# 結合韻律與聲學訊息之強健性漢語語者驗證系統

張文杰[2]，陳鼎允[1]，陳子和[2]，曾志仁[1]、廖元甫[1]，莊堯棠[2]

[1]國立台北科技大學電子工程學系

[2]國立中央大學電機工程學系

Email: yfliao@ntut.edu.tw

**摘要**

在本論文中，我們探討強健式漢語文字特定(text-dependent, TD)與文字不特定(text-independent, TI)語者驗證系統，主要是針對漢語的聲調語言特性，提出潛在韻律分析(latent prosody analysis, LPA)及高斯混合模型(Gaussian mixture model, GMM)兩種方式，分別用來建置每位語者的韻律行為模型及能量與音高軌跡(pitch contour)的動態變化模型。實驗結果顯示在使用 ISCSLP-SRE 語料之漢語文字特定與文字不特定語者驗證實驗情況下，使用韻律訊息(prosodic information)來輔助傳統使用頻譜特徵(spectral features)之語者驗證系統，可有效提升系統效能。

## 1. 序論

語者驗證在現今的語音處理中為重要的分支研究項目之一 [1] ，目前有相當多的研究不斷地持續發展中。尤其從 1996 年開始，NIST 機構每年都會藉由舉辦語者辨認評估(speaker recognition evaluation, SRE)來提供一個共同的測試平台 [2] ，以促進語者辨認技術演進及各種演算方法的實用性，更讓全世界最新穎的想法得以在競賽裡獲得驗證。相較於外國語言，漢語的語者辨認競賽還在起步階段，在 2006 年舉辦的中文口語語言處理國際會議(ISCSLP)中，首度建立了漢語語言的語者競賽機制 [3] ，讓此領域的研究人員能夠同時在擁有一樣的資源下，透過中文語言資源聯盟(Chinese Corpus Consortium, CCC) [4] 所提供的資料庫，切磋漢語語者的辨認技術與研究。

語者驗證技術在現實生活中可以有許多的應用，例如可以藉由電話連接到銀行或是信用卡等客服中心，並直接透過使用者的聲音來驗證身份以即時提供便利的私人服務。然而使用者若任意使用不同的電話話筒或通道，則會有電話話筒與通道環境不匹配問題，而導致傳統以頻譜特徵為主之語者驗證系統效能降低。為了改善電話話筒與通道不匹配問題，近年來有許多人利用韻律訊息來強化傳統以頻譜特徵為基礎之語者驗證系統 [5-8] 的效能，韻律特徵(prosodic feature)不僅含有語者訊息並已被認定是不易受到電話話筒與通道不匹配的影響，而且在西方語言的研究中亦有很多的文獻證實其效果。因此在本論文中我們將著重在討論如何利用韻律特徵來強化漢語語者驗證系統的效能，主要是考慮到漢語屬於一種聲調(tonal)語言，其本質上依賴聲調的不同來區別出同音異字詞，故韻律特徵對漢語的影響應較西方語言強烈。

一般來說頻譜特徵代表是較短程(short term)且低階層的聲學訊息，都是和發音器官相關的實體線索，其中被廣泛使用的梅爾頻率倒頻譜係數(Mel-frequency cepstral coefficients, MFCCs)是可以擷取並傳達出發音腔道(vocal tract)的分佈；韻律特徵則通常作為聲門資訊(glottic source)的特徵參數，不僅是較長程(long term)且高階的特徵並含有語者本身特殊的訊息，如音高軌跡及音調(intonation)等，因此兩者各是呈現語音訊號中不同的訊息。在韻律訊息改善不匹配問題的方法

中，對於短程韻律方面通常會使用高斯混合模型來統計韻律訊息，能捕捉到如音高與能量的分佈、音高與能量的斜率以及音高與能量的持續時間等韻律特徵，而長程韻律模型則通常有 N-gram 及 discrete hidden Markov model(DHMM) [6] 兩種方法，可以表現出韻律訊息隨時間的長程變化。不過長程韻律模型通常受限於大量語料的需求問題，因為要有充分語料才能有效描述韻律的特性，所以針對這點缺失我們將提出潛在韻律分析方法來得到可靠的韻律訊息。

本文章中，我們會在系統前端的頻譜特徵使用 mean subtraction, variance normalization, and ARMA filtering (MVA) [9] 去除部份通道不匹配的問題，接著語者驗證系統將運用不同模組來整合頻譜與韻律訊息。文字不特定條件下，有三種模組用作語者確認系統的建構，包括目前被視為標準作法的 a maximum a posteriori (MAP)-adapted GMM (MAP-GMM) [10] 、音高與能量之高斯混合模型，以及潛在韻律分析模組。而文字特定則有另外三種模組來構成，包括文字限定的語者高斯混合模型，隱藏式馬可夫模型(hidden Markov model, HMM)以及音高與能量之高斯混合模型。而後端改良型的測試分數正規化(test normalization, T-norm) [11] 則可以對分數作調整。最後我們利用 MIT 林肯實驗室所發展的 LNKnet [12] 軟體做不同模組分數上的結合。

在 MVA 對頻譜特徵的處理主要是將特徵向量作一種正規化，雖然近年來有很多特徵正規化的方法，如 feature warping [13] 及 histogram equalization(HEQ) [14]都能有很好效果，但是 MVA 的良好表現與簡單使用是我們在此優先考量的原因。而文字不特定語者驗證中的 MAP-GMM 是透過通用背景模型(universal background model, UBM)調適出語者個別的高斯混合模型，使每個語者模型所含蓋的聲學特性更具完整性，如此對於文字內容的變異性就能廣泛接納。而韻律特徵由兩方面著手，短程韻律用高斯混合模型對能量與音高軌跡建置其動態變化模型，長程則用所提出的潛在韻律分析更有效地得知韻律行為，其主要是將語者驗證問題轉換為類似文件檢索（document retrieval）的問題，統計出韻律序列的組合並建立韻律空間(prosody space)，再透過 probabilistic latent semantic analysis（PLSA）[15-16] 的空間維度簡化後來呈現語者的韻律模型。

文字特定語者驗證任務對使用者的說話內容是有其限制，所以對語音事件之聲學變化有詳細考慮的隱藏式馬可夫模型是必需的，這樣才能善用系統對使用者先天的限制條件。當然高斯混合模型在頻譜上對語者特性的描述仍是不可或缺的角色，因為用高斯密度函數表示語者的聲學類別仍可反應出語者特性分佈，與隱藏式馬可夫模型分屬不同角度的分析。而在韻律訊息方面，考慮到語料長度的缺乏，僅對短程韻律方面使用高斯混合模型來描述音高與能量軌跡的動態變化。至於系統後端我們也考慮到分數的變化性，來自語者之間說話內容或是長度的不同都會造成影響，且訓練和測試環境的不匹配更是一大主因，所以使用改良式測試分數正規化(modified test normalization, MT-norm)來調整目標語者(target speaker)模型的分數，拉開目標語者與冒充語者(impostor)之間的分佈，進而改善正確率並更簡易產生驗證所用的門檻值。

