音框同步之雜訊補償方法在汽車語音辨識之應用

Frame Synchronous Noise Compensation for Car Speech Recognition

簡仁宗　林敏順

國立成功大學資訊工程學系

Email : jtchien@mail.ncku.edu.tw

摘要

自動語音辨識(ASR)系統應用在雜訊環境下時，由於雜訊語音與模型參數間的不匹配，將導致辨識率明顯的下降。音框同步補償方法可從測試語音中以音框為單位作補償，每個音框先計算出語音與模型參數之線性等化因子，再根據等化因子的大小將模型參數的平均值調整量對應出來，自動找出語音隱藏式馬可夫模型之參數調整量，將模型參數調整後再作辨識。本論文提出一種強健性的方法做平均值調整函式的估測，從實驗結果得知，使用本方法可有效提升在汽車噪音環境下語音辨識的正確率。在九十公里和五十公里汽車環境下免持麥克風之詞彙辨識系統都有明顯的改善。

1. 簡介

語音是人與電腦間最自然的溝通方式，要讓電腦聽的懂人講話的聲音一直是人類努力的目標，此目標達成與否要視電腦語音辨識技術的開發成熟度而定，雖然傳統錄音間所錄得的語音已經可以達到很高的辨識率，但真正實用之語音辨識系統其應用之場所一定有程度不等的噪音存在[6][7]，若此辨識系統不做任何調整，測試語音與模型參數間的不匹配將導致辨識率明顯的下降。因此我們以汽車環境下免持聽筒之語音辨識噪音補償來進行研究。

一般而言，不同汽車上的噪音大小不同，要用來訓練(training)噪音語音模組之語料庫將會非常龐大且不易取得是不切實際的方法；另一方面汽車環境是屬於高雜訊的地方，例如、引擎輪胎轉動引起的噪音、風嘯聲、收音機或音響的噪音、汽車內說話的回音等等，所以在

測試(testing)時與語音模組不吻合的情形會非常嚴重，這些都會使語音辨識技術更加困難，我們所提的方法就是要解決上面的問題。我們以安靜房間錄下的語料庫訓練出一組語音模組，再以實際汽車裡錄得的少數語料訓練出平均值與變異數調整函式，當測試語料在測試時會根據噪音程度自動對應出平均值及變異數的調整量作補償。

我們所提出的方法是依據在汽車環境下免持麥克風語音辨識之噪音補償方法[1]做改良，這個方法在做估測模型參數調整函式時需要將噪音語料做維特比(Viterbi)切音，由於噪音語料庫經Viterbi切音之後不準確，如再依據等化因子估測調整函式效果不佳，為了改善這個缺點我們以新的方法來估測調整函式。本補償方法是可以音框同步的，主要是當聲音錄得的同時不需要等所有語音資料都收集好再進行辨識，可以以一個音框為單位，錄得一個音框後就可直接計算等化因子，再根據等化因子對應出平均值調整量，將乾淨語音模組依不同環境作不同之調整，在經過實際測試後的確可有效地改善汽車環境下的語音辨識率。

2. 噪音補償方法

這裡的噪音補償方法的理論基礎是從美國喬治亞理工學院學者Carlson 和 Clements 於1994年提出的“以投射為主之相似度量測 (projection-based likelihood measure)” [3][4]方法所延伸出來的，根據1989年AT&T 貝爾實驗室 Mansour 和 Juang [9]的觀察結果發現，任何乾淨語音的倒頻譜向量受到白色雜訊(white noise)的干擾，其向量的大小值會縮小且相位大致不變的特性，投射為主之相似度量測就是根據這個觀察結果所發展出來的抗噪音辨識演算法。但在實際汽車環境下的噪音並非白色雜訊，因此我們導入一個等化因子(equalization factor) λ_e，我們將根據等化因子訓練出平均值調整函式，再根據等化因子大小自動對應出其平均值調整量，再語音辨識時結合進去。以下為等化因子求法和相似度量測時結合平均值補償量之算法。

2-1 等化因子

一個觀測樣本$\mathbf{c}_t$與乾淨語音所訓練出來的隱藏式馬可夫模型$\Lambda_{s,m}=(\mu_{s,m},\Sigma_{s,m})$ 做相似度

