結合聲學與韻律訊息之強健性語者辨認方法

Combination of Acoustic and Prosodic Information for Robust Speaker Identification

[1]廖元甫，[1]莊智顯，[2]陳子和，[2]莊堯棠

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

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

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

摘要

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

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

1. 緒論

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

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

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

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

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

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

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

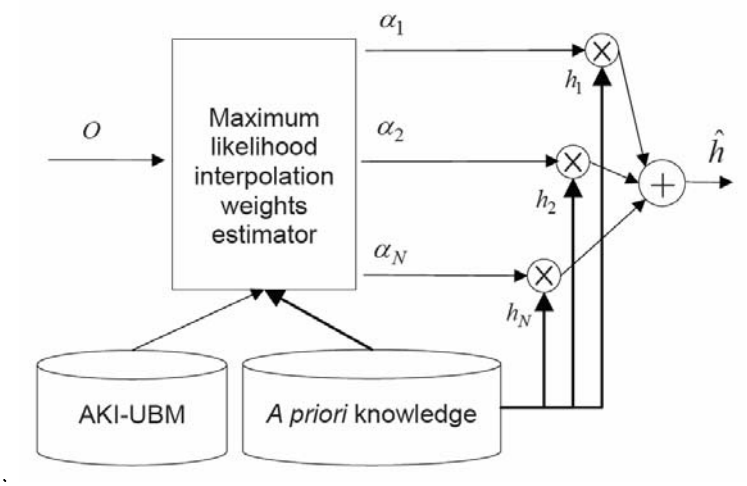
2. 最大相似度先驗知識內差法（ML-AKI）

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

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

其中 α_n 為內插的權重，h_n 為模型領域上的先驗知識。

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



圖一、ML-AKI 的架構圖

2.1. 基於 MLLR 之話筒特性先驗知識

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

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

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

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