由於頻譜特徵與韻律特徵是呈現訊號中不同的訊息，所以考慮其之間可能的互補特性，則文字不特定與文字特定語者驗證的不同模組必須整合，而我們是透過多層感知機(multi-layer perceptrons, MLPs)與 development 的測試語料來決定驗證系統的合併方式，在系統求得的分數上作非線性組合，以達到利用韻律特徵來強化漢語語者驗證系統之目的。

本文內容安排如下：第二章節描述在漢語語言裡所使用的各種方法，並討論韻律特徵在系統中的輔助作用；第三節則詳細說明潛在韻律分析的方法；第四節是系統運用在 ISCSLP2006-SRE 的實驗結果；最後則為結論。

## 2. 文字特定與文字不特定之語者驗證系統架構

圖一及圖二所示分別為文字不特定與文字特定驗證之整合架構，對於文字不特定來說，將有音框(frame)和語者兩種層級一起使用，主要是因為考量到註冊與測試語料數量的關係，在語者層級所需的量遠比音框層級大得多，況且現實狀況中總是只能獲得有限的語料量；反觀文字特定的情況則只是採取音框層級的方式，因為該語料的長度都非常的簡短。

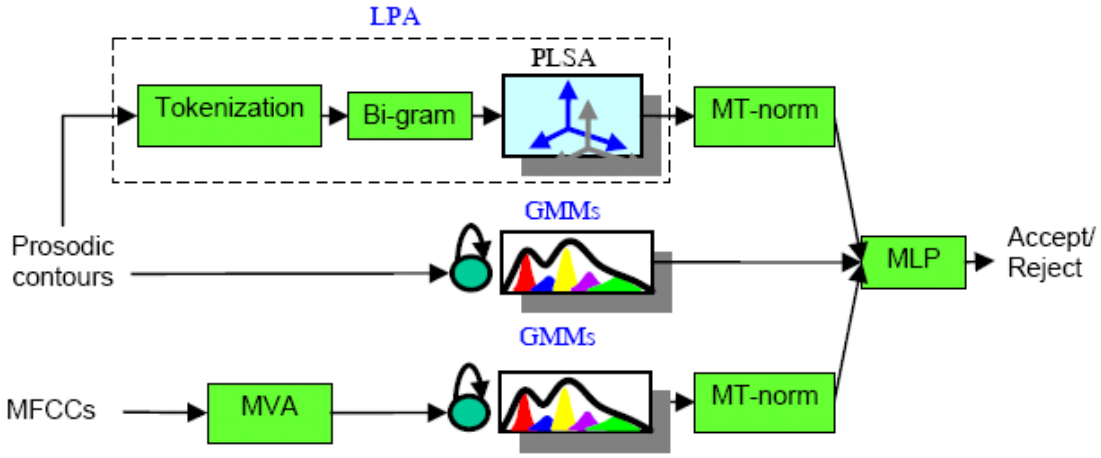
文字不特定任務是由三種不同模組來構成，如圖一所示，首先是高斯混合模型將音高與能量軌跡的動態變化與所提出之潛在韻律分析方法做一合併，完整獲取每位語者的韻律行為，最後則是以 MAP-GMM 完成系統在頻譜特徵的主體。利用 MAP-GMM 取代原本的文字限定語者高斯混合模型，原因在於實際應用情況中不可能要求使用者在註冊時錄製大量的語音，以致於每一個人的訓練語料可能有一些聲學特性沒被涵蓋到，在測試時可能會造成系統效能下降，並且文字不特定的確認是無法限制測試語者說話的內容，所以建立出來的語者模型不僅要能代表該註冊語者的特性，還要能夠涵括在不同聲學情況下的語者變異性。

為了克服電話話筒與通道不匹配的影響，韻律訊息的使用仍是我們主要考量，雖然使用高斯混合模型可以用來統計韻律訊息，但一般只能補捉到音高與能量變化等短程的韻律訊息，其所得到的改善幅度仍然有限，而對於補捉較長程韻律訊息變化的方法通常有 DHMM 和 N-gram 兩種，可是都需使用大量的訓練與測試語料，對此我們提出潛在韻律分析的方法是能在有限的語料情況下得到可靠的韻律訊息。

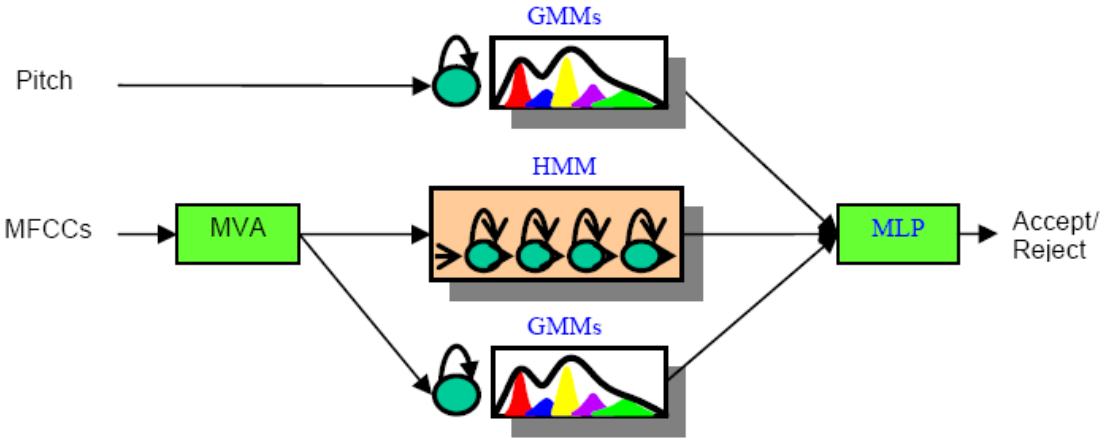
在文字特定條件下，有另外三種模組用作語者驗證系統的建構，如圖二所示，包括獲知音高與能量軌跡動態變化的高斯混合模型、模型化梅爾頻率倒頻譜係數暫態軌跡所用的隱藏式馬可夫模型以及統計梅爾頻率倒頻譜係數分佈的高斯混合模型。一般來說，語者驗證系統都會採用梅爾頻率倒頻譜係數與高斯混合模型的搭配，其中梅爾頻率倒頻譜係數已經將語音的頻譜特徵做了良好描述，然後透過由許多高斯密度函數組成的高斯混合模型來表示語者特性的分佈，而這裡並不如文字不特定中使用 MAP-GMM 來建立語者特定模型，因為藉助通用背景模型補強的聲學特性反而會對文字特定產生困擾，造成語者模型無法針對文字特定任務進行驗證。

另外圖二的文字特定語者驗證中，我們可知測試語者說話內容是有限制性的，它必須符合宣稱語者在系統中註冊語料的語句內容，除此內容外的語句都將一律拒絕，即便是真實語者說出不同樣的內容也是無法接受的，利用這種系統使用上的限制條件，隱藏式馬可夫模型會是更適合用來建立模型的方法，因為隱藏式馬可夫模型對梅爾頻率倒頻譜係數之暫態軌跡可以有詳細的描述，而高斯混合模型中並未考慮到語音事件的聲學變化。

