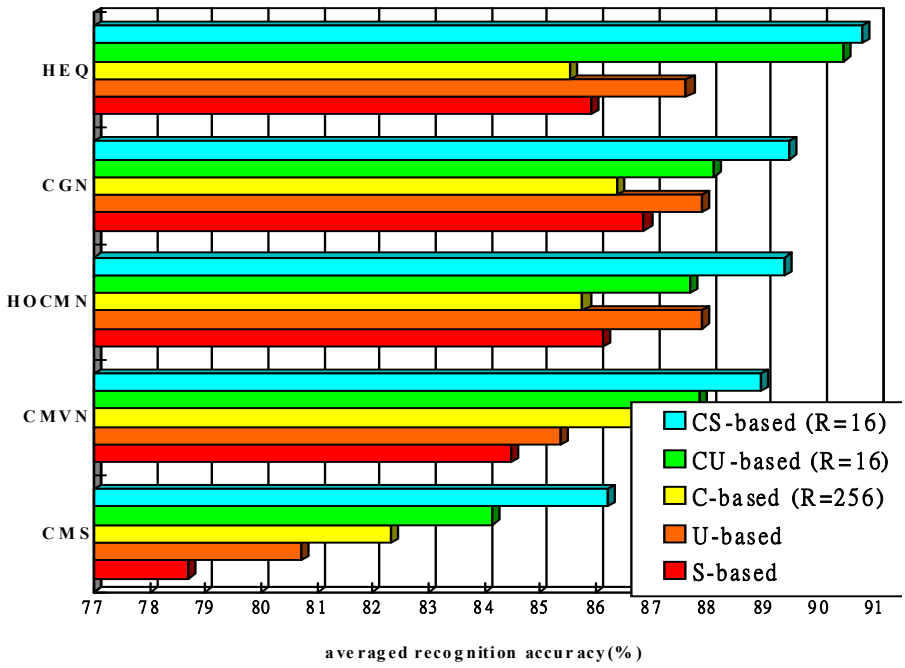
5、各種統計圖等化法之實驗結果

表五列出了各種不同型態的統計圖等化法(HEQ)之辨識率，其結果與表三與表四類似，我們明顯看出，雖然碼簿式的 HEQ(C-HEQ)效果不盡理想，然而當我們把碼簿與整句特徵合併，或把碼簿與片段特徵合併，所分別對應的 CU-HEQ 與 CS-HEQ，其帶來的辨識率提昇程度皆十分顯著，明顯超越了整句式 HEQ(U-HEQ)與片段式 HEQ(S-HEQ)，而其中又以 CS-HEQ 表現最佳。平均辨識率高達 90.76%，相對錯誤降低率爲 67.49%。

method	Set A	Set B	Set C	Average	RR
Baseline	71.92	68.22	77.61	71.58	—
C-HEQ (R=16)	80.15	80.41	76.03	79.43	27.62
C-HEQ (R=256)	86.23	85.77	83.71	85.54	49.12
U-HEQ	86.95	88.39	87.40	87.62	56.44
CU-HEQ (R=16)	90.21	91.16	89.37	**90.42**	66.29
CU-HEQ (R=256)	88.76	89.68	87.85	88.95	61.12
S-HEQ	85.10	86.82	85.64	85.90	50.39
CS-HEQ (R=16)	90.57	91.54	89.57	**90.76**	67.49
CS-HEQ (R=256)	89.11	90.21	88.35	89.40	62.70

表五、各種 HEQ 的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

(三) 綜合討論



圖一、各種語音特徵正規化法在不同的特徵統計估測法下的總平均辨識率之比較

圖一是本章所有方法之總平均辨識率比較圖，從此圖與前面的五個表中，我們大致觀察到以下幾種情形：

(1) 整句式方法皆比分段式方法還要好，其可能原因爲，分段式方法是把整段語句分割成許多相鄰的片段語句，雖然可達到近似即時處理的效果，但也因片段資料量的不足導致統計估測值較不準確，因此造成辨識率的下降。在碼簿式方法方面，除了 C-CMS 與 C-CMVN 之外，效果皆低於整句式與分段式方法，其可能原因爲，虛擬雙通道中的雜訊估測只以整段語句的前幾個音框當雜訊的代表，所得之雜訊特性可能較不精準，造成建立的的碼簿比較不精確。

(2) 併合式方法幾乎皆比整句式、分段式與碼簿式方法來的好，例如：CU-HEQ (90.42%)與 CS-HEQ (90.76%)優於 U-HEQ (87.62%)、S-HEQ (85.9%)與 C-HEQ (85.54%)。這結果證實了整句式、分段式與碼簿式方法，獨自所提升的辨識率較小，但結合了碼簿式與整句式(或分段式)的併合式方法，將使辨識率大幅地提升。

