行政院國家科學委員會專題研究計畫　成果報告

子計畫一：無線通訊環境下國語語音之分散式辨認(3/3)

計畫類別：整合型計畫
計畫編號：NSC91-2219-E-002-035-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立臺灣大學電信工程學研究所

計畫主持人：李宇眩

計畫參與人員：吳承晃、王柏淵、何維鴻、黃婉甄、徐禎助、陳彥倫、陳鼎堯、蔡昇甫、陳厚昕、陳禮鈞、黃燦煌、王佐

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 2 月 22 日

第一章　緒論

1．1　研究主題和動機

無線通訊的興起，滿足了人們隨心所欲，隨時隨地交流訊息，傳遞資訊的渴望：例如，不管走到那裡，你都可以和好朋友們交換心情；在緊急事件的當下，透過無線通訊，我們也可以掌握到最快、最新的資訊。而通訊技術不斷地進步，使得我們不但可以利用無線網路進行對談，也可以透過它連上網際網路，來存取想要獲得的資訊。例如，你可能想要隨時下載在你的電子信箱中的郵件；或是當你在會議進行中，發現有某一筆文件忘了帶，你可以透過行動電話呼叫家中的電腦為你傳送過來。此外，無線通訊即將邁入一個新的進程－－第三代行動通訊，其提供的頻寬，可以讓我們享受許多多媒體的服務，屆時，我們可以利用行動電話下載許多的影音資訊，不需要個人電腦和網路線，一樣可以存取網際網路上的資源。但是有一個問題隨之而來，那就是：我們勢必要對行動電話下很多複雜的指令；但是行動電話的體積為了可攜性，方便性的緣故是愈做愈小，也就是說，我們可以在行動電話上安排輸入的按鍵數目將愈來愈少，未來在功能增多情況下將不敷使用。另外，行動電話鍵盤的設

計，本來是用來輸入電話號碼，但在進入高頻寬的第三代行動通訊的時代後，行動電話的角色將不只是通話的工具，為了要應付更多更複雜的指令，原先的鍵盤設計不能符合新的用途。所以尋找一個方便，快速又能處理繁雜指令的輸入介面，是我們所關注的。而利用語音作為輸入介面不啻為一個好方法：語音輸入符合了我們上述方便、快速的原則；同時，由於語音辨識的技術漸趨成熟，對於連續數字（Digit String）的辨識已有讓人相當滿意的成果，大字彙連續語音辨識（Large Vocabulary Continuous Speech Recognition，LVCSR）的結果亦逐步改善中——因此我們可以利用語音來處理繁雜的指令。

根據上面所描述，若採用語音做為新的輸入介面，許多衍生的問題也隨之發生：行動電話的體積太小，其計算能力以及儲存用的記憶體將嚴重受限，使得我們若要在行動電話上處理整個語音辨識的程序是有困難的。我們可以使用一部可以負荷語音辨識計算量、記憶體可以容納辨識所需的資訊的遠端伺服器幫助呼叫辨識服務的用戶處理辨識程序，也就是說，我們要把繁複的語音解碼工作分散在伺服器端和行動電話用戶端兩者之間。那麼，我們該如何在整個系統配置整個音辨識的工作呢？又，當我們配置整個語音辨識工作，意

味著我們必須透過無線通道的環境傳送某些資訊：這些資訊在無線通道中勢必受到各種雜訊的干撓，包括無可避免在電子儀器中的白色雜訊（White Noise）、因多通道衰減（Multi－Path Fading）造成的群集錯誤（Burst Error）等等。在以上所述的條件底下，語音的辨識正確率將會受到多大的影響？有什麼可以改善錯誤的方法？以上的問題，在語音辨識研究的領域中已成為一個新興的主題，稱作分散式語音辨識（Distributed Speech Recognition），而本報告也將針對分散式語音辨識所要面對的問題做一些探討。