以頻譜特徵為主的系統來說，隱藏式馬可夫模型與高斯混合模型的結合已經可以獲得還不錯的結果，然而在訓練與測試環境不匹配的狀況下，仍需加入不同觀點的韻律訊息來強健系統，因此我們考慮音高與能量軌跡的動態變化，利用有聲音(voiced)的區段中取出每一個音框的對數(log)音高及對數能量，並估計對數音高及能量的一階微分來建立高斯混合模型，而由於韻律特徵是比較不受話筒或通道的影響，所以可以補強原頻譜系統的缺失。




圖一、文字不特定語者驗證方法之方塊圖。




圖二、文字特定語者驗證方法之方塊圖。

最後值得一提的是系統不論特定或不特定的任務，對取自於梅爾頻率倒頻譜係數的特徵向量我們都利用 MVA 去除部份通道不匹配問題，因為頻譜上受通道造成的偏移量相當於時間上的旋積性(convolutional)噪音，而梅爾頻率倒頻譜係數對平均值的削減正可以對抗旋積性噪音下之失真，至於變異數正規化與濾波器的使用則分別可以對抗加成性(additive)與高功率加成性噪音下的失真。在文字不特定語者驗證系統後段的分數方面，更透過改良式測試分數正規化做補償，將同儕語者模型(cohort model set)分數的平均值與變異數來調整目標語者模型的分數，經由減去平均值可以使冒充語者分數的分佈中心移至原點，而除以變異數則能將冒充語者分數分佈之標準差限定為一，這樣的方法不僅可以拉開目標語者與冒充語者之間的分佈進而改善正確率，還能讓決定接受與否的門檻值更容易產生。另外，多層感知機可用來結合各模組之語者的測試分數，進一步把頻譜系統及韻律系統融合，以便強化驗證系統之效能。

### 3. 潛在韻律分析

在語音訊號中，韻律訊息的動態變化受到各種潛藏因素的影響更甚於語者本身特性，譬如說話速度、情緒轉變以及說話內容等等，因此我們所觀察到韻律軌跡表象的變異量是相當大。而跟西方

語言比較之下可知漢語屬於一種聲調語言，隱藏於內的聲調更是個關鍵的因素，將會大大地影響韻律軌跡的變化。

一般來說，prosody state N-gram 語者模型 [6-8] 已經是以韻律特徵為主之驗證系統所採納的方法，然而想要能夠可靠地估測出此 N-gram 語者模型，則擁有大量的訓練與測試語料通常是先決條件，譬如說知名的 NIST2001-SRE Extended Data Task 中，分別使用 8 句和 2 句約兩分鐘的對話句子作為訓練與測試。然而在我們所提出的潛在韻律分析方法裡，大量語料將不會是必需條件，因為相同的語料庫下已能成功地運用在文字不特定之語者驗證 [8]，且平均來說僅僅只需共兩分鐘及三十秒的訓練與測試語料量，所以我們將嘗試著套用此方法在屬於聲調型的語言，特別是在漢語語言上的表現尚未能明確地得知。

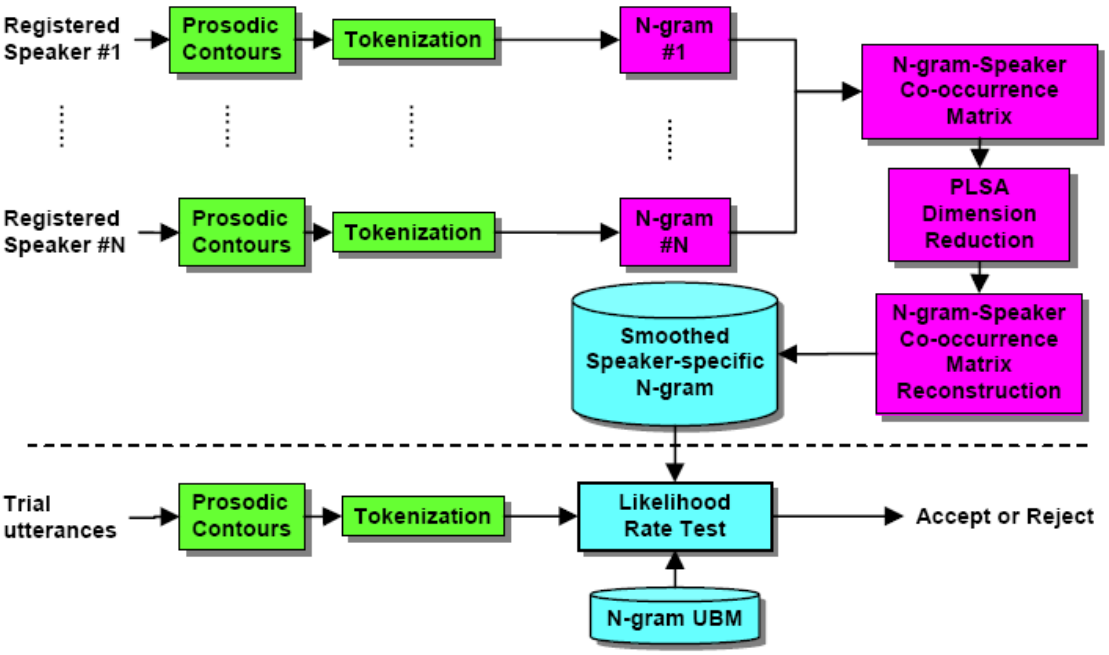
關於潛在韻律分析的基本構想是利用 PLSA 概念找出一個低維度的韻律資訊空間以表示語者的特性所在位置，主要是為了擷取出重要的韻律線索來鑑別語者之間的不同，再者是讓語者特定的韻律狀態 N-gram 語者模型能更可靠的建立。圖三是潛在韻律分析方法在語者驗證應用的方塊圖，首先必須把輸入語句的韻律軌跡經由 Tokenization 自動轉換成韻律狀態序列，並在訓練階段中建立起 N-gram 語者關係矩陣(co-occurrence matrix)，目的是集合每位語者的韻律行為特性來學習韻律狀態資訊和語者之間的相互關係。

圖四則說明 Tokenization 如何自動標記及轉換成韻律狀態序列。由 piece-wise curve fitting 先把每一段傳入的韻律軌跡擷取其韻律特徵向量，且許多鄰近的區段將串連成一個龐大的韻律特徵 supervector，而考量到音節為最小的韻律單位，所以採用五種音節層次的韻律特徵參數，包括一個母音區段的音高斜率(pitch slope)和長度的延長變化(lengthening factor)、兩個母音間的對數能量差和音高跳躍(pitch jump)以及兩個音節之間的暫停長度(pause duration)。此外為了移除語句發音內容（context-information）對韻律變化的影響，必須將韻律特徵參數做正規化的動作，藉由整個訓練語料所統計出來之韻律特徵參數的平均值及標準差，移去任何非韻律特性的影響。於是一個以向量量化為基礎，透過 Expectation- Maximization(EM)演算法訓練好的韻律模型便可以自動地把輸入語句所構成的 supervector 作符號的標記，並且再轉換成一連串的韻律狀態序列。