量測時(其中s為狀態的引數，m為混合數的引數)，我們通常都用高斯機率密度 $P(\mathbf{c}_t|\Lambda_{s,m}) = N(\mathbf{c}_t;\mu_{s,m},\Sigma_{s,m})$ 來量測，然而，在噪音的干擾下，此相似度量測的平均值向量 $\mu_{s,m}$ 部份應該自動匹配掉雜訊語音，所以在平均值向量乘上一個線性調整因子 λ 形成以下的相似度量測：

$$P(\mathbf{c}_t|\lambda,\Lambda_{s,m}) = N(\mathbf{c}_t;\lambda\mu_{s,m},\Sigma_{s,m}) = (2\pi)^{-N/2}\left|\Sigma_{s,m}\right|^{-1/2}\cdot\exp\left(-\frac{1}{2}(\mathbf{c}_t-\lambda\mu_{s,m})^T\Sigma_{s,m}^{-1}(\mathbf{c}_t-\lambda\mu_{s,m})\right) \quad (1)$$

其中 N 是向量的維度 $\mu_{s,m},\Sigma_{s,m}$ 是隱藏式馬可夫模型的平均值及變異數。

應用最佳相似度 **(maximum likelihood, ML)** 法則，可以推導出最佳等化因子λ_e如下所示:

$$\lambda_e = \arg\max_{\lambda}\log P(\mathbf{c}_t|\lambda,\Lambda_{s,m}) = \frac{\mathbf{c}_t^T\Sigma_{s,m}^{-1}\mu_{s,m}}{\mu_{s,m}^T\Sigma_{s,m}^{-1}\mu_{s,m}} \quad (2)$$

此λ_e是 $\mathbf{c}_t$ 在 $\mu_{s,m}$ 上之投射量，為隱藏式馬可夫模型引數s和m的函式，及觀察樣本$\mathbf{c}_t$的相關的函式。將式(2)的λ_e代回式(1)即為投射為主之相似度量測的計算方法。

2-2 相似度量測之參數調整

由於平均值部份用簡單的等化因子做調整會產生程度不等的偏差，我們認為平均值向量的偏差 $\mathbf{b} = \mathbf{c}_t - \lambda_e\mu_{s,m}$ 也應一併補償，平均值補償函式$\mathbf{b}(\lambda_e)$是與 λ_e 有關，我們對所有的$\mathbf{c}_t$與隱藏式馬可夫模組$\Lambda_{s,m}$會針對各λ_e值去統計平均值應對應的調整量$\mathbf{b}$，當語音辨識進行辨識之相似度量測時先計算出λ_e值後就可根據平均值補償函式$\mathbf{b}(\lambda_e)$自動對應出平均值調整量來作補償，以提高語音在噪音環境下的辨識率。式(3)為結合平均值補償函式後之相似度量測的計算方法：

$$P(c_t|\lambda_e,\Lambda_{s,m},b(\lambda_e)) = N(c_t;\lambda_e\mu_{s,m}+b(\lambda_e),\Sigma_{s,m}) =$$

$$(2\pi)^{-N/2}\left|\Sigma_{s,m}\right|^{-1/2}\exp\left(-\frac{1}{2}(c_t-\lambda_e\mu_{s,m}-b(\lambda_e))^T\Sigma_{s,m}^{-1}(c_t-\lambda_e\mu_{s,m}-b(\lambda_e))\right) \quad (3)$$

本方法又稱為平均值補償過相似度量測(Mean Compensated Likelihood Measure, MCLM) 。

我們將對所有的c_t與隱藏式馬可夫模組$\Lambda_{s,m}$針對各λ_e值去統計平均值的調整量，在我們的實驗中λ_e值是介於-2到4之間，每0.01為一個區間(section)共600個section，當語音辨識進行辨識之相似度量測時先計算出λ_e值後自動對應出平均值調整量作補償，以提高語音在噪音環境下的辨識率。

3. 調整函式之估測

在原始平均值和變異數調整函式的估測過程中[5]，我們需要準備一組乾淨語料庫，訓練出乾淨的語音模組，以及一組以人工方式加上不同噪音型態和噪音分貝的噪音語料庫，但要在汽車環境下做自動語音辨識系統(ASR)，這樣的作法是不太實際，因為我們無法在汽車上錄下一組與實驗室同步的語音資料，這很困難而且有諸多限制，最好的方法就是實際準備一組在噪音環境錄下的少量語料庫用來訓練調整函式。對於調整函式的估測我們提出兩種研究方法並在實驗結果中列出其辨識的結果。

