國語電話語音辨認之強健性特徵參數及其調整方法

黃儀芬 王小川

國立清華大學電機工程學系

摘要

本文對經電話線之國語語音辨認，就兩種不同的強健性解決策略進行探討。(1) 在強健型特徵參數萃取上，採用倒頻譜平均值正規化法（CMN）及自相關係數頻域正規化法（DFT-MN-AUTO）。 (2) 在特徵參數調整法方面，採用特徵參數扣減法（SBR）、階層式特徵參數扣減法（HSBR）及統計式對應法（SM）。並提出叢集模型概念及其改良架構。。以國語語音資料庫 MAT-160 進行實驗，結果顯示在有通道效應情況下，CMN 效果最佳，若有背景雜訊情況，則是 DFT-MN-AUTO 效果最好。在特徵參數調整方法方面，以訓練端及測試端皆進行階層式特徵參數扣減法運算，並採用加權式叢集模型，可得最佳辨識率。

一、 緒論

近年來，隨著電腦運算速度和儲存設備的大幅進展，使語音辨識技術得有突破性的發展，也因此，讓電腦、儀器聽懂人說的話，成爲一個非常具有實現可能性的課題。目前，在安靜環境下針對特定語者模式的技術已有相當不錯的研究成果，市面上相關的產品也紛紛出現。但是，若要更廣泛的使用語音辨識的技術，便必須再進一步發展針對不特定語者的辨認技術以及克服訓練環境語料與測試環境語料不匹配的問題。就經電話線之國語語音辨認而言，已有若干研究與探討[1,2,3,4]。

語音辨認流程可分成訓練端及測試端兩個部分。在訓練端，透過語音與詞庫資料，經由各種演算法的訓練，去預估與統計出聲學模型及語言模型。在測試端，則應用所得之模型，求出辨認結果。在聲學模型訓練的部份，著重在語料的狀態估測與切割，普遍被使用的演算法有 k 均值切割演算、維特比最佳路徑搜尋、EM 演算法…等。在語言模型訓練部份，則以樹狀結構演算最爲普遍。

電話線語音和在安靜環境下錄製的麥克風語音相較，具有較複雜的失真問題。其失真來源包含：

- 發話端之環境雜訊（ambient noise）
- 電話線通道效應（channel effect）
- 頻寬限制（bandwidth limitation）

加上各種語音環境下皆存在的語者發音變異（speaker variation）問題，構成電話線語音辨認必得處理強健性的課題。

解決電話線語音辨識強健性的問題，可從不同領域著手[5,6,7,8]。假設 x 爲訓練環境之語料波形訊號，X 爲訓練環境之語料特徵參數，Λ_x爲訓練環境下之聲學模型；y 爲訓練環境之語料波形訊號，Y 爲訓練環境之語料特徵參數，Λ_y爲訓練環境下之聲學模型。增加強健性的方法大致可分類成三種：

- 具強健能力的特徵參數求取法 $R_\tau(\cdot)$
- 在特徵參數域中估測失真偏移量 $F_v(\cdot)$，將 Y 扣除此一偏移量，拉近與 X 距離，以期匹配 Λ_x。
- 估測模型轉換函式 $G_\eta(\cdot)$，使 Λ_x 能接近 Λ_y。

目前最普遍的語音模型訓練方式爲求取梅爾倒頻譜參數及使用隱藏式馬可夫模型。根據生理實驗結果，人耳聽覺在頻譜效應上並非線性，梅爾倒頻譜參數的求取，即是根據相對的轉換公式，以一系列非等距非對稱的「三角帶通濾波器」模擬耳蝸在接收訊號時的情形，求出頻譜振幅，再經由對數化及餘弦轉換，可得多維的梅爾倒頻譜參數。而差分化參數可以描述發音過程中的變動特性，因此也被廣泛的採用。

語音訊號具有暫態穩定的性質，故對於同一音節或同一句話而言，雖然每次發音的長短不同，但聲道變化的過程是相似的，語音波形訊號爲此聲道共鳴後的表現。因此，可將波形訊號視爲一雙重隨機程序的結果，以隱藏式馬可夫模型來模擬。

二、 國語語音之模型設計

國語是單音節的語言，每一音節由聲母、韻母、和聲調所組成。聲母有 22 類（包含一個空聲母），韻母有 40 類（包含二個空聲母分別對應於捲舌音與齦後音），聲調

有 5 類。若不考慮聲調化，可將之歸納成 412 個音節類型。同一音節內，聲母和韻母會相互影響，產生耦合的現象；此情形在跨音節語音的韻尾與聲母、複韻的兩個韻母之間亦會出現。

只考慮音節內耦合的現象，並採用右音相關聲韻母模型。其中 22 類聲母和 6 類首韻韻母共可歸納出 94 個右音相關聲母，同理空聲母也依據首韻韻母特性分成 6 類。每個聲韻母模型分別以 3 個和 4 個狀態表示，加上 1 個靜音狀態模型。共可得 464 個狀態模型。另鑑於男女音質的差異，實驗中皆採用男女生分類模型。

三、 語音資料庫

本文中用來訓練模型的語料庫共有兩套。第一套 MAT-160[2,13]，是電話線環境下的錄音語料，係國科會補助之語音收集計畫(MAT 計畫)所錄製，其內容經過安排設計，使能涵蓋國語語音中可能出現的音節，爲主要實驗使用之語料庫。第二套爲 MIC-101，係中華電信研究所錄製，是麥克風環境下所錄製的語料庫，內容爲 2 至 4 字的短字詞彙。表 1 及表 2 分別說明其內容。