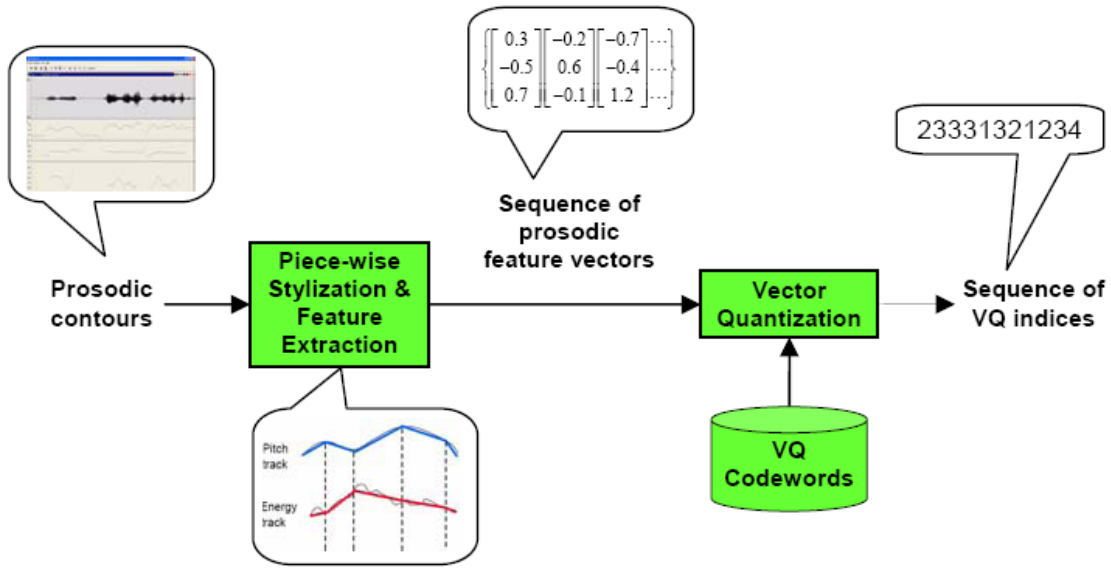
獲得韻律狀態 N-gram 語者關係矩陣後，由於訓練語料與測試語料之資料量的不足，在受此限制之下以韻律訊息所建構出的 N-gram 語者模型可能不夠具有統計特性，沒辦法準確的訓練出代表語者韻律特性的語者模型，所以必須再經過 PLSA 找出降低維度的韻律資訊空間，如圖五所示，而其分解定義如下，

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i) = P(d_i)\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \qquad (1)$$

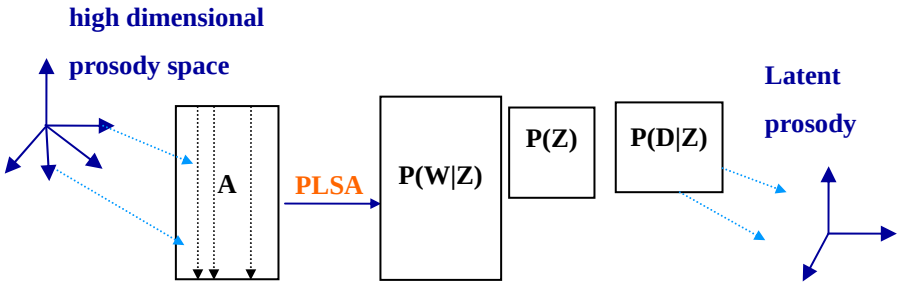
其中 $d_i$ 、 $w_i$ 、 $z$ 所代表的分別是 document、keyword、latent prosody factors， $P(d_i, w_j)$ 所代表的是 document( $d_i$ )與 keyword( $w_i$ )之間的結合機率，且 document 和 keyword 對應到韻律特徵的關係分別為語者及 N-gram term。這樣一來較為可靠的 N-gram 關係矩陣就能藉由幾個少數的特徵韻律向量(eigen-prosody vector)順利重建，達到語者韻律模型平滑化處理之目的。最後在測試階段我們只需將重建空間產生的 N-gram 語者模型和測試語句計算出相似度比值(likelihood ratio)，即可完成驗證的任務。




圖三、潛在語意分析方法輔助系統之方塊圖。




圖四、自動標記及轉換成韻律狀態序列的方塊圖。



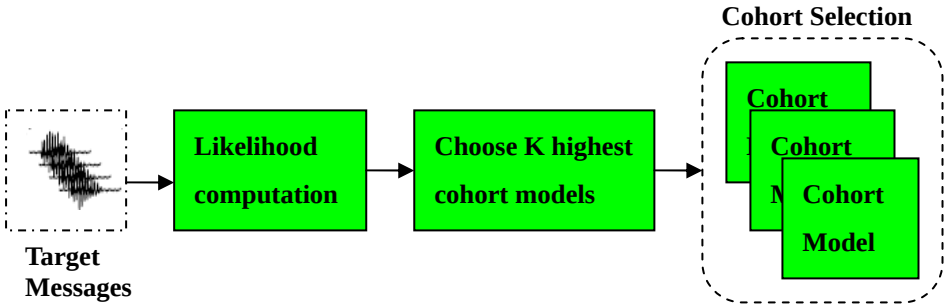

圖五、韻律特徵空間降維。

### 4. 改良式測試分數正規化

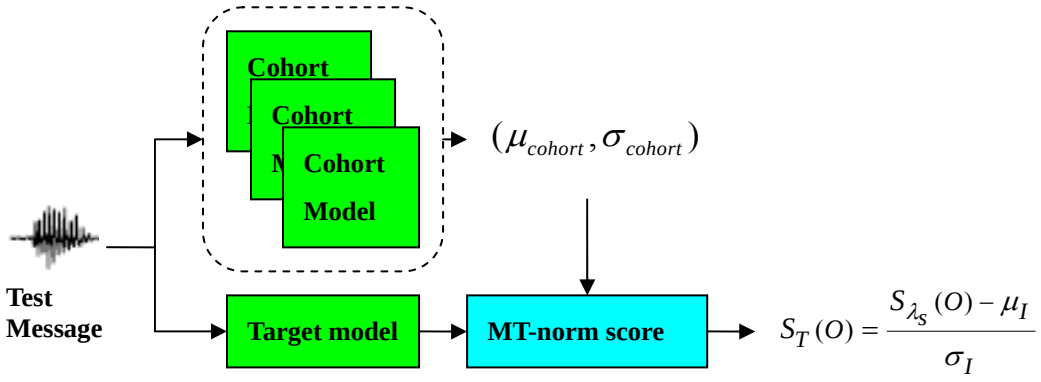
在改良式測試分數正規化的原理是利用一群相似於目標語者模型的同儕語者模型，估計出相似於每位不同目標語者的冒充語者，這和原始測試分數正規化方法有所不同，因為正規化所用的參數不再由同一組同儕語者模型得到，而是針對每個目標語者模型找出各別對應的同儕語者模型，如此才能找出每個目標語者模型真正的冒充語者群，這樣的方式亦可以帶來減少運算量的好處，因為對於和目標語者模型較不相似的同儕語者模型可以不再考慮其影響。而我們的同儕語者模型估測是根據訓練語料對每個語者特定模型量測 log likelihood score 而得，如圖六所示，這和 [11] 中利用距離的估測方式是有所不同的，主要是藉由 log likelihood score 的高低來決定同儕語者模型，並選出前面 K 個同儕語者模型來計算參數。接著每位不同目標語者依據相對應的同儕語者模型計算出分數的平均值與變異數作為調整目標語者分數的參數，其定義如下，

$$S_T = \frac{S_{\lambda_S} - \mu_I}{\sigma_I} \tag{2}$$

其中 $S_{\lambda S}$ 為測試語料與語者模型 $\lambda_s$ 所計算的 log likelihood score，$\mu_I$ 與 $\sigma_I$ 分別代表測試語料相對於同儕語者模型分數的平均值與變異數，$S_T$ 為經過測試正規化後的分數。從(2)式中看到測試正規化減去 $\mu_I$，這動作可以將冒充語者分數的分佈之中心移至原點，亦即同時拉大目標語者與冒充語者的分數分佈，而除以 $\sigma_I$ 則可以將冒充語者的分數分佈之標準差限定為一，進而提升正確率。而整個改良式測試分數正規化的架構則如圖七所示。