3-1 噪音語料庫

汽車噪音環境下的語料庫是實際在汽車行駛中錄下的語音資料；有時速零公里怠速路況、五十公里正常路況及九十公里高速路況三組；以免持聽筒遠距離麥克風(hands free far talking)方式錄音，錄音時引擎發動、關閉汽車音響、關閉車窗、冷氣開到最小，使用MDWalkman(MZ-R55)錄音設備，麥克風為高抗雜訊麥克風型號為(SONY ECM-717)，麥克風置於副駕駛坐前的置物箱上。語料庫內容以人名和命令為主；其中有人名100個、命令14個。以這樣的錄音環境下我們總共錄了1271句，由五男五女所錄下。

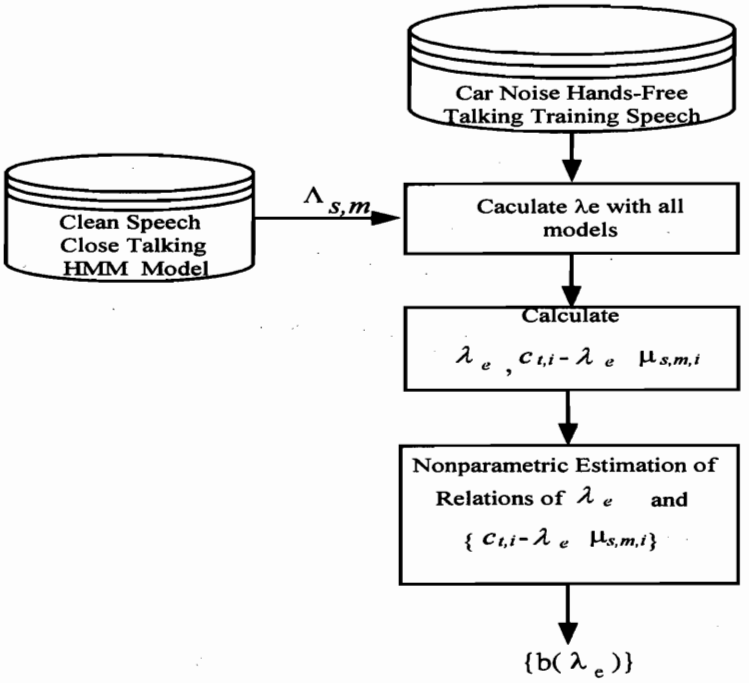
這1271句總共是由兩組語料庫所組成，這兩組語料分別是用來訓練補償函式用的訓練語料庫，和用來做測試用的測試語料庫。第一組是訓練補償函式用的語料庫，汽車是TOYOTA COROLLA 1.8分別由兩男兩女實際開車所錄下的語料總共有480句，零公里有122句、五十公里158句、九十公里200句。第二組是測試用語料庫，汽車是裕隆尖兵1.6分別由三男三女實際開車所錄下的語料總共有791句、零公里有204句、五十公里263句、九十公里324句。在估測補償函式方面我們又把第一組資料額外分出一些訓練語句，由原本兩男兩女再分出一男一女，分出的部份共有222句，零公里54句、五十公里74句、九十公里94句，我們將觀察訓練語句的多寡與辨識率的關係。

3-2 調整函式的估測方法

由於與傳統相類似的方法[1]需要以維特比(Viterbi)切音之後再根據λ_e去統計補償函式，但是噪音語料庫經維特比(Viterbi)切音後會有誤差，雖然仍具有很重要的統計特性，但統計特性不可靠。為了改善這個缺點我們以新的方法來統計補償函式，仍是以實際在汽車環境錄下的噪音語料庫來訓練，共有兩組語料庫；兩男兩女及一男一女語料。圖一、為我們提出的平均值調整函式$b(\lambda_e)$估測方法。與傳統相類似的方法[1]做比較，不同的部份在於每一個訓練音框c_t我們會與所有的語音模組都計算λ_e，之後根據λ_e收集訓練音框，我們把維特比(Viterbi)這個切音程式取代掉了，其進行步驟描述如下：