MAT-160	
語料庫來源	中華民國計算語言學學會
錄音環境	透過電話網路經由個人電腦錄製
取樣頻率	8kHz
取樣位元數	16bits
語者	女 79 人、男 81 人
每語者語料內容	12 句單音節、30 句短字詞彙、10 句平衡長句
總句數 / 總時數	8320 句 / 5.01 小時

表 1：MAT-160 語料庫

MIC-101	
語料庫來源	中華電信研究所
錄音環境	透過麥克風在安靜環境下經由個人電腦錄製
取樣頻率	8kHz
取樣位元數	16bits

語者	女 51 人、男 50 人
每語者語料內容	50 句短字詞彙
總句數 / 總時數	5050 句 / 1.44 小時

表 2：MIC-101 語料庫

訓練語料爲 TEST-500，是 MAT 計畫中抽取之語音資料，在 1998 年語音辨認評比時採用之自行測試語料[14]，內容如下：

TEST-500	
語料庫來源	中華民國計算語言學學會
錄音環境	透過電話網路經由個人電腦錄製
取樣頻率	8kHz
取樣位元數	16bits
語者來源	女 15 人、男 15 人，與 MAT-160 無重複
語料內容	50 句單音節、150 句短句詞彙（text dependent） 300 句平衡長句（text independent）
總句數 / 總時數	500 句（音節總數 4736） / 0.45 小時

表 3：TEST-500 語料庫

四、 基本特徵參數及實驗

本研究中所設計之語音辨認基本系統架構中，特徵參數使用梅爾倒頻譜係數，求取法則如下表所示，此法所得的特徵參數標示爲 MEL：

音框長度	256 points
音框位移	128 points
MEL 特徵參數	12-order MFCC + 12-order-delta-MFCC + 1-order-delta-log-energy + 1-order-delta-delta-log-energy

表 4：MEL 特徵參數取法

爲瞭解在測試環境匹配或不匹配時的辨認結果，先就上述之訓練語料庫與測試語料庫，作基礎實驗。在實驗中，辨識率皆依照下列方式計算：

辨認率＝(正確音節總數－錯誤音節總數)／正確音節總數

錯誤音節總數＝取代音節總數＋刪除音節總數＋插入音節總數

將 MAT-160 與 MIC-101 訓練語料庫，以 MEL 爲特徵參數，分別訓練出聲學模型，以 TEST-500 爲測試語料，辨認結果如表 5。

	以 MAT-160 爲訓練語料			以 MIC-101 爲訓練語料		
	混合數			混合數		
	4	8	16	4	8	16
插入率(%)	3.53	3.82	3.61	4.22	3.86	3.93
刪除率(%)	2.22	2.11	1.94	6.65	7.45	7.47
取代率(%)	49.30	47.61	45.68	78.15	77.45	76.44
辨認率(%)	44.95	46.45	48.75	10.98	11.23	12.16

表 5：以 MAT-160 及 MIC-101 爲訓練語料之實驗結果

由此實驗結果可看出，當測試環境與訓練環境接近時，可得較佳之辨認率。而當測試環境不匹配時，辨認率即急驟下降，在混合數等於 4 和等於 16 的情況下，分別降低了 33.97%和 36.59%，降幅達 75%以上。混合數的多寡亦是決定因素，混合數的增加會提升辨認正確率，其改善主因爲降低取代型錯誤。在以 MAT-160 爲訓練語料的情況下，混合數=16 可較混合數=4 提升 3.8%的辨識率，幅度約爲 8%。

爲改進辨認正確率，本文針對(1) 強健型特徵參數萃取，以及 (2) 特徵參數調整法，作更進一步之探討。

五、 強健型特徵參數之萃取

5-1. 倒頻譜平均值正規化法（Cepstral Mean Normalization, CMN）

倒頻譜平均值正規化法是一種減低通道效應影響的方法，其原理如下：假設 x(m,n) 爲語者實際在第 m 個音框所發的第 n 點訊號值，y(m,n)爲 x(m,n)經過通道後的語音訊號值，則 x(m,n)與 y(m,n)的關係可表示如下：

$$y(m,n) = x(m,n) \otimes h(m,n) \qquad (1)$$

若假設在語者所念的一句語料內通道效應爲穩態，可移去 h(m,n)的音框引數，

成爲 h(n)。不同的通道，使 h(n)值產生差異，因此，以 y(m,n)訓練出來的模型將受通道效應的所產生的影響，若能移除 h(n)的影響，語音模型將更爲精確。以 X(m,k)、Y(m,k)及 H(k)分別代表 y(m,n)、x(m,n)及 h(n)在頻譜上的表現。

$$Y(m,k) = X(m,k) \cdot H(k) \tag{2}$$

取對數值之後，通道效應變成加法性。

$$\log Y(m,k) = \log X(m,k) + \log H(k) \tag{3}$$

等號兩邊各求平均值，可得

$$\frac{1}{M}\sum_{m=0}^{M-1} \log Y(m,k) = \frac{1}{M}\sum_{m=0}^{M-1} \log X(m,k) - \frac{1}{M}\sum_{m=0}^{M-1} \log H(k) \tag{4}$$

$$\frac{1}{M}\sum_{m=0}^{M-1} \log Y(m,k) = \frac{1}{M}\sum_{m=0}^{M-1} \log X(m,k) - \log H(k) \tag{5}$$

原對數值減其平均值，通道效應即被消除。

$$\log Y(m,k) - \frac{1}{M}\sum_{m=0}^{M-1} \log Y(m,k) = \log X(m,k) - \frac{1}{M}\sum_{m=0}^{M-1} \log X(m,k) \tag{6}$$