(3) 在併合式方法中，我們將 α 都設爲 0.5，這表示著碼簿與整段語句的統計資訊之使用比例相等，而沒有任何偏差。雖然 $\alpha = 0.5$ 未必是最佳的設定參數，但至少代表了我們無須精微地挑選此參數值，便能得到明顯的併合效益。

(4) 在兩種併合式方法中，碼字數目 *R*=16 所對應的辨識結果明顯比 *R*=256 所對應的辨識結果有明顯的改善，此現象在 CMS、CMVN、HOCMN、CGN 與 HEQ 皆是如此，此可能原因在於碼簿資訊與整句語音資訊(或片段語音資訊)之間的不一致性。由於碼簿只呈現純語音部分的資訊，而整句或片段特徵可能同時包含了語音與非語音部分的資訊，因此增加碼字數目，使碼簿所對應的語音專屬資訊越多且越詳細，這將會使雙方的不一致性越來越明顯。然而此結果卻反而成爲我們所提出之併合式方法的優點，因爲這代表了我們在各種併合式倒頻譜統計正規化法中，只要使用較小的碼字數目，就可得到較佳的辨識結果，而且大幅降低演算法本身的運算複雜度。

(5) 在併合式方法中，以碼簿/分段式方法的辨識效能最高，它補足了分段式方法之辨識率較差的缺點，而仍保有分段式方法之近似及時處理的優點，因此極具實用價值。

五、結論

在本論文中，我們提出了兩種併合式的特徵統計估測法，它們是將語音特徵之碼簿與整句或分段的語音特徵適當地結合，進而估測出特徵的各項統計值，我們將此新方法分別使用在五種強健性語音特徵統計正規化技術上：倒頻譜平均消去法(CMS)、倒頻譜平均值與變異數正規化法(CMVN)、高階倒頻譜動差正規化法(HOCMN)、倒頻譜增益正規化法(CGN)以及倒頻譜統計圖等化法(HEQ)，我們發現，跟整句式、分段式與碼簿式統計估測法相較之下，這兩種併合式的估測法皆能更明顯地提昇語音特徵正規化技術的效能，更有效地改善雜訊環境下的語音辨識率，因此可推論，我們所提出的新方法得以得到更精確的語音特徵的統計特性。

在未來的相關研究上，我們有以下幾個方向發展：

(1) 我們希望利用併合式特徵統計估測法，運用在其他的倒頻譜統計正規化技術上，例如：倒頻譜形狀正規化法(CSN)或其他階層的 HOCMN 等技術，以觀察其效能。

(2) 在實驗結果中，我們發現了併合式的倒頻譜統計正規化技術在摺積性雜訊環境下的辨識改善程度，較差於加成性雜訊環境，因此我們期望能結合消除通道效應的方法，例如相對頻譜法(RASTA)等，使其提升辨識結果。
(3) 除了本論文所用的數字語音資料庫外，我們將嘗試把所提的新方法運用在其它較大字彙的語音資料庫上，進一步驗證這些方法的實用價值。

參考文獻

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Volume 29, Issue 2, pp. 254-272, 1981

[2] C.-P. Chen, K. Filaliy and J. A. Bilmes, "Frontend post-processing and backend model enhancement on the Aurora 2.0/3.0 databases", *in Proc. International Conference on Spoken Language Processing* (*ICSLP*), pp. 241-244, 2002

[3] R. Haeb-Umbach, "Investigations on inter-speaker variability in the feature space", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 397-400, 1999

[4] C.-W. Hsu and L.-S. Lee, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 197-200, 2004

[5] S. Yoshizawa, N. Hayasaka, W. Naoya, and Y. Miyanaga. "Cepstral gain normalization for noise robust speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 209-212, 2004

[6] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust speech recognition", *European Conference on Speech Communication and Technology* (*Eurospeech*), pp. 1135-1138, 2001

[7] J.-W. Hung, "Cepstral statistics compensation and normalization using online pseudo stereo codebooks for robust speech recognition in additive noise environments", *IEICE Trans. Information and Systems*, pp. 296-311, 2008

[8] T.-H. Hsieh, “Feature statistics compensation for robust speech recognition in additive noise environments”, *M.S. thesis*, National Chi Nan University, Taiwan, 2007

[9] K.-C. Wu, “Study of cepstral statistics normalization techniques for robust speech recognition in additive noise environments”, *M.S. thesis*, National Chi Nan University, Taiwan, 2008

[10] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *in Proc. of ISCA IIWR ASR2000*, Paris, France, 2000

[11] http://htk.eng.cam.ac.uk/