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

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

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

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

2.2. 最佳化內插權重值求取

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

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

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

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

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

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

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

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

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

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

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

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

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

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

接著可將式子(8)表示如下

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

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

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

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

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

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

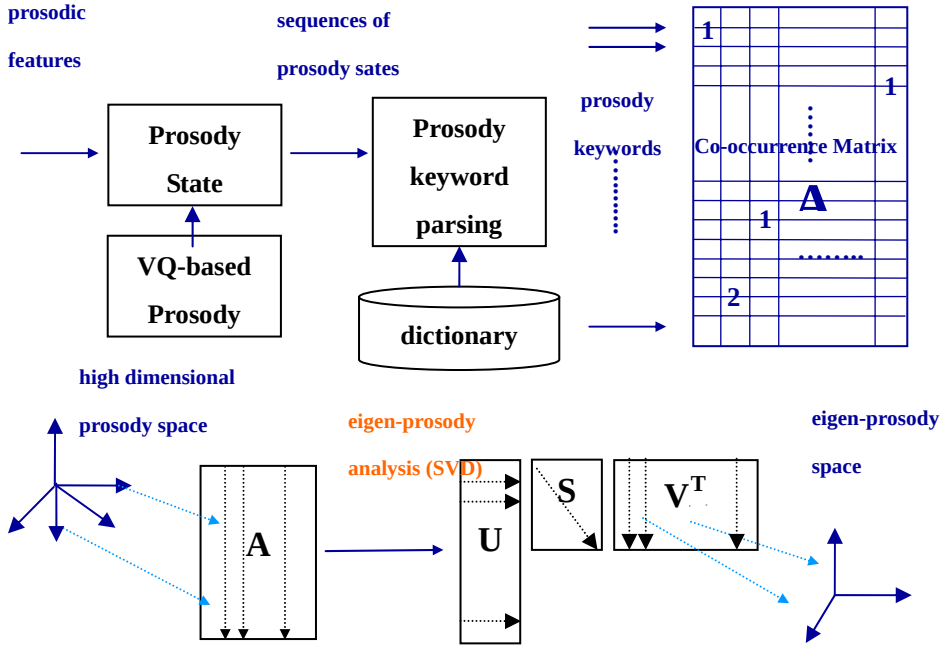