１．２　語音辨識的基本說明

在介紹分散式語音辨識系統前，我們先對傳統的語音辨識系統架構做個描述。如圖１．１及圖１．２所示，**語音辨識的流程可以分成二大階段，第一個階段是訓練階段，第二個階段是辨識階段；而每一階段都可以分成二個部分，第一個是聲學特徵部分，第二個是模型參數部分。抽取聲學特徵的原因在於：對於辨識相關的資訊大部分在某些頻譜相關的係數上，我們只須取出該部分資訊即可**。其中梅爾倒頻譜係數（Mel-Frequency Cepstral Coefficient，MFCC）由

於在雜訊環境下是個夠強健的語音特徵抽取法，是最常被使用的方法。梅爾倒頻譜係數的抽取方式如圖１・３所示，語音信號先經過

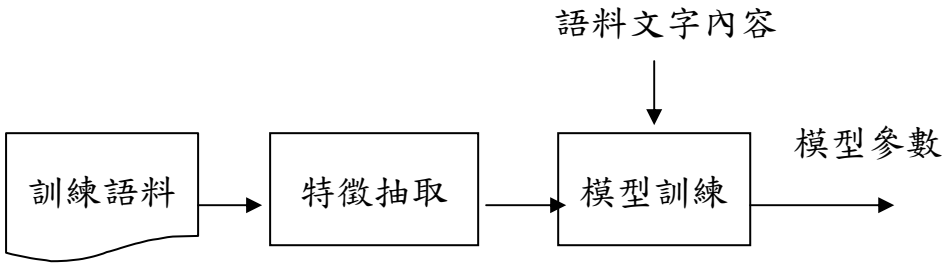


圖１・１　訓練階段

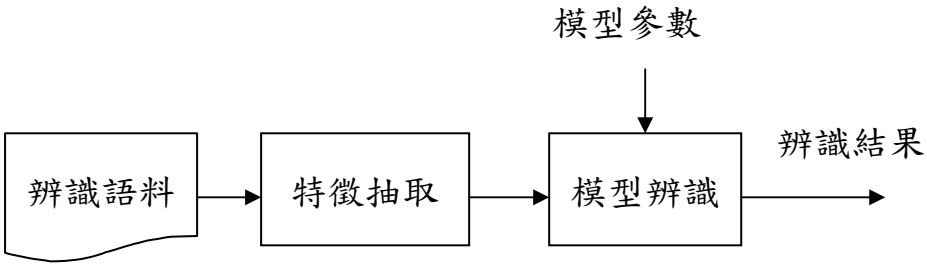


圖１・２　辨識階段

預強調（Pre - Emphasize）加強信號在高頻易受噪音干擾的部份，接著經過漢明窗（Hamming Windowing）使語音信號有短時穩定（Short-Time Stationary）的特性；利用 FFT 把信號轉到頻譜領域後，以梅爾濾波器組（三角形）取出各個共振峰的資訊，然後取對數得到聲帶濾波器的係數，再取 IDCT 得到梅爾倒頻譜係數。另外，有些研究也使用線性預估倒頻譜係數（Linear Predictive Cepstral Coefficients，LPCC）來做為特徵參數，如圖１・４，這個方法是先將語音信號做線性預估分析（Linear Predictive

Analysis）得到線性預估係數（Linear Predictive Coding Coefficients，LPC），再轉到反頻譜領域。這兩種方法在本報告中都將討論到。而在模型的部分，在第一階段中，對於每一個想要