轉換成倒頻譜(Cepstrum)，即等於倒頻譜值減其平均值，此法稱爲倒頻譜平均值正規化(Cepstrum Mean Normalization, CMN)。

在基礎實驗所使用的 MEL 特徵參數，其中 12 階 MFCC 可視爲 log *Y(m,k)* 的線性組合，因此適用倒頻譜平均值正規化法。依據(6)式得出新的 12 階 MFCC 係數，再與其他的 12 維度未變動的特徵向量構成新的 26 階特徵參數，在接下來的敘述中，此法得到之語音參數標示爲 MEL-CMN。

5-2. 自相關係數頻域正規化法（DFT-MN-AUTO）

自相關係數頻域正規化法爲自相關係數微分法（RAS）[5,6]的變形，兩者的差異在：RAS 法爲針對消除背景雜訊而設計，DFT-MN-AUTO 法爲針對消除通道效應而設計。

若是考慮通道效應，假設 x(m,n)爲語者實際在第 m 個音框所發的第 n 點訊號值，y(m,n)爲 x(m,n)經過通道後的語音訊號值，則 x(m,n)與 y(m,n)的關係可表示如下：

$$y(m,n) = x(m,n) \otimes h(m,n) \tag{7}$$

若假設在語者所念的一句語料內通道效應爲穩態，可移去 h(m,n)的音框引數成爲 h(n)。

$$y(m,n)=x(m,n)\otimes h(n) \tag{8}$$

在自相關域上，可表示成：

$$r_{yy}(m,k)=r_{xx}(m,k)\otimes h(-k)\otimes h(k) \quad ,0\le m\le M-1,0\le k\le 2N-1 \tag{9}$$

$$r_{yy}(m,k)=\begin{cases}\sum_{j=0}^{N-1-k} y(m,j)y(m,j+k) & ,0\le k\le N-1\\ 0 & ,k=N\\ r_{yy}(m,2N-k) & ,N+1\le k\le 2N-1\end{cases} \tag{10}$$

將第 m 個音框之自相關係數 $r_{yy}(m,k)$取 2N 點作 DFT 演算，即變成功率頻譜：

$$S_{r_{yy}}(m,f)=S_{r_{xx}}(m,f)\cdot|H(f)|^2 \tag{11}$$

取對數值之後，

$$\log S_{r_{yy}}(m,f)=\log S_{r_{xx}}(m,f)+2\log|H(f)| \tag{12}$$

等號兩邊各求平均值，可得

$$\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{yy}}(m,f)=\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{xx}}(m,f)+\frac{2}{M}\sum_{m=0}^{M-1}\log|H(f)|$$

$$\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{yy}}(m,f)=\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{xx}}(m,f)+2\log|H(f)| \tag{13}$$

原對數值減其平均值，通道效應即被消除。

$$\log S_{r_{yy}}(m,f)-\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{yy}}(m,f)=\log S_{r_{xx}}(m,f)-\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{xx}}(m,f) \tag{14}$$

據此可得調整後的自相關係數，其求法如下：

$$\overline{r_{yy}}(m,k)=InverseDFT\left\{\exp\left(\log S_{r_{yy}}(m,f)-\frac{1}{M}\sum_{m=0}^{M-1}\log S_{r_{yy}}(m,f)\right)\right\}$$

$$,0\le m\le M-1,0\le k\le 2N-1,0\le f\le 2N-1 \tag{15}$$

由(15)式得出經由頻域正規化調整的自相關係數，因自相關係數串之對稱特性，取前 N 點結果作爲最終訊號串輸出 $r_{yy}^{+}(m,k)$。

以 $\overline{r_{yy}^{+}(m,k)}$取代原始的語音波形訊號，求取 MFCC 特徵參數並訓練相關模型；值得注意的是，當以前述 MEL 法求取 MFCC 時，參數 c_0 通常是捨棄不用的，因其代表的是訊號的強弱，對於語音特徵來說並不具鑑別度。但以自相關係數所求出的 c_0 包含其它的語音訊息，故在接下來的實驗裡，嘗試兩種作法，一爲保留 c_0（取係數 c_0~c_{11}），標示爲 DFT-MN-AUTO-C0；另一作法如同 MEL，捨棄 c_0，取 c_1~c_{12}爲 MFCC，標示爲 DFT-MN-AUTO-C1。如圖 1 所示，語音波形訊號自相關係數之求取，我們採用一快速演算法：