圖六、同儕語者模型的估測。




圖七、改良式測試分數正規化方塊圖。

## 5. 漢語之語者驗證實驗結果

### 5.1. ISCSLP2006-SRE 語料庫

在此語料庫中，不論是文字特定與文字不特定的語者驗證任務，都是來自中文語言資源聯盟所提供的 development 與 evaluation 資料庫，而此語料庫所有的聲音檔案都是採用 8kHz 取樣頻率，且為 16bits 單聲道的 PCM 格式。至於 evaluation 的語料庫中，其真實語者與冒充語者的測試樣本比例為 1 比 20。

*5.1.1 文字特定語者驗證之語料庫*

development 的資料取自於 CCC-VPR3C2005，語料庫包含了男女性各 5 位的個別資料量，每個人的聲音透過三種不同麥克風通道來獲得，分別用“micl”、“micr”及“micu”三種符號來表示，其中每位語者在分別通道上有五種句子會重複錄製 4 遍，而另外二十一種句子只會各錄製一次。對於 evaluation 的部份則共有 591 位註冊語者，每一位都有相同內容的三個句子，平均用來註冊的語句長度約有 4.5 秒，最後用來試驗的句子共有 11181 個且平均時間長度為 5.2 秒。值得一提的是每句發話開始都有很長的靜音，此外，某些雖然由相同語者所發出但卻為不同語句內容的句子，我們應該視為冒充語者並加以拒絕掉。

*5.1.2 文字不特定語者驗證之語料庫*

development 的資料取自於 CCC-VPR2C2005-1000，語料庫只包含了 300 位男性語者，每位語者含有兩種語句，分別由電話線(PTSN)及手機通道(GSM)所製成，所以總共有 600 個句子在內。evaluation 的部份則共有 800 位註冊語者，每一位都只會有一句從電話線或是手機通道所提供的語料，平均用來註冊的語句長度約有 36.2 秒，最後用來試驗的句子共有 11800 個且平均時間長度為 15.9 秒。

### 5.2. 實驗條件

本文中對於所有的頻譜特徵為主之語者驗證系統都用 39 維的梅爾頻率倒頻譜係數作為特徵參數，包括前 13 維倒頻譜係數(包含 $C_0$)及其差分值 Δ-MFCCs 與二次差分值 Δ2-MFCCs，至於音高與能量的軌跡則是藉由 snack 軟體套件中的 ESPS 音高擷取演算法來求得 [17] ，同時也計算其差分值及二次差分值，最後則將音高與能量連同梅爾頻率倒頻譜係數一併作為使用。

另外在驗證系統的合併方式，我們是運用共有 120 個隱藏節點的多層感知機，將頻譜特徵與韻律特徵在 evaluation 測試語料上所得到的分數作一結合。這部份的步驟是須先把 development 的語料區分為訓練和測試使用，其訓練語料部分用來訓練出各個系統的模型參數，而其測試語料的部份則用來得取每個系統的辨識結果，接著繼續再利用其測試語料的部份建置出多層感知機的各項參數，然後便可將各個系統的融合參數固定套用到之後 evaluation 測試語料所得到的分數上，而多層感知機的各項參數是由實驗數據所得，因此在後面的文字不特定與文字特定語者驗證實驗結果，所呈現的是系統之最佳效果。

語者驗證系統的錯誤率有兩種：一種是錯誤拒絕率(False Rejection Rate, FR)，即正確語者的分數小於門檻值造成拒絕的錯誤率。另一種是錯誤接受率(False Acceptance Rate, FA)，即仿冒語者的分數高於門檻值造成接受的錯誤率。FA、FR 這兩種錯誤率是一種取捨(tradeoff)的關係，若把門檻值提高，則錯誤拒絕率將會提高，而錯誤接受率則會降低；若門檻值降低，則錯誤拒絕率將會降低，而錯誤接受率則會提高。所以系統最後效能的量測主要是透過相等錯誤率(equal error rate, EER)及決策成本函數(decision cost function, DCF)來衡量。

相等錯誤率是一種評估語者驗證系統的方式，所謂的相等錯誤率就是錯誤拒絕率與錯誤接受率相等時的機率值，但在某些特殊的情形中，錯誤拒絕與錯誤接受的後果和重要性並不相等。舉例來說，語者驗證應用在金融交易的情況，為了避免冒領盜用，因此錯誤接受的機率必須減至最低。而決策成本函數則是被定義成一種錯誤機率的加權總和，如下所示。

$$C_{DET} = C_{Miss} \cdot P_{Miss} \cdot P_{Target} + C_{False} \cdot P_{False} \cdot \left(1 - P_{Target}\right) \tag{3}$$

其中 $C_{Miss} = 10, C_{False} = 1, P_{Target} = 0.05$ 。

另外，呈現錯誤拒絕率及錯誤接受率的方式則使用偵測錯誤交易曲線圖(Detection Error Tradeoff Curve, DET Curve)，此種方式是假設目標語者和仿冒語者的對數相似度比分數為兩各個不同的高斯分佈，隨著門檻值的變化表現出相對應錯誤拒絕率及錯誤接受率的曲線變化。