3. 韻律特徵分析（EPA）

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

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

韻律訊息關鍵詞作詢問(query)，以擷取最相似的註冊語者。

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



圖二、EPA 的架構圖

3.1. VQ 韻律模型與自動韻律狀態標記

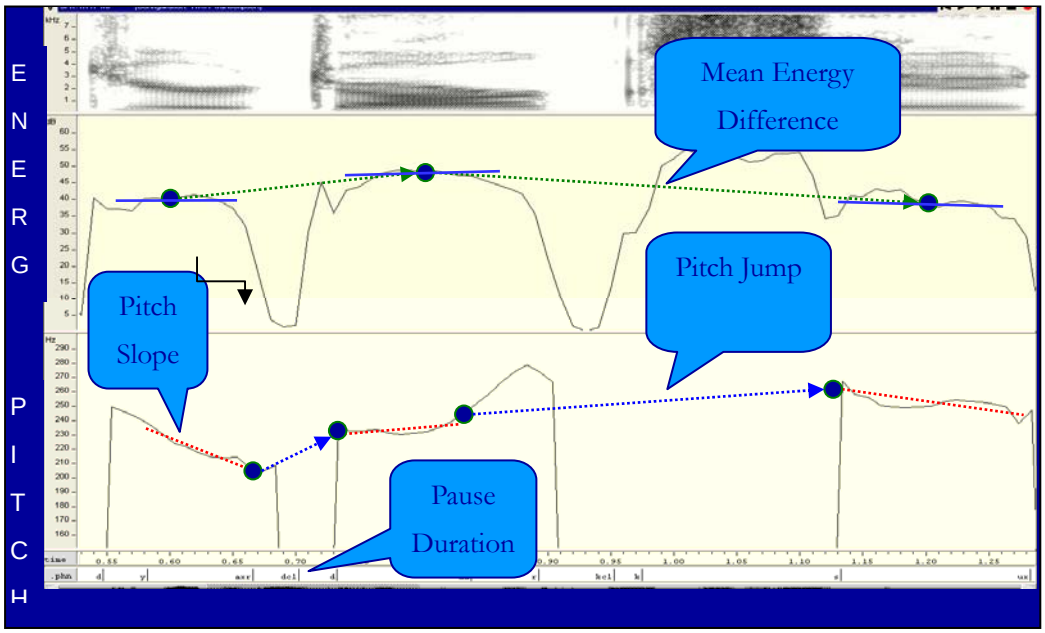
3.1.1. 韻律特徵參數求取與正規化

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

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

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

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



圖三、音節層次之韻律參數求取

3.1.2. 以 VQ 為基礎之韻律訊息模型

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

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

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

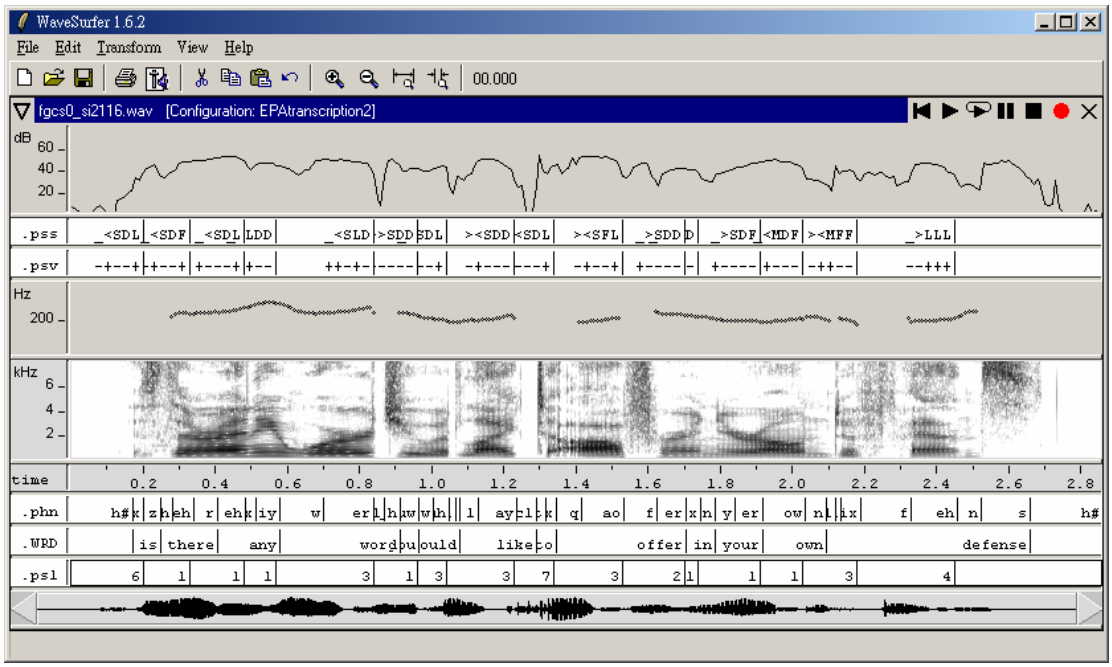
Feature/State	1	2	3	4	5	6	7	8
Pitch slop	-0.1	0.7	-0.1	-0.2	0.1	0.3	-0.2	-2.5
Energy diff.	-0.4	-0.5	-0.8	-1.9	-0.1	0.2	1.3	0.1
Pitch jump	-0.2	-0.2	1.3	1.4	-0.1	-0.9	0.3	-0.6
Lengthening	0.3	-0.5	0.3	1.4	0.1	-0.1	0.1	0.1
Pause	0.4	-0.5	0.5	2.6	0.2	-0.3	0.3	0.1

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

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

3.1.3. 自動韻律訊息標記

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



圖四、典型的自動韻律狀態標記範例。

3.2. EPA 分析步驟

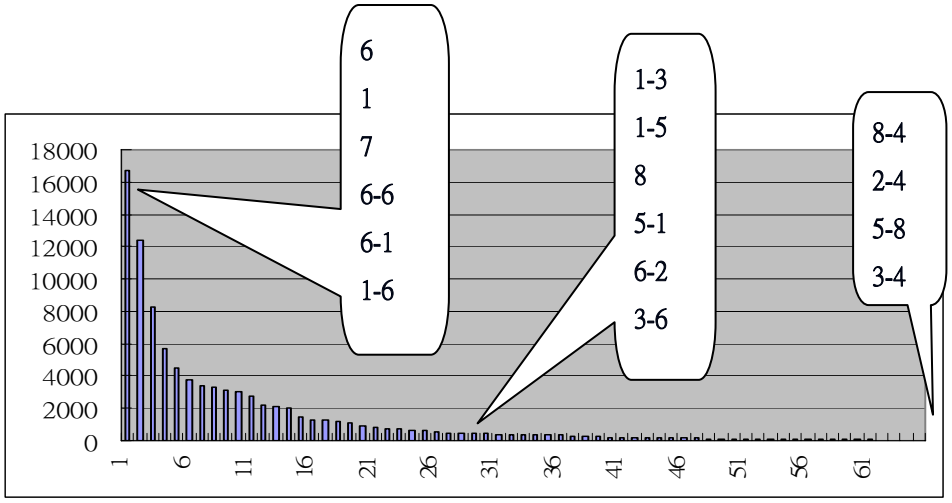
利用 EPA 輔助語者辨認包括四個步驟，包括（1）從所有語者的韻律狀態序列擷取韻律關鍵