- 將波形訊號經由 DFT 轉換至頻域。
- 計算頻域訊號的能量參數。
- 將能量參數經由 IDFT 轉回時域，即可得自相關係數，自相關係數的長度爲 2N 點。

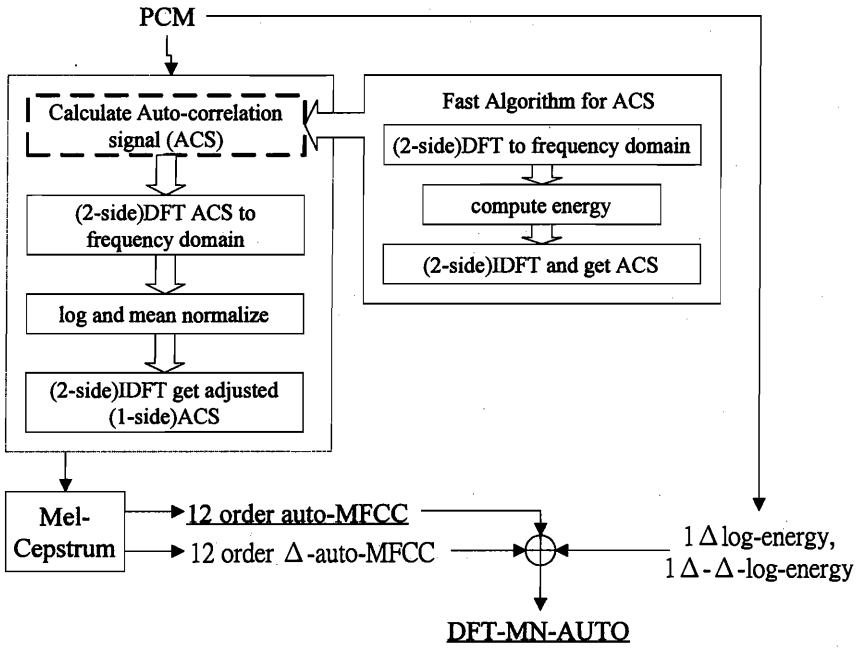


圖 1：DFT-MN-AUTO 特徵參數求取流程

5-3. 強健型特徵參數實驗

(1) 測試語料與訓練語料環境匹配實驗

以 MAT-160 爲訓練語料，並分別以 MEL-CMN、DFT-MN-AUTO-C0 及 DFT-MN-AUTO-C1 爲特徵參數，測試結果如表 6 所示。

特徵參數	MEL-CMN			DFT-MN-AUTO-C0			DFT-MN-AUTO-C1		
	混合數			混合數			混合數		
	4	8	16	4	8	16	4	8	16
插入率 (%)	2.05	2.36	1.84	1.71	2.17	2.13	3.25	3.29	2.93
刪除率 (%)	2.05	1.92	1.71	1.96	1.79	1.56	2.07	1.73	1.48
取代率 (%)	43.96	41.89	41.28	46.26	44.45	44.13	46.18	44.89	43.77
辨認率 (%)	51.94	53.82	55.17	50.06	51.58	52.17	48.50	50.08	51.82

表 6：不同特徵參數之測試結果

由實驗可看出 MEL-CMN 特徵參數求取法的表現最佳，DFT-MN-AUTO-C0 其次，DFT-MN-AUTO-C1 最末。但與表 5 相對照，皆可獲一定程度之提升。三種強健型特徵參數中，DFT-MN-AUTO-C1 的插入型錯誤比率較 MEL-CMN 與 DFT-MN-AUTO-C0 高，因此，若搭配較精準的切音方法，應可得更佳的表現。

(2) 去除波形訊號偏移量之延伸實驗

觀察了訓練語料庫與測試語料庫波形訊號的情形，發現錄製方式不同的關係，波形訊號的平均值會產生不同的偏移特性，其統計結果如圖 2 及圖 3 所示。

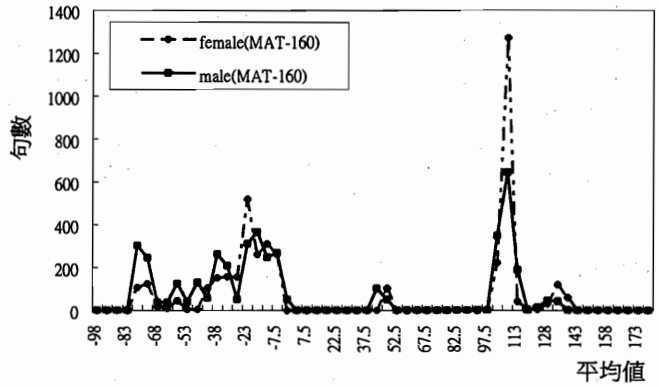


圖 2：MAT-160 波形訊號偏移特性

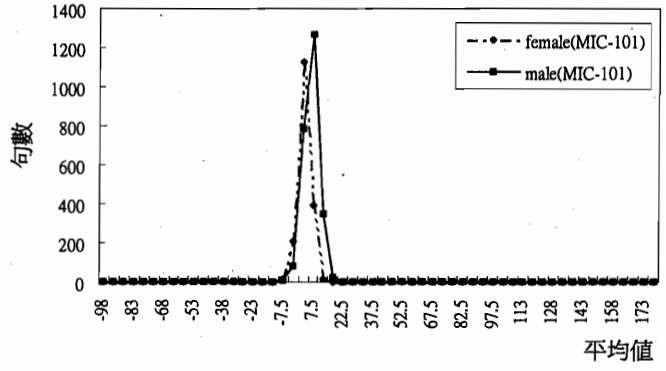


圖 3：MIC-101 波形訊號偏移特性

根據聲波的物理特性，一句聲音的波形訊號平均值應該在零附近，一如圖 3 所示透過麥克風錄製乾淨語音的統計結果。但我們發現 MAT-160 語料的波形訊號偏移特性並非如此，其可能原因爲：MAT（臺灣地區國語語音資料庫）語料是經由電話線連接到國內各大學語音實驗室進行錄音工作。各大學所使用的錄音界面是透過個人電腦上的 Dialogic Card 取樣並儲存的，若此卡的 DC 值略有偏移，則該大學所錄製的語音皆受此偏移量影響。MAT-160 語料庫便是從 MAT 計劃所錄製的聲音中整理與挑選出較佳之 160 人而成，因此，包含了數所學校錄音的結果。對照圖 2，可發現 MAT-160 的波形訊號偏移有數個鋒值，恰可符合上述所做的假設。所以，一個多處地點收音的語料庫應就系統間校準的問題多加注意。

爲了消除錄製系統間的差異對語料所造成的影響，在進行模型訓練及辨認之前，以句爲單位，對每句語料扣減其波形訊號的平均值，得出新的語料，此法標示爲 PCM-MN。以 PCM-MN 法處理語音訊號之後，再求特徵參數，其實驗結果如下，