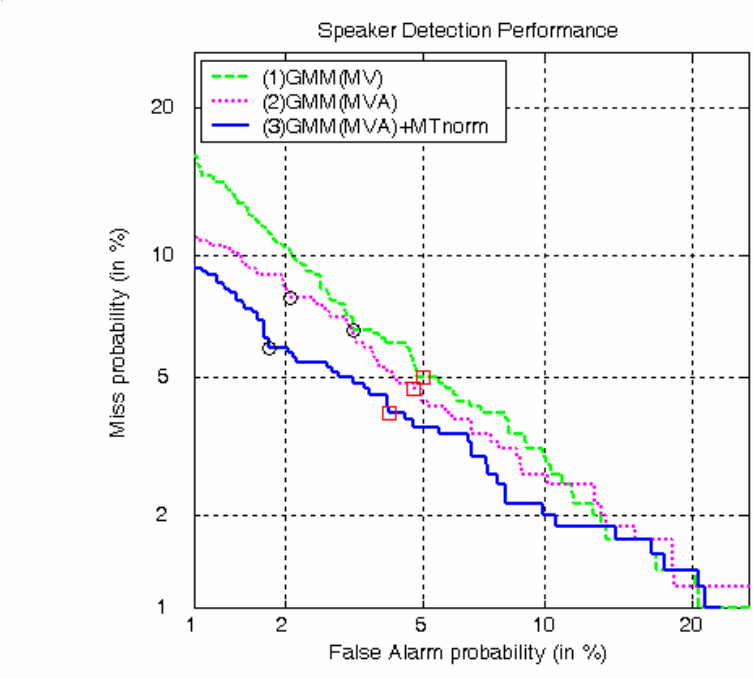
### 5.3. 文字不特定與文字特定語者驗證實驗結果

#### *5.3.1 文字不特定語者驗證結果*

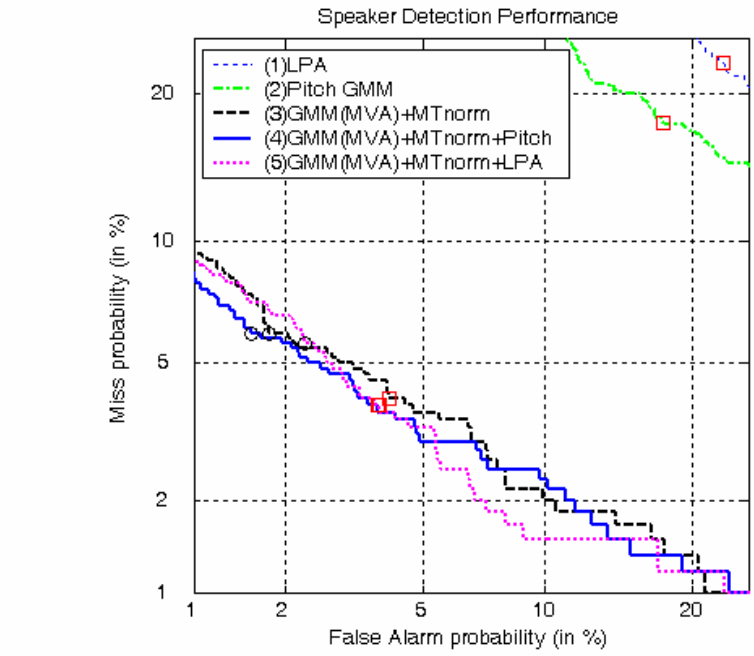
首先，頻譜特徵為主的 MAP-GMM 驗證結果會作為各項模組比較的標準，其通用背景模型一律為所有註冊語者訓練語料集成並用 1024 混合數組成，而語者特定的高斯混合模型是利用其註冊語料向通用背景模型調適得到。針對系統前端有兩種特徵正規化的方法會作為考量，包括 cepstral mean and variance normalization (MV)及 MVA，這兩種方式的結果在圖八可看出，MVA 的效果明顯比 MV 好了許多，因此我們將 MAP-GMM 和 MVA 的組合作為最基本的語者驗證系統。接著系統後端的改良式測試分數正規化則以 320 個相似語者為主，由圖八可看到此方法對驗證統確實有很大影響並大幅改善 MAP-GMM 的結果，由此可知 MVA 與改良式分數正規化法不僅相當有效且是互相補償。所以我們在文字不特定語者驗證中，以 MVA 與改良式分數正規化法和 MAP-GMM 的結合方式，作為頻譜特徵方面最佳的架構，爾後再加入韻律特徵的輔助。

在圖九中我們看到兩種韻律模型化的方法被用來和頻譜特徵最佳之效果做一結合。以高斯混合模型方式來說，該語者特定 64 混合數的韻律模型是直接由其註冊語料訓練而成，而非透過背景模型來調適，其驗證結果的相等錯誤率及決策成本函數分別為 17.7%和 0.223；另外潛在韻律分析方式則使用到 bi-gram 模型及 11 個狀態的向量量化(8 個為音高與能量使用，3 個為 pause segments 所用)，可以讓潛在韻律空間中的文件大小從 112(11*11-9)個維度減少至 30 個，這表示說每位語者其 N-gram 模型的平均參數量可從 112 降到僅僅只有 34.2 維，而其所帶來的好處是大幅的簡化了系統的複雜度，其驗證結果的相等錯誤率及決策成本函數分別為 22.7%和 0.272。

在韻律特徵方面，音高與能量之高斯混合模型及潛在韻律分析的相等錯誤率分別為 17.7%和 22.7%，這樣的結果以強化頻譜特徵為主的輔助角度來看已經是很不錯的，而將這兩種韻律模型化的方法與頻譜特徵最佳結果合併後，分別能讓驗證系統再從相等錯誤率 4.0%及決策成本函數 0.047 改善至 3.8%與 0.045 以及 3.8%與 0.050，可見韻律特徵對頻譜特徵的系統確實是產生輔助的效益。此外，在圖一及圖九的結果可看到文字不特定的語者驗證中，對於韻律特徵的使用並未先做特徵正規化的處理，因為我們主要是先決定好頻譜特徵方面驗證效果最好的架構後，再藉由韻律特徵的輔助作用，用以強化使用頻譜特徵之語者驗證系統的效能，因此沒有討論韻律特徵在未正規化之頻譜特徵下對於系統的影響，相對的文字特定語者驗證也是如此。




圖八、文字不特定語者驗證系統，在頻譜特徵上使用不同前後端處理方式的 DET 曲線圖。




圖九、包括 5 種不同文字不特定語者驗證系統之 DET 曲線圖。

*5.3.2 文字特定語者驗證結果*

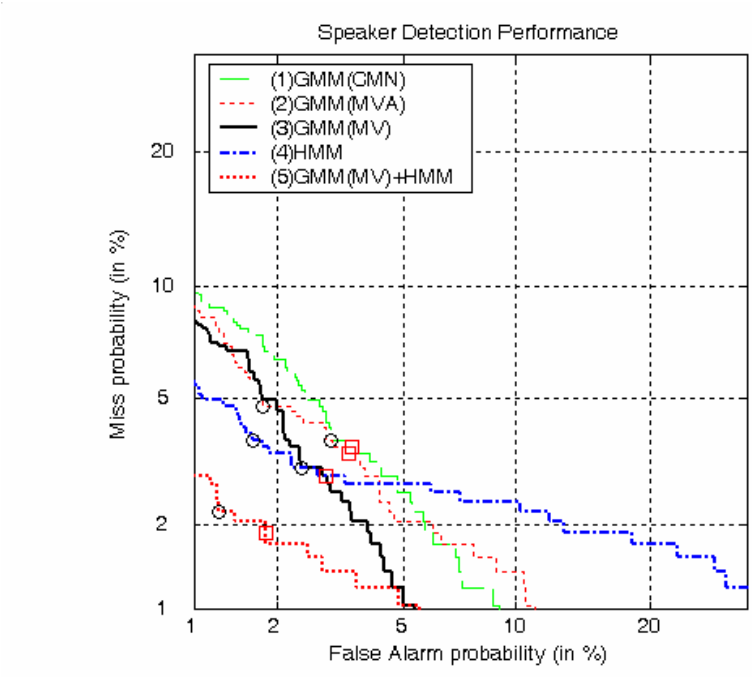
在頻譜特徵的實驗中，針對系統前端處理則考量 cepstral mean normalization(CMN)、MVA 及 MV 三種方式，從圖十的結果可知 MVA 的驗證結果略勝 CMN，然而我們發現在文字特定任務裡，MV 的表現更優於 MVA，這可能是因為在文字特定的語料庫是由麥克風錄製而成，且所有測試