(1)首先，對於輸入的雜訊語音，每一個音框c_t會與所有的語音模組$\Lambda_{s,m}$都計算λ_e。等化因子的值限定在-2到4之間。

(2)等化因子λ_e計算出來後，同時也將受雜訊影響的平均值調整量$c_{t,i}-\lambda_e\mu_{s,m,i}$統計出來，$i$ 表示特徵向量的維度引數。



圖一、平均值調整函式估測流程圖

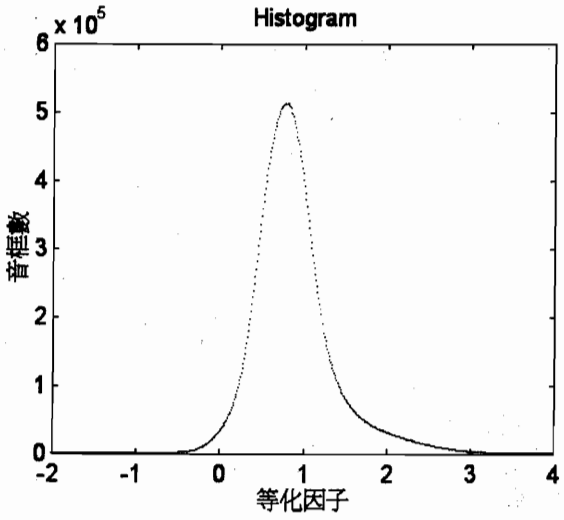
(3)從等化因子λ_e與調整量$\{c_{t,i}-\lambda_e\mu_{s,m,i}\}$的關係分佈圖(Scatter Diagram)中，估測出平均值補償函式$b(\lambda_e)$。

圖二為使用改善後估測調整函式方法的音框數與等化因子關係圖，橫軸為等化因子值介於-2 到 4 之間，縱軸為音框數出現的頻率，大部份的音框λ_e值介於 0.5 到 2 之間，也是形成一個近似常態分配的曲線。圖三為使用改善後估測調整函式的平均值調整量與等化因子在 LPC cepstrum 第一維下的關係圖，其中，橫軸為等化因子，縱軸為平均值對應的調整量。

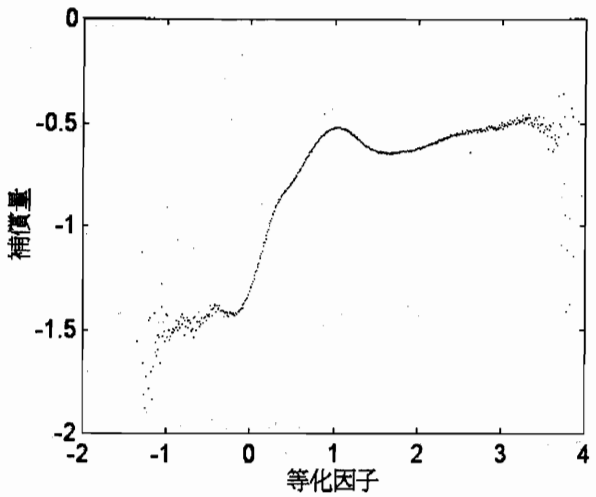
4. 實驗結果

4.1 調整參數的存取及辨識架構

語音辨識系統中，語音需先經過取樣及量化成為數位資料，實驗中所有語料的語音取樣頻率均為8kHz，以及以16bit表示每個數位點，我們可以從有效的語音取樣中抽取適當的語