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

3.2.1. 韻律關鍵詞詞典

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

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



圖五、韻律關鍵詞詞典詞頻統計。

3.2.2. 語者—韻律關鍵詞關係矩陣

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

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

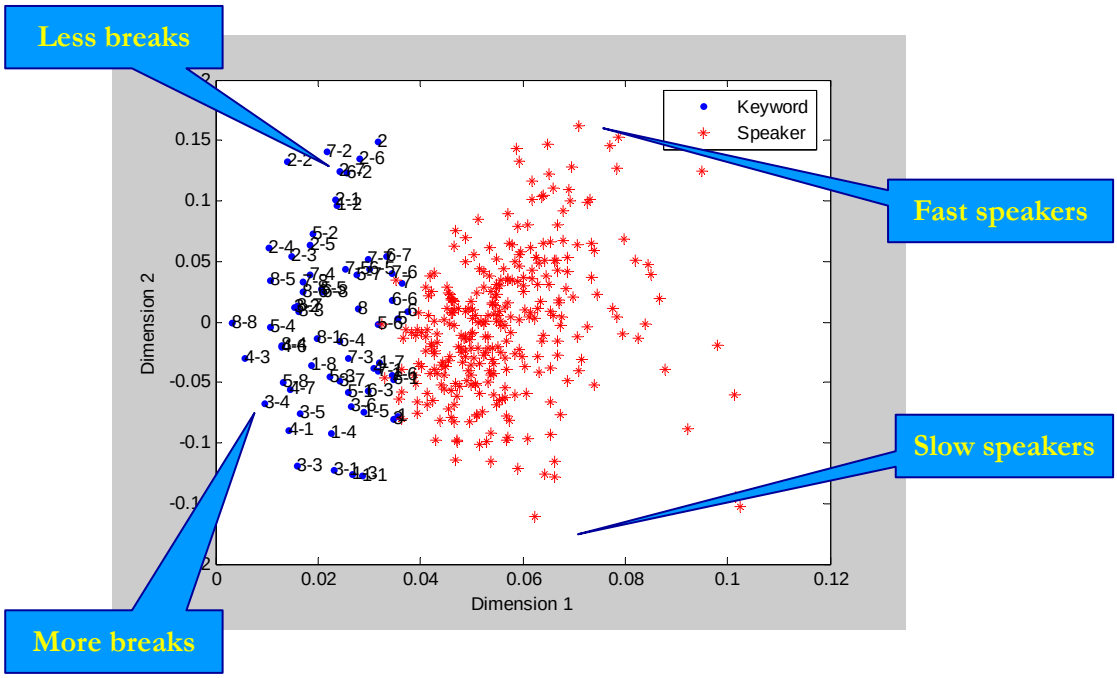
3.2.3. 特徵韻律訊息分析

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

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

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

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



圖六、語者韻律特徵空間。

3.2.4. 語者辨認分數的測量

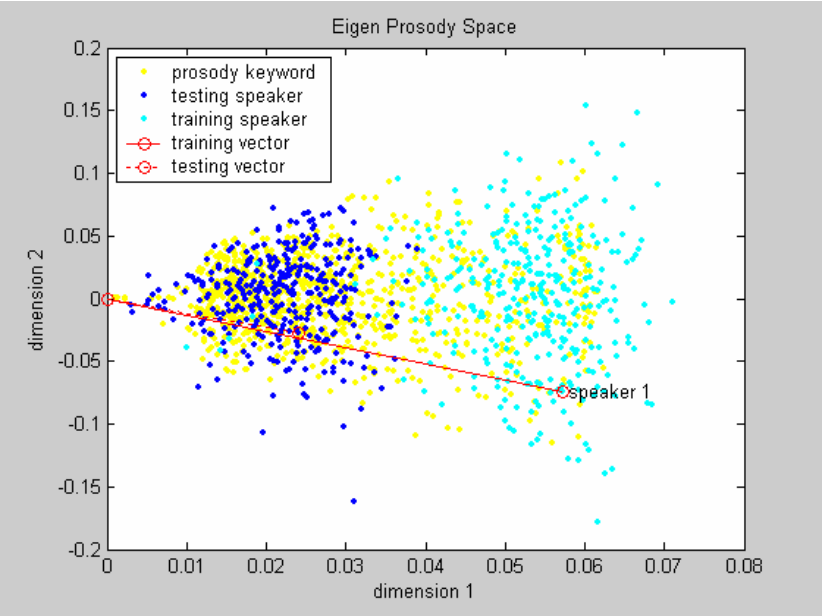
最後可以把語者辨認的問題，看成相當於文件擷取的問題。將輸入之測試語者的測試句子，同樣轉成韻律狀態索引序列，利用韻律關鍵詞詞典剖析後，當做一虛擬詢問向量y_Q，再利用式子(15)，將其投影至特徵-韻律語者空間，得到詢問向量(query vector) v_Q