時所用的麥克風特性都在訓練過程中遇過，如此現象在隱藏式馬可夫模型系統中將是不謀而合。

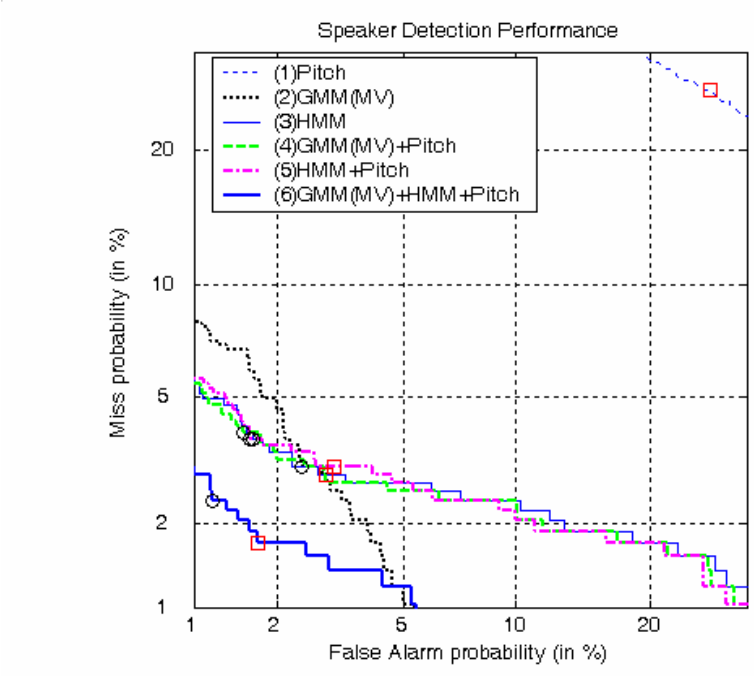
文字特定語者驗證中，有文字限定的語者高斯混合模型及隱藏式馬可夫模型兩種模組被用來建構驗證系統，包含 16 混合數的高斯混合模型及 8 混合數的隱藏式馬可夫模型，而隱藏式馬可夫模型的狀態數目是根據註冊語料中文字的多寡來做調整。從圖十的結果看到，雖然兩種方法結果的曲線趨勢有所不同，但高斯混合模型最佳的相等錯誤率及決策成本函數分別為 2.9%和 0.038，和隱藏式馬可夫模型 2.9%和 0.034 的表現卻是差不多。圖十亦可看見有趣的結果是當這兩個系統結合後會對結果產生強勁的改善，相等錯誤率及決策成本函數分別為 1.9%和 0.023，而這或許就是對這兩種系統的互補性做了最佳的驗證，因為高斯混合模型僅需少量語料便能對每個音框的倒頻譜係數分佈作模型化，反觀隱藏式馬可夫模型多量需求才能仔細描繪出倒頻譜係數的暫態軌跡，可見兩者所長不同於語料量所供應的大小。

當頻譜特徵為主的最佳系統建立好之後，再來就是關於韻律特徵與頻譜特徵的結合，相較於文字不特定的任務，我們只運用了音高及能量的 8 混合數高斯混合模型。圖十一顯示韻律特徵和上面兩套系統的合併結果，與文字限定語者高斯混合模型的結合是相等錯誤率及決策成本函數分別為 2.9%和 0.034，而隱藏式馬可夫模型則是 3.1%和 0.034，雖然由此看到效果沒能有顯著的改善，不過有趣的在於我們將所有的方法結合後，整個系統的錯誤率又再下降些許，如圖十一所示，而這似乎又證實了韻律特徵對頻譜特徵的互補在文字特定的系統上仍舊是很有效果的。

另外在圖二及圖十的結果不難發現，文字特定語者驗證系統並未呈現改良式分數正規化的結果，這是因為改良式分數正規化方法中，不論我們使用多少相似語者的數量，都未能讓系統有所改善。如此的結果和文字不特定語者驗證雖不一致，但由於文字特定語者驗證的測試語料通道特性都是在訓練階段看過的，因此改良式分數正規化無法在文字特定語者驗證上有所貢獻，可能就是測試語料與訓練時通道一致的關係，故針對此語料庫的文字特定語者驗證並不適合使用改良式分數正規化。



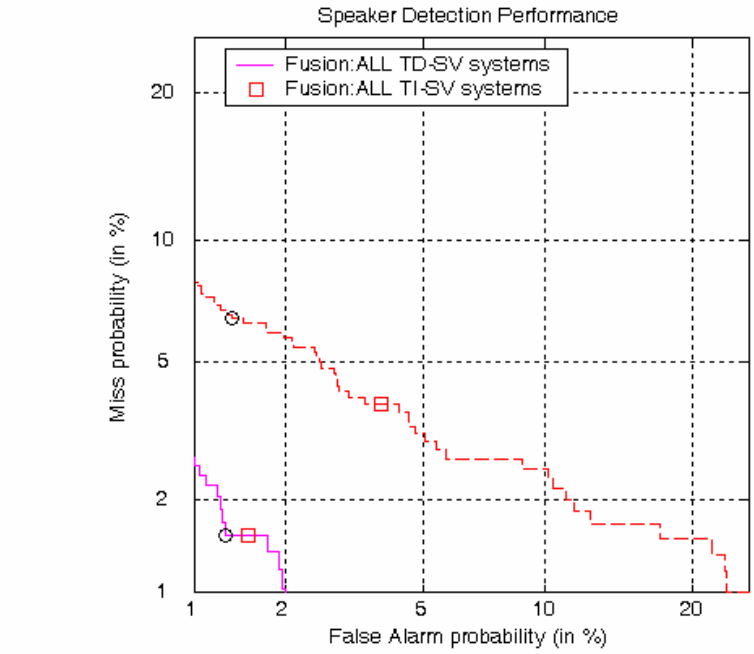

圖十、文字特定語者驗證系統，在頻譜特徵上使用不同處理方式的 DET 曲線圖。




圖十一、包括 6 種不同文字特定語者驗證系統之 DET 曲線圖。

*5.3.3 文字不特定與文字特定之語者驗證結果比較*

在此我們將融合文字不特定與文字特定語者驗證系統中所有的模組，因為不僅只考慮頻譜特徵與韻律特徵之間的互補特性，不同模組之間的相關性亦可為有效的特徵，所以再藉由多層感知機不同於一般線性組合的方式，將所有模組作全面性合併，結果如圖十二所示。




圖十二、結合系統所有模組的 DET 曲線圖。

此外，表一與表二呈現文字不特定與文字特定語者驗證系統的所有結果。從表一中的比較我們發現韻律特徵的作用必須建構在頻譜特徵的系統上，儘管 22.7%和 17.7%的相等錯誤率及 0.272 與 0.223 決策成本函數不能和頻譜特徵的 4.0%和 0.047 相比，但兩者合併後的相等錯誤率及決策成本函數為 3.8%和 0.045，確實改善了文字不特定語者驗證系統正確率。而表二中更顯示文字特定語者驗證的韻律特徵對於系統強化有很不錯的幫助，讓系統從頻譜特徵最佳結果的 1.9%與 0.023，大幅改善至 1.5%與 0.020。