特徵參數	MEL-CMN			DFT-MN-AUTO-C0			DFT-MN-AUTO-C1		
	混合數			混合數			混合數		
	4	8	16	4	8	16	4	8	16
插入率 (%)	2.68	2.96	2.58	1.94	2.11	2.03	3.29	3.59	3.42
刪除率 (%)	2.09	1.79	1.52	2.01	1.75	1.56	2.13	1.73	1.58
取代率 (%)	44.53	43.07	41.81	45.16	43.52	42.93	45.02	43.86	42.48
辨認率 (%)	50.70	52.17	54.10	50.89	52.62	53.48	49.56	50.82	52.51

表 7：不同特徵參數，採用 PCM-MN 法之測試結果

由實驗可看出， PCM –MN 法對於使用自相關係數計算特徵參數的方法（DFT-MN-AUTO-C0 和 DFT-MN-AUTO-C1）具正面的效益，在混合數等於 4 與 16 的情形下，DFT-MN-AUTO-C0 有 0.83%與 1.17%的提升，DFT-MN-AUTO-C1 有 1.06%與 0.49%的提升。其原因是具 DC 偏移量的波形訊號對於自相關係數的影響是全面的，若假設受到錄音設備偏移的影響爲 $y'(m,n)=y(m,n)+bias$，在自相關域則爲 $r_{yy}'(m,k)=r_{yy}(m,k)+ bias^2+bias*\Sigma[y(m,n)+y(m,n+k)]$，最末一項造成的雜訊不僅不能從之後的方法消除，而且因與訊號相乘關係，影響值不小。當進行 PCM-MN 時，偏移值造成的影響便大幅被壓抑，故可提升辨識率結果。

然而，對 MEL-CMN 特徵參數而言，則具負面影響，在混合數等於 4 與 16 的情

形下，辨識率下降了 1.24%與 1.07%。波形訊號偏移量對頻域的影響僅在頻率爲零處，其餘頻域皆不受干擾，故對 MFCC 影響本就不大。當額外加上 PCM-MN 時，事實上強迫所有波形訊號的平均值皆爲零，對照圖 3 麥克風錄音的情況，可知波形訊號的平均值實際並非皆爲零，而是在零附近呈現一窄分佈情況，加上扣減平均值的過程中忽略了偏移值時變的可能性，故 PCM-MN 亦造成外加的雜訊來源。對 MEL-CMN 而言，因其必須收集所有音框之 MFCC 求其平均，再扣減回每一音框，故某些音框若遭較強干擾，會透過 CMN 的過程連帶影響其他音框，進而降低了整句語料之特徵參數可靠性。

(3) 訓練語料與測試語料環境不匹配實驗

將訓練語料換成 MIC-101，測試訓練語料與測試語料在環境不匹配情形下，各種強健型特徵參數的表現如下。

特徵參數	MEL-CMN			DFT-MN-AUTO-C0			DFT-MN-AUTO-C1		
	混合數			混合數			混合數		
	4	8	16	4	8	16	4	8	16
插入率 (%)	3.40	3.25	3.04	3.38	3.25	3.23	5.47	5.41	5.30
刪除率 (%)	2.98	3.00	2.81	3.12	2.96	2.68	2.72	2.66	2.70
取代率 (%)	64.29	63.66	63.05	66.79	66.91	66.49	66.89	65.60	64.76
辨認率 (%)	29.33	30.09	31.10	26.71	26.88	27.51	24.92	26.33	27.24

表 8：不同特徵參數之測試結果

實驗中，MEL 特徵參數在環境不匹配的情形下，混合數 4 與 16 的辨識率爲 10.98%與 12.16%，以此對照，使用強健型特徵參數 MEL-CMN 與 DFT-MN-AUTO 皆可獲得改善，其中又以 MEL-CMN 的效果最佳。在混合數爲 4 與 16 的條件下，MEL-CMN 可提升辨識率 18.35%與 18.94%，DFT-MN-AUTO-C0 可提升 15.73%與 15.35%，DFT-MN-AUTO-C1 可提升 13.94%與 15.08%。

(4) 外加雜訊實驗

除了通道效應外，背影雜訊亦是造成辨識率下降的主因。因所使用的測試語料 TEST-500 可視爲較乾淨的電話線語音，爲模擬雜訊情形，對語音訊號加上不同強度的白雜訊（WHT），其實驗結果如下：

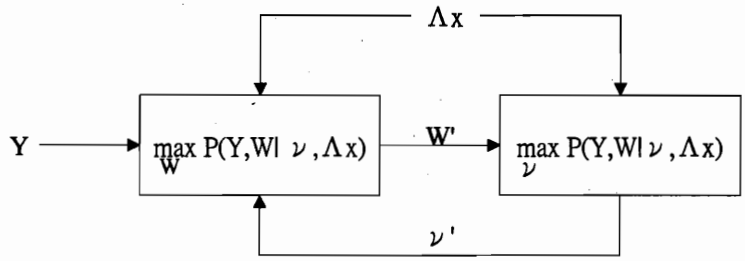
特徵參數	MEL-CMN			DFT-MN-AUTO-C0			DFT-MN-AUTO-C1		
	混合數			混合數			混合數		
	4	8	16	4	8	16	4	8	16
無外加雜訊	51.94	53.82	55.17	50.89	52.65	53.48	49.56	50.82	52.51
SNR= 20 dB	37.36	40.87	42.14	38.86	40.98	42.24	39.07	41.74	42.27
SNR= 10 dB	17.33	19.10	21.37	22.70	25.02	27.47	25.21	27.13	28.28
SNR= 0 dB	1.44	2.26	4.08	4.88	6.07	6.57	8.24	7.40	8.79