圖二、等化因子之機率密度函式



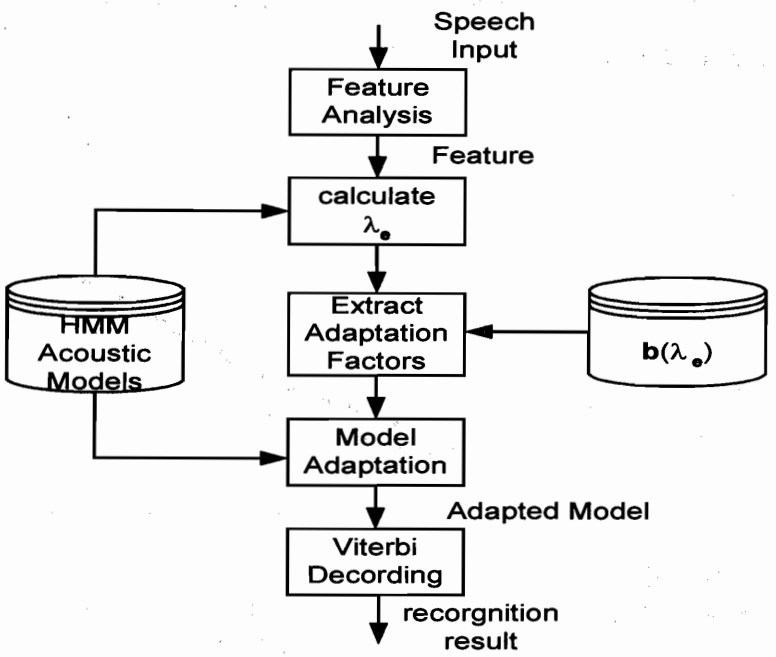
圖三、平均值調整函式在 LPC cepstrum 第 1 維下的分佈圖

音特徵值，我們的語音特徵每個音框的參數為12階 LPC cepstrum 和12階 delta LPC cepstrum 和1階 delta log energy 和1階 delta delta LOG Energy[8]。共26階。本研究方法在做補償函式分析時，會對每一階均建立一個查詢表(table)，我們的特徵參數有26維，所以平均值調整函式方面有26個查詢表，變異數調整函式方面也是26個查詢表。

訓練乾淨的隱藏式馬可夫模型參數(HMM)的語料庫(database)是以近距離麥克風之方式錄下的乾淨語料，這組語料庫共有5050句由50男及51女在安靜辦公室房間裡所錄下的，每一個人各唸二字詞、三字詞或四字詞共50句。使用的馬可夫模型參數有408個音節模組，以次音節前後相關的方式建構出467個狀態及一個背景雜訊狀態；每個狀態依實際的音框多寡分成不同數量的混合數，每個狀態至多有四個混合數。

將兩組調整函式儲存在記憶體中，於雜訊語音辨識時，就可依等化因子對應出平均值與變異數調整量，將乾淨語音模組依不同環境作不同之調整。圖四為噪音環境下的語音辨識架構；以下為整個架構的說明：

(1)當語音輸入求取參數後，由維特比解碼器(Viterbi Decoding)找出一條最佳路徑。在求取最佳路徑時，需先參考隱藏式馬可夫(HMM)語音模型的參數值依式子(2)計算出等化



圖四、系統架構流程圖

因子λ_e。

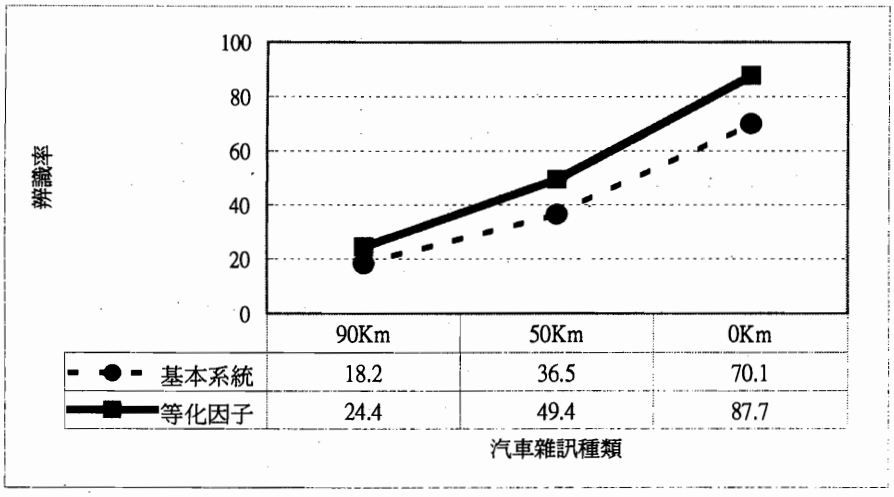
(2)根據λ_e可以由資料庫找出$b(\lambda_e)$這兩個參數；之後再進行相似度量測，此最佳路徑就是經過補償後的最佳路徑。