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

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

者，如圖七所示。

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



圖七、語者在特徵韻律空間的分佈與辨認分數量測。

4. EPA，ML-AKI＋MCE 和 MAP-GMM/CMS 融合方法

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

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

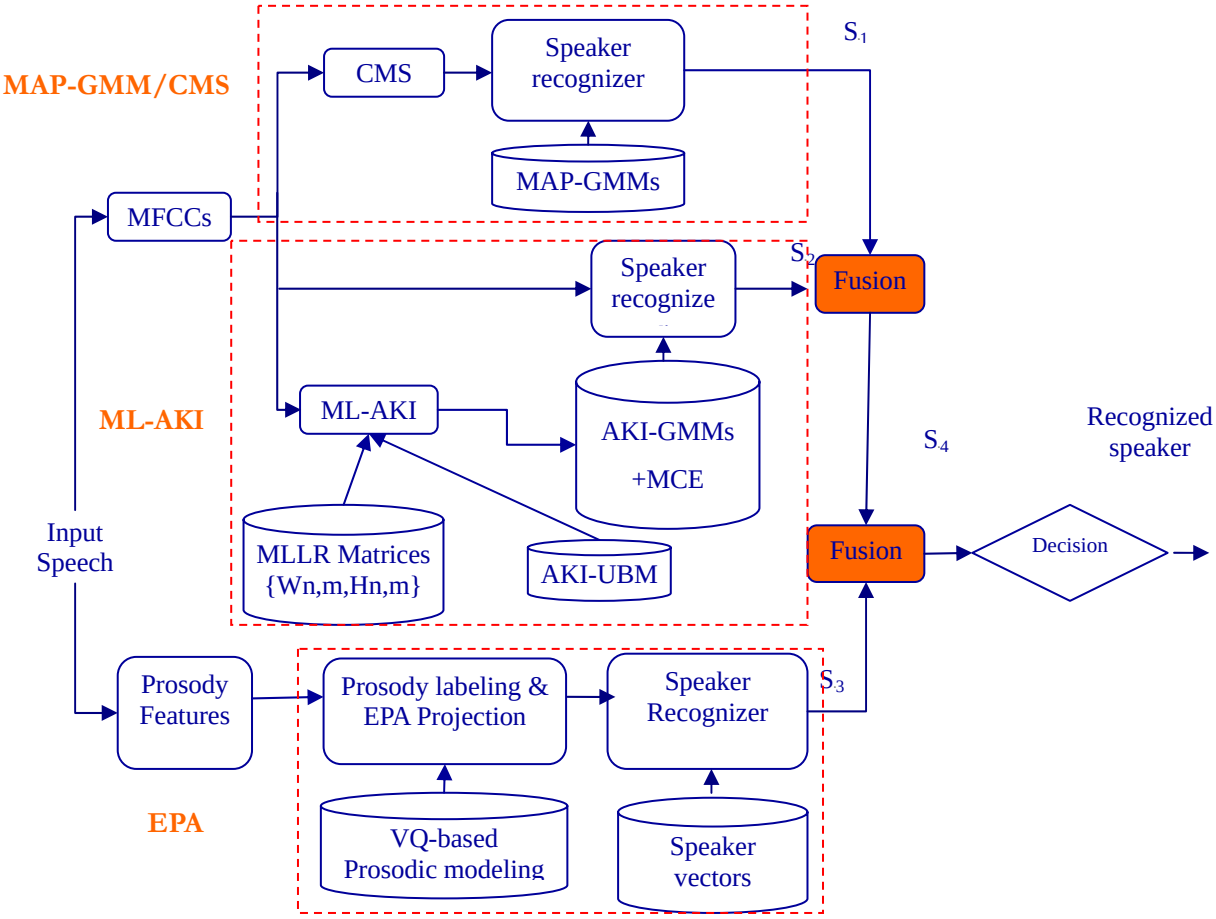
$$S_f = \lambda \cdot \frac{(s_1 - \bar{s}_1)}{\sigma_{s_1}} + (1-\lambda) \cdot \frac{(s_2 - \bar{s}_2)}{\sigma_{s_2}} \qquad (16)$$