表 9：不同特徵參數，有外加雜訊時之辨認結果

由實驗的結果可發現，DFT-MN-AUTO 法對抗雜訊的能力較強。當信號雜訊比大於 20dB 時，採用自相關係數便可獲得好處。原因是在求取自相關係數的過程中，訊號加乘可增大其訊號雜訊比值。其中 DFT-MN-AUTO-C1 的效果略好於 DFT-MN-AUTO-C0，應與所加入的訊號爲白雜訊有關，自相關域上白雜訊會影響低頻部份，而 DFT-MN-AUTO-C1 的求取正好可避開這部份的影響。。

六、 特徵參數調整法

特徵參數調整法的主要概念爲：在特徵參數域中估測與參考模型之間的失真偏移量，據之調整觀測序列特徵參數，以期符合訓練環境情狀。圖 4 爲示意圖[9]，其中 Y 代表測試語料特徵參數序列，W'代表辨識後的音節字串（包含錯誤可能），Λ_x 代表參考模型，ν 爲特徵參數調整參數。



$$(\nu', W') = \underset{(\nu, W)}{\operatorname{argmax}}\, p(Y, W|\nu, \Lambda x)$$

圖 4：特徵參數調整法示意圖

6-1. 訊號偏移消去法（Signal Bias Removal, SBR）

訊號偏移消去法[7]爲一種估測測試環境與模型間頻道偏移量，再對原特徵向量扣除偏移量的方法，以消除測試與訓練環境間不同頻道效應的影響，簡介作法如下：

假設$Y=\{y_t\}$爲測試語音觀測序列，$X=\{x_t\}$爲符合訓練環境的語音觀測序列，兩者關係如下：

$$y_t = x_t + \bar{b} \tag{16}$$

偏差量$\bar{b}$以下列式子估出：

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu}_{s(t)}) \tag{17}$$

其中$\hat{\mu}_{s(t)}$代表第 t 個音框中，相對於訓練碼本有最大觀測機率之狀態平均值。

經由多次遞迴，SBR 可估出更準確的偏差量，在第 n 次遞迴時匹配訓練環境的觀測序列爲：

$$\tilde{X} = \{y_t - (\bar{b}^n + \bar{b}^{n-1} + ... + \bar{b}^0)\} \tag{18}$$

6-2. 層次式訊號偏移消去法（Hierarchical Signal Bias Removal, HSBR）

層次式訊號偏移消去法[8]亦是一種在特徵參數域進行調整的強健性補償方法，與 SBR 不同的是：不是對整句測試語料估出一個可能的頻道偏移量，而是對每一個音框求取音框相關（frame dependent）的偏移量。基本假設爲在一段語句中所受到的頻道偏差不全是穩態的，不同音段內的頻道情形應是變化不定的，因此針對不同音框採用不同的偏移量來模擬頻道失真。

爲估出「音框相關偏差」（frame-dependent bias），假設$Y=\{y_t\}$爲測試語音觀測序列，$X=\{x_t\}$爲符合訓練環境的語音觀測序列，$B=\{\bar{b}_t\}$則爲要估測的音框相關偏差序列，三者關係如下：

$$y_t = x_t + \bar{b}_t \tag{19}$$

求取$\bar{b}_t$之前，先將測試語料與參考模型碼本相較較，找出每個音框所屬的叢集。假設音框被分散到 M 個叢集中，第 i 個叢集偏移量定義成：

$$b_i^c = \frac{1}{T_i}\sum_{l=1}^{T_i}(y_{t_i(l)} - \hat{\mu}_i), \qquad 1 \le i \le \mathrm{M} \tag{20}$$

其中$y_{t_i(l)}$表示測試語句中屬於叢集 i 的音框，T_i爲屬於叢集 i 的音框數目，而$\hat{\mu}_i$則

爲叢集 i 的平均值向量。音框相關偏移量可經由各個叢集偏差 b^c 的線性組合得出：

$$\overline{b_t} = \frac{\sum_{i=1}^{M} b_i^c w_{t(i)}}{\sum_{j=1}^{M} w_{t(i)}} \tag{21}$$

其中叢集權重 $w_{t(i)}$ 是每個音框分別對叢集 i 的平均值做加權歐氏距離（Weighted Euclidean distance）的倒數。最後估出匹配訓練環境的觀測序列：

$$\tilde{X} = \{(y_t - \bar{b}_t)\} \tag{22}$$

此法也和 SBR 一樣，可遞迴地減少環境間的差異性。

6-3. 統計式對應法（Stochastic Matching, SM）

統計式對應法[9,10,11]是一種基於最大機率估測（ML estimation），來求取測試環境與訓練環境特徵參數域失真的方法。因其亦爲音框獨立的偏移扣減方式，故基本假設如訊號偏移消除法，所不同的是，統計式對應法包含機率概念，除了用到參考模型的狀態平均值外，也考慮變異數的影響：

假設 $Y = \{y_t\}$ 爲測試語音觀測序列，$X = \{x_t\}$ 爲符合訓練環境的語音觀測序列，兩者關係如下：

$$y_t = x_t + \bar{b} \tag{23}$$

偏差量 $\bar{b}$ 以下列式子估出：

$$\bar{b} = \frac{\sum_{t=1}^{T} \frac{(y_t - \hat{\mu}_{s(t)})}{\Sigma_{s(t)}}}{\sum_{t=1}^{T} \Sigma_{s(t)}} \tag{24}$$

其中 $\hat{\mu}_{s(t)}$ 和 $\Sigma_{s(t)}$ 代表第 t 個音框中，相對於訓練碼本有最大觀測機率之狀態平均值及變異數。同樣的，統計式對應法亦與前兩種方法一樣，可遞迴地減少環境間的差異性。