(3)與114個詞彙做樣型比對，找出一個最有可能的詞彙。

4-2 實驗結果

本實驗之測試語料庫包括零公里怠速路況共204個測試語句，五十公里正常路況下共263個測試語句，九十公里高速路況共有324個測試語句，它們都是由遠距離麥克風錄得的，所使用的汽車是裕隆尖兵 1.6，由三男三女所錄下的。實驗結果部份，我們列出不作補償及加入λ_e補償的辨識結果，其中90公里測試語料部份由原先18.2%增加到24.3%；50公里測試語料部份由原先36.5%增加到49.4%；0公里測試語料部份由原先70.0%增加到87.7%；我們可以發現辨識結果都有明顯的改善。圖五為基本系統及加入λ_e作補償辨識結果的比較圖。

實驗中用來訓練補償函式語料庫部份有兩組(兩男兩女及一男一女)訓練語料，從表一實驗結果中可以看出以兩男兩女的語料訓練調整函式會比一男一女的訓練語料有較好的辨識率，主要是因為較多的訓練語料可以訓練出較佳的調整函式。不管是90公里路況或是50公里路況，都可以看出這個現象。而且我們發現以改良後估測調整函式會比以原始的估測方法[1]的辨識率要好。我們以改良後估測調整函式的方法對這兩組訓練語料所得的比較結果如圖六所示。其中等化因子的範圍都是訂在-2到4之間。

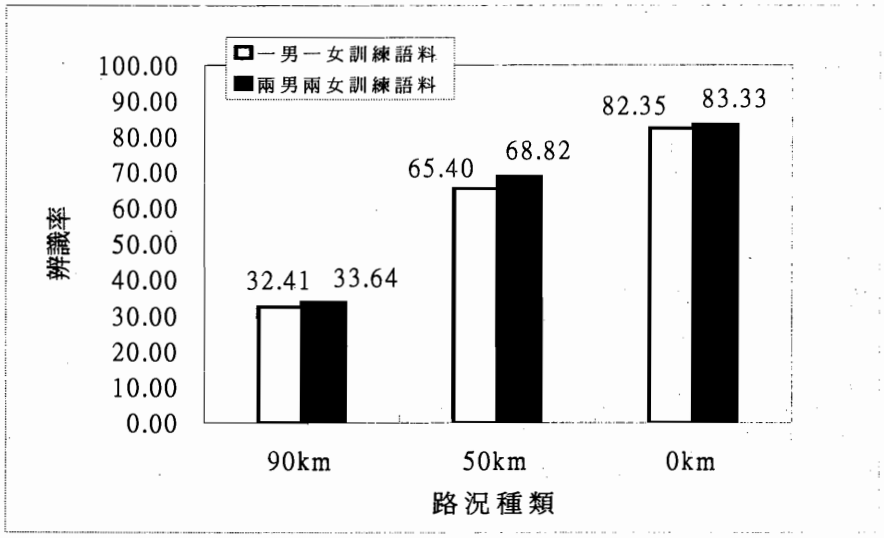


圖五、基本系統及加入λ_e作補償辨識結果的比較

		基本系統	.+λ_e	一男一女訓練語料 .+λ_e+b	兩男兩女訓練語料 .+λ_e+b
原始的估測方法	90km語料	18.2	24.3	27.7	31.5
	50km語料	36.5	49.4	57.7	63.9
	0km 語料	70.1	87.7	82.4	86.8
改良的估測方法	90km語料			32.4	33.6
	50km語料			65.4	68.8
	0km 語料			82.4	83.3

表一、訓練語料的多寡對辨識率的影響

我們作效能評估所使用的電腦為**P2-350**，**RAM**為**128MB**，表二為每句語音辨識所需消耗的時間比較表。



圖六、訓練語料與辨識率比較圖

效能評估	基本系統	.+λ	.λ+b
秒/句	0.41	0.67	0.74

表 二、語音辨識的消耗時間比較表(單位：秒)

在λ_e的計算中，每句大約要耗掉**0.26**秒，再加上其他**function call**的時間，基本系統與作平均值補償大約差**0.74-0.41=0.33**秒，因為查表動作沒有佔**CPU**很多時間，所有作λ_e補償與作

λ_e及**b**(λ_e)補償的時間相差不多，但我們必需額外付出一些記憶體來儲存補償函式，如果等化因子的範圍為**-2**到**4**，平均值及變異數補償函式需要**2**(函式個數)*600(λ_e量化的點數)*26(特徵向量維度)=**31200**筆浮點數大小，總共要**121KBytes**。