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

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

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

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



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

5. 語者辨認實驗

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

5.1. HTIMIT 語料庫

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

5.2. 實驗條件

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

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

5.3. 實驗結果

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

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

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

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

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

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

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

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

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

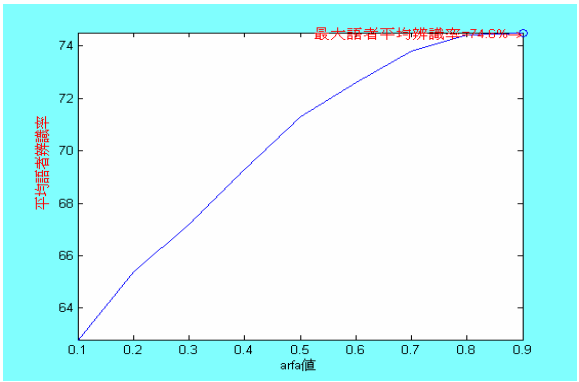
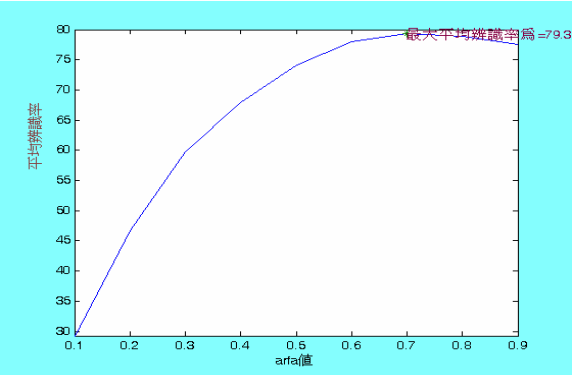


圖 九、融合 ML-AKI＋MCE 和 MAP-GMM/CMS 分數時權重 λ 值與辨認率之關係。

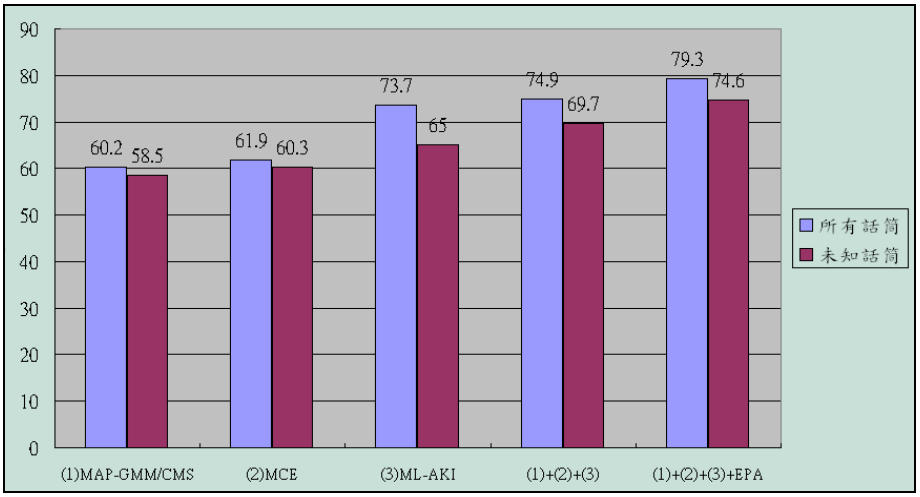


圖十、融合聲學訊息與韻律訊息時 α 值與辨認率之關係。

5.4. 實驗總結與討論

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

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



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

6. 結論

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

7. 參考文獻

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

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

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

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

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

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

【7】 S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, "Indexing by Latent Semantic Analysis," Journal of the Society for Information Science, vol. 41(6), pp. 391-407.1990.

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

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

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

【10】 http://www.speech.kth.se/snack/

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

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

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

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