6-4. 特徵參數調整法實驗

特徵參數調整法可分別對測試端與訓練端進行。

在測試端的做法爲：

1. 每句測試語料以一階動態規劃演算法得出最佳狀態序列；

2. 利用步驟 1 所得序列進行特徵參數調整；

3. 重複進行步驟 1 與步驟 2，可得不同遞迴數的辨識結果。

在訓練端的做法爲：

1. 對每句訓練語料以維特比演算法得出最佳狀態序列；

2. 利用步驟 1 所得序列進行特徵參數調整；

3. 對新的特徵參數進行叢集，得出調整模型；

4. 重複進行步驟 1 到步驟 3，可得不同遞迴數的調整模型。

在實驗中，分別對於 (1)測試端作特徵參數調整，以及(2) 訓練端與測試端的語料皆作特徵參數調整，進行測試。所使用的訓練語料皆爲 MAT-160，以 MEL 特徵參數作測試，使用混合數 16 的女男分類模型，其未作特徵參數調整時之辨認率爲 48.75%，作特徵參數調整時之結果在表 10。

		遞迴數				
		1	2	3	4	5
測試端作特徵參數調整	SBR	53.00	53.59	53.84	53.95	54.05
	HSBR	52.85	53.44	53.80	53.97	54.01
	SM	52.72	53.53	53.86	53.82	53.95
訓練端與測試端皆作特徵參數調整	SBR	54.27	54.65	54.52	54.43	54.46
	HSBR	53.80	54.62	54.65	54.77	55.03
	SM	53.36	54.52	54.48	54.50	54.54

表 10：特徵參數調整之測試結果

由實驗的結果可看出，只在測試端進行特徵參數調整，則不論是特徵參數扣減法、階層式特徵參數扣減法或統計式對應法，所獲致的改善成果都差不多，在遞迴數爲 5 的情形下，經由特徵參數調整可提升整體辨識率 5.20% ~ 5.30%，提升的幅度約爲 10.50%。

訓練端與測試端皆作特徵參數調整時，訓練端的模型是經由 5 次遞迴特徵參數調整後，進行叢集訓練而得出。比較實驗的結果，可發現，在訓練端亦進行模型參數的調整對於辨識率有 0.42% ~ 1.02%的提升。這是因爲在訓練端進行特徵參數調整時，訓練語料透過扣減偏移量的計算會使得模型模糊的程度下降（即變異數值下降），而使模型更加強健。這點可由階層式特徵參數扣減法提升最多可看出來，該法在求取特徵參數偏移量時與參考模型的精確度關連性最大。

結果顯示三種方法中，階層式特徵參數扣減法的辨識效果最好，這是因爲此法多加考慮了特徵參數偏移量的時變性，其偏移值計算是音框相關的，而特徵參數扣減法和統計式對應法的偏移值則是音框獨立的，辨認率結果也顯示沒有很大的差別。

6-5. 叢集模型建構同類參考碼本

在上述實驗中，每一音框的特徵參數是相比於最佳解碼序列中對應之狀態混合數，但是，相對應的狀態混合數並非皆是正確值，根據觀察辨識結果，發現造成辨識錯誤大部份的原因皆爲聲母或韻母的取代型錯誤，並且大半落在混淆音的叢集範圍內，故若對可能的混淆狀態混合數進行特徵參數調整的運算，對於辨識率應有所助益。因此，希望以叢集模型[12]的架構，替代原先的對應狀態混合順列。

叢集模型的概念是：對於參考模型內所有狀態的所有混合數進行叢集分類。叢集的結果可得：1.叢集模型，由所有落在該叢集的狀態混合數以線性方式合併得出；2.各狀態混合數的叢集映射表。

在測試端，計算特徵參數偏移量時，每一個音框相對應之狀態混合數以其映射叢集模型替代，辨認流程如圖 5：

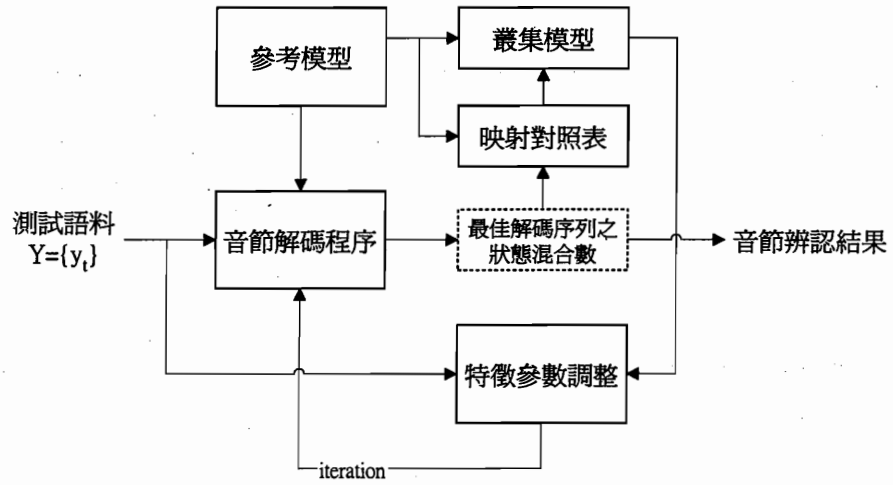


圖 5：叢集模型辨認的基本架構

在圖 5 中，匯入特徵參數調整方塊的模型參考值是透過一映射對照表，選取最相

近的叢集模型匯入特徵參數調整方塊作爲參考模型，進行特徵參數調整運算。此法標示爲 VQModel。一個混合數 16 的模型，所有狀態的所有混合數總共約有 4500 個，接下來的實驗中，將之叢集成 256 個碼本，每個碼本約可分到 10~40 個混合數。所使用的訓練語料爲 MAT-160，以 MEL 特徵參數作測試，使用混合數 16 的女男分類模型，其未作特徵參數調整時之辨認率爲 48.75%。此法先對只作測試端特徵參數調整之 SBR 方法進行測試，其結果並不理想，表 11 所示之辨認率沒有改進。