如果我們以改良後的方法為主，並把等化因子的範圍限制在**0~3**之間，我們發現使用較小的λ_e範圍降低了些許辨識率，但所需的記憶體為**1**(平均值補償函式)*300(λ_e量化的點數)*26(特徵向量維度)=**7800**筆浮點數，約為**30KBytes**，可大幅減少記憶體的使用量。

實驗結果最後部份我們混合訓練語料及測試語料，以觀察不同汽車對辨識率的影響。表三是以之前的語料庫組合方式所得的最後結果，訓練語料與測試語料是由不同的車子錄得。而現在我們把含有這兩台汽車的語料都混合在一起，混合後的訓練語料有**526**句；其中零公里部分有**118**句，五十公里部分有**206**句，九十公里部分有**202**句。而混合後的測試語料共有**788**句；其中零公里部分有**203**句，五十公里部分有**262**句，九十公里部分有**323**句。表四為我們混合不同汽車語料後的辨識結果。

改良後的方式估測調整函式			λ_e範圍: 0 ~3	λ_e範圍: -2~4
	Baseline	.+λe	.+λ_e+b	.+λe+b
90km語料	18.2	24.4	32.4	33.6
50km語料	36.6	49.4	62.0	68.8
0km 語料	70.1	87.8	82.8	83.3

表三、辨識結果的比較

改良後的方式估測調整函式			λ_e範圍: 0 ~3	λ_e範圍: -2~4
	Baseline	.+λe	.+λe+b	.+λ_e+b
90km語料	26.70	33.85	40.99	41.61
50km語料	48.85	63.35	72.51	77.86
0km 語料	76.47	84.31	87.2	87.25

表四、混合不同汽車語料後的辨識結果

5. 及時展示系統

為了實際評估此演算法的效能，最好的方法就是直接線上做語音辨識，我們並不是直接開車然後在車上作展示；而是先錄製一段汽車背景雜訊，用喇叭播放出來以模擬實際的汽車噪音環境，而麥克風的位置是以免持式遠距離麥克風為主，大約與說話者的距離25到35公分之間，線上錄下一段語音做即時語音辨識。

在展示系統設計過程中我們遇到了一些問題，如背景雜訊是由喇叭播出與實際汽車環境是不相同的，我們不能以汽車上訓練出的補償函式直接用在展示系統中，所以我們採用線上錄音線上訓練補償函式的方式來解決；另一個問題是前後背景雜訊太長會嚴重地導致辨識結果不佳。為了解決這個切音問題我們使用兩階段維特比(Two Pass Viterbi Decoding)辨識方式，並且線上錄下一段背景雜訊以訓練出背景模組(Background Model)並配合Two pass Viterbi Decoding做語音辨識。