表一、8 種不同文字不特定語者驗證系統之結果比較。

| | ERR (%) | DCF |
|---|---|---|
| (1) LPA | 22.7 | 0.272 |
| (2) Pitch GMM | 17.7 | 0.223 |
| (3) GMM(MV) | 5.0 | 0.064 |
| (4) GMM(MVA) | 4.7 | 0.060 |
| (5) GMM(MVA)+MT-norm | 4.0 | 0.047 |
| (6) GMM(MVA)+MT-norm+Pitch | 3.8 | 0.045 |
| (7) GMM(MVA)+MT-norm+LPA | 3.8 | 0.050 |
| (8) Fusion ALL | 3.8 | 0.045 |

表二、10 種不同文字特定語者驗證系統之結果比較。

| | ERR(%) | DCF |
|---|---|---|
| (1) Pitch | 25.9 | 0.297 |
| (2) GMM(CMN) | 3.6 | 0.047 |
| (3) GMM(MVA) | 3.4 | 0.041 |
| (4) GMM(MV) | 2.9 | 0.038 |
| (5) HMM | 2.9 | 0.034 |
| (6) GMM(MV)+Pitch | 2.9 | 0.034 |
| (7) HMM+Pitch | 3.1 | 0.034 |
| (8) GMM(MV)+HMM | 1.9 | 0.023 |
| (9) GMM(MV)+HMM+Pitch | 1.7 | 0.023 |
| (10) Fusion ALL | 1.5 | 0.020 |

**5.4. 結果討論**

在文字特定語者驗證系統中，如圖二所示，可發現並沒有運用潛在韻律分析方法，這是因為受限於語料量的因數，過於貧乏的語料無法建立有效的韻律特徵，雖然潛在韻律分析在文字不特定上有不錯表現，但對於文字特定語者驗證，仍需藉由調適或其他方式來解決語料問題，而這將是我們往後發展韻律特徵的目標。另外對於潛在語意分析方法輔助系統，如圖三所示，我們先透過自動標記及轉換成韻律狀態序列的方式，如圖四所示，再由 PLSA 作後續處理，這是因為目前尚無

法直接用 PLSA 獲得韻律軌跡對應於語者特性的關係，所以往後針對 PLSA 分析能力的運用將會作改善，以期能更完整保留語者的韻律訊息。

前面的實驗結果顯示，雖然韻律訊息的確能輔助傳統使用頻譜特徵之語者驗證系統，有效提升系統效能，尤其是在文字特定語者驗證系統上，但我們發現與韻律特徵在英文語者驗證系統的貢獻相比，此漢語語料庫的語者驗證，似乎沒能因為屬於聲調語言而突顯出韻律特徵的關鍵性，這可能攸關於些許條件上的差異，好比在此語料庫中的語料量較為簡短，且說話內容是以閱讀句子為主，而非一般的對話形式，造成每位語者的韻律特徵會較有相似之處。此外我們並未針對漢語中的聲調做特別處理，而聲調對於音高軌跡是有極大的影響力，所以韻律特徵的求取並無法很精確獲得。再者，雖然沒有對聲調做正規化，僅利用粗略的韻律特徵卻可對頻譜特徵的系統有不錯的改善，相較於英文中以較高相等錯誤率為基礎的改善來說，韻律訊息在漢語語者驗證的結果應該算是有良好貢獻。

## 6. 結論

在本文章，我們針對文字特定與文字不特定的語者驗證任務提出我們的語者驗證系統，集合眾多方法之長處來完成，包含了前端特徵正規化的 MVA、後端的改良式測試分數正規化、頻譜特徵的高斯混合與隱藏式馬可夫模型以及韻律特徵的潛在韻律分析與高斯混合模型，而最後的多層感知機更是用來耦合各個相異的驗證系統，使得表現的效果有更進一步的改善。由實驗結果來看，有幾項要點是值得注意的，首先是 MVA 與改良式分數正規化法確實成功地在語者驗證上補償通道不匹配的問題，再者是不論文字特定與文字不特定的情況下，韻律特徵都能與頻譜特徵作完善的結合。而本論文的結果亦顯示了在漢語語者驗證上仍有進一步發掘出關於韻律線索的必要。

## 7. 致謝

本研究受國科會專題研究計畫及教育部計畫補助，計畫編號分別為 NSC 94-2213-E-027-003 與 A-94-E-FA06-4-4。

## 8. 參考文獻

1. M. Faundez-Zanuy, E. Monte-Moreno, “State-of-the-art in speaker recognition”, *IEEE Aerospace and Electronic Systems Magazine*, Vol. 20, Issue 5, pp. 7-12, 2005.
2. “NIST - Speaker Recognition Evaluations”, http://www.nist.gov/speech/tests/spk/.
3. The CSLP Speaker Recognition Evaluation (SRE) 2006, http://www.iscslp2006.org/.
4. Chinese Corpus Consortium (CCC), http://www.cccforum.org/.
5. “NIST 2001 Speaker Recognition Evaluation - Extended Data task”, http://www.nist.gov/speech/tests/spk/2001/extended-data/.
6. D.A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones and B. Xiang, “The SuperSID project: exploiting high-level information for high-accuracy speaker recognition,” *Proc. ICASSP’03*, Vol. 4, pp. 784-787, 2003.
7. A.G. Adami, R. Mihaescu, D.A. Reynolds and J.J. Godfrey, “Modeling prosodic dynamics for speaker recognition,” *Proc. ICASSP’03*, Vol. 4, pp. 788-791, 2003.
8. Z.H. Chen, Z.R. Zeng, Y.F. Liao, and Y.T. Juang, “Probabilistic Latent Prosody Analysis For Robust Speaker Verification,” *ICASSP’06,* 2006.

9. C.P. Chen and J. Bilmes, "MVA Processing of Speech Features", to appear in *IEEE Trans. on Speech and Audio Processing.*
10. D.A. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, Vol. 10, pp. 19-41, January 2000.
11. D. Sturim and D.A. Reynolds, "Speaker Adaptive Cohort Selection for Tnorm in Text-Independent Speaker Verification", *ICASSP'05,*2005.
12. LNKnet Pattern Classification Software, http://www.ll.mit.edu/IST/lnknet/.
13. J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," *Proc. ISCA Workshop on Speaker Recognition - 2001: A Speaker Oddyssey*, June 2001.
14. A. de la Torre, A. M. Peinado, J. C. Segura, J. L. P-C, M. C. Benitez, A. J. Rubio, "Histogram Equalization of Speech Representation for Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*. Vol. 13, pp. 355-366, May 2005.
15. T. Hofmann, "Probabilistic Latent Semantic Analysis," *Proc. Uncertainty in Artificial Intelligence* 1999, 1999.
16. T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, pp. 177-196, 2001.
17. K. Sjölander, "Snack sound toolkit", http://www.speech.kth.se/snack/.