		遞迴數				
		1	2	3	4	5
只作測試端特徵參數調整(SBR)	未使用叢集模型	53.00	53.59	53.84	53.95	54.05
	使用叢集模型	53.46	53.86	53.91	53.82	53.84

表 11：使用叢集模型之實驗結果

這是因爲一個叢集模型的參數值爲落入此叢集中的所有狀態混合數平均而得，故若某測試音框對應的最佳狀態混合單位若映射到叢集的邊緣，在計算時會硬是將它拉到中心的位置，而造成大的誤差。然而，比對其音節辨認輸出時可發現，叢集模型的方式可以補償一些基本 SBR 法無法改善的部份。

爲了改善因叢集造成模型值的偏移，進而影響補償效果，因此需對叢集模型辨認的基本架構進行改良。所選取的改良作法爲：對於同一叢集的所有狀態混合數求取音框對應偏移值，再以相似機率值（likelihood probability）進行加權；因此一個音框的偏移量是經由其最佳解碼序列所落入叢集內的所有狀態混合數，計算距離並加權相加而產生的。這個方法捨棄了叢集參考模型，僅使用了映射對照表。此法標誌爲 WVQModel，實驗結果如表 12。

		遞迴數				
		1	2	3	4	5
測試端作特徵參數調整	SBR	53.12	53.82	54.14	54.16	54.22
	HSBR	53.23	53.59	54.05	54.20	54.18
	SM	52.83	53.67	53.70	53.89	54.08
訓練端與測試端皆作特徵參數調整	SBR	54.22	54.65	54.67	54.60	54.60
	HSBR	55.29	55.11	55.19	55.22	55.17
	SM	53.32	54.24	54.69	54.73	54.73

表 12：使用改良叢集模型之實驗結果

比較實驗結果，加權式模型叢集的方法皆可獲致比直接序列方法較佳的辨認率。

七、 結論

在本文中，討論了兩種不同的強健性策略：強健型特徵參數萃取及特徵參數調整法。所採用的三種的強健型特徵參數萃取方法中， MEL-CMN 對抗通道效應的能力最好，DFT-MN-AUTO-C1 對抗雜訊的能力最佳。在訓練環境與測試環境皆爲電話網路的情況下，MEL-CMN 可得最佳辨識率 55.17%。在測試的三種特徵參數調整方法中，若只在測試端單方作特徵參數調整，這三種方法的效果都差不多，辨識率約可提升 5.25%。若是在訓練與測試雙方作特徵參數調整，可再提升辨識率 0.50%~1.00%。另外，提出叢集模型概念及其改良「加權式叢集模型法」，此法可再提升辨識率約 0.20%。綜上所述，以階層式特徵參數扣減法，在訓練與測試雙方作特徵參數調整，並使用加權式叢集模型運算，可得最佳辨識率 55.29%。

參考文獻

[1] 簡仁宗, “電話環境下語音辨認之研究”，國立清華大學電機工程研究所博士論文，民國八十六年六月

[2] 邱榮樑，“經電話網路之連續語音辨認及辨認技術評比方法之建立”，國立清華大學電機工程研究所碩士論文，民國八十七年六月

[3] 涂英傑，“電話環境下國語語音辨認之強健性問題”，國立臺灣大學電機工程研究所碩士論文，民國八十七年六月

[4] 謝華君，“電話網路上國語連續音節辨認的初步研究”，國立清華大學電機工程研究所碩士論文，民國八十六年六月

[5] You, Kuo-Hwei and Hsiao-Chuan Wang, “Robust Features Derived from Temporal Trajectory Filtering for Speech Recognition under the Corruption of Additive and Convolutional Noises”, Proceedings ICASSP, 1998, pp. 577-580.

[6] You, Kuo-Hwei and Hsiao-Chuan Wang, “Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences”, Speech Communication vol. 28, no. 1, pp. 13-24, 1999.

[7] Rahim, Mazin G. and Biing-Hwang Juang, “Signal Bias Removal by Maximum Likelihood

Estimation for Robust Telephone Speech Recognition", IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, pp. 19-30, 1996.

[8] Rahim, Mazin G., Biing-Hwang Juang, Wu Chou and Buhrke E., "Signal conditioning techniques for robust speech recognition", IEEE Signal Processing Letters, Vol. 3, No. 4, pp. 107-109, 1996.

[9] Sankar, Ananth and Chin-Hui Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 3, pp. 190-202, 1996.

[10] Siohan, Olivier and Chin-Hui Lee, "Iterative Noise and Channel Estimation Under the Stochastic Matching Algorithm Framework", IEEE Signal Processing Letters, Vol. 4, NO. 11, pp. 304-306, 1997.

[11] Chien, Jen-Tzung, Hsiao-Chuan Wang and Lee-Min Lee, "Estimation of Channel Bias for Telephone Speech Recognition", Proceedings ICSLP 1996, pp. 1840-1843.

[12] Lawrence, Craig and Mazin Rahim, "Integrated Bias Removal Techniques for Robust Speech Recognition", Proceedings, EuroSpeech 1997, pp. 2567-2570.

[13] Wang, H. C., "MAT- A project to collect Mandarin speech data through telephone networks in Taiwan," Computational Linguistics and Chinese Language Processing, vol. 2, no. 1, pp. 73-90, 1997.

[14] Wang, H. C., "Speech research infra-structure in Taiwan – from database to performance assessment," Proceedings, 1999 Oriental Cocosda Workshop, 1999, pp. 53-56.