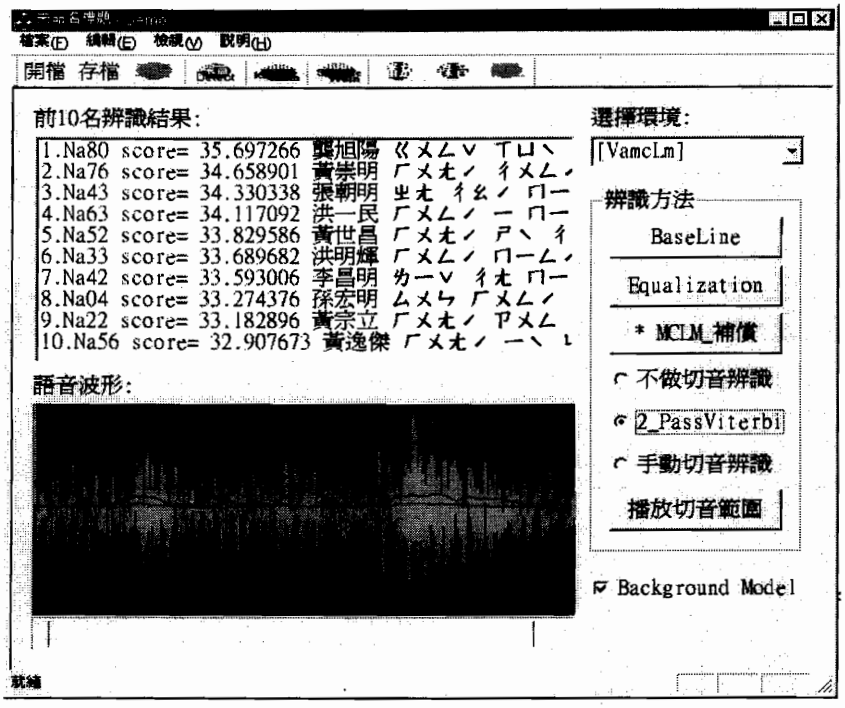
5-1 背景模組

背景模組(Background Model)的訓練方式，會依實際錄得的聲音長短，訓練出不同混合數(Mixture)的背景模組，例如64個混合數或32個混合數等等。實際的訓練方式我們是以向量量化(Vector Quantization)分析程序作群的分類，在做語音辨識時將原始乾淨語音訓練出來的背景模組用新的背景模組取代，在做模型調整時此背景模組不做調整。

5-2 兩階段維特比辨識

在噪音環境下為了可以更準確地切出語音音段，我們使用兩階段維特比的語音辨識方式。所謂兩階段維特比是針對測試語音執行兩遍維特比(Viterbi Decoding)程式，第一遍維特比先段出語音的起始點及結束點，第二遍維特比再將段出後的語音進行辨識。為了可以準確地段出語音音段我們會結合在前一節所介紹的背景模組(Background Model)，這樣可以使切音的結果更加準確。圖七是我們在Windows平台上發展出的一套系統展示介面，在這套介面上我們可線上錄製一段聲音，線上訓練出背景雜訊模組及平均值補償函式，也可線上錄音放音並及時求出辨識結果。辨識的方法有：Baseline不作任何補償的方法；Equalization以等化

因子作補償的方法；MCLM結合等化因子及平均值補償函式的辨識方法。我們並以辨識結果的前10名來觀察不同的辨識方法的辨識效果。



圖七、Demo 系統介面

6. 結論

依我們的演算法先計算等化因子，在進行相似度量測時把這兩個調整函式結合進去，可大幅回升在噪音環境下語音辨識正確率。我們以實際的噪音環境下錄下少數語料庫訓練出調整函式，就可以有很好的語音模型調整效果。

實驗結果顯示我們以改良後的方法辨識率有提升，而且可以降低記憶體的需求量，將來如果要實際實做成晶片，把它應用在汽車噪音環境裡的對話系統或撥號系統，這是一個很實用的演算法。因為本方法需要額外的CPU時間計算等化因子λ_e，以及少量的記憶體空間儲存調整函式，相信在未來電腦硬體技術會快速發展，這些額外的需求將不成問題。

致謝:

感謝工研院電通所前瞻技術中心在本研究上的協助

參考文獻

[1] Jen-Tzung Chien and Ming-Shung Lin (1999)，"Noise Compensation approach to hands-free speech recognition in the car", Proc Workshop on Distributed System Technologies and Application , pp.80-87, Taiwan-Tainan (in Chinese)

[2] Boll, S. F. (1979), "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoustic, Speech, Signal Processing*, Vol. 27, pp. 113-120.

[3] Carlson, B. A. and Clements, M. A. (1991), "Application of a weighted projection measure for robust hidden Markov model based speech recognition", *IEEE Proceedings of International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 921-924.

[4] Carlson, B. A. and Clements, M. A. (1994), "A projection-based likelihood measure for speech recognition in noise", *IEEE Trans. Speech and Audio Processing*, Vol. 2, pp. 97-102.

[5] Chien, J. T., Wang, H. C. and Lee L. M. (1998), "A novel projection-based likelihood measure for noisy speech recognition," *Speech Communication*, Vol. 24, no. 4, pp. 287-297, July 1998.

[6] Gong, Y. (1995), "Speech recognition in noisy environments: A survey", *Speech Communication*, Vol. 16, pp. 261-291.

[7] Lee, C. H. (1997), "On feature and model compensation approach to robust speech recognition", *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 45-54.

[8] Lee, C. H., Giachin, E., Rabiner, L. R., Pieraccini, R. and Rosenberg, A. E. (1992), "Improved acoustic modeling for large vocabulary continuous speech recognition", *Computer Speech and Language*, Vol. 6, pp. 103-127.

[9] Mansour, D. and Juang, B. H. (1989), "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoustic, Speech, Signal Processing*, Vol. 37, pp. 1659-1671.