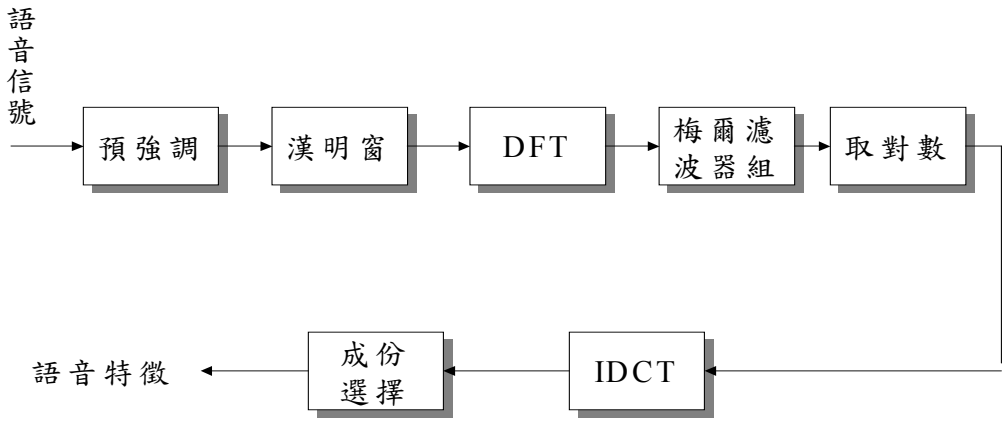


圖１・３　梅爾倒頻譜係數的求法

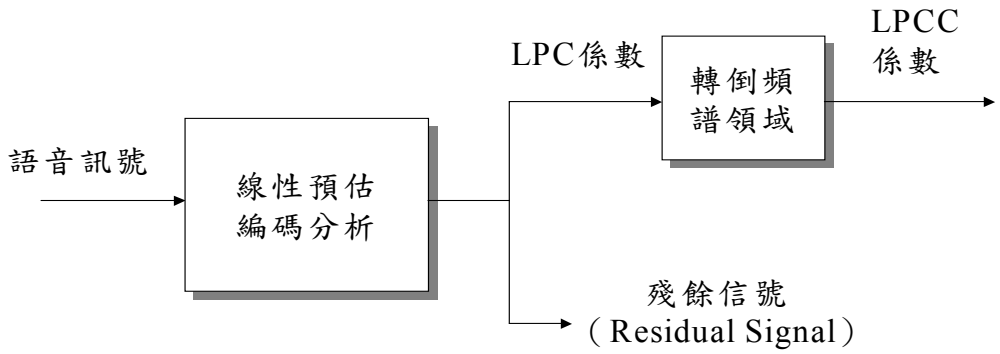


圖１・４　線性預估編碼係數的求法

辨識的單位，例如：音素（Phone），音節（Syllable），字（character）

，詞（Word），都可以建立一個隱藏式馬可夫模型（Hidden Markov Model），而每個模型的參數，則由訓練語料中求得；而在第二階段中，我們就使用第一階段中所得到的模型來做辨識。簡單的說，辨識就是根據所收到的語音信號，利用前面所找出的模型來做評分，分數最高者就選定該模型所代表的單位為最後的辨識結果。

1．3 分散式語音辨識的相關問題

分散式語音辨識的問題如前面所述，在用戶端所使用的行動電話有功能上的限制，要直接在行動電話上處理整個語音辨識的工作是有困難的。因此我們必須把語音辨識的架構給“分散”在伺服器端以及用戶端。**如何配置**就是顯而易見的第一個問題。另外，整個系統是架構在無線網路的環境中：**在無線傳輸中，我們必須考慮頻寬的限制**，因此我們必須選擇對語音辨識有所幫助的資訊來做傳輸，甚至必須對這些資訊作壓縮，才能夠有效的節省頻寬。另外，**無線通道會帶給訊號錯誤**，這些錯誤將導致辨識正確率的下降；我們必須建立錯誤更正的機制，才能使語音辨識的工作不受通道環境造成的錯誤的影響。另外，由於行動電話使用的環境具有許多的背景雜訊

，例如：嘈雜的人聲，汽車的引擎聲等，這些會造成語音辨識正確率下降的因素，仍然存在於分散式語音辨識系統中。以上所述都是目前最被熱烈討論，有關分散式語音辨識的問題。

1．4 研究主題和主要成果

在這篇報告當中，我們將探討上一章節所提出的幾個問題：首先針對分散式語音辨識系統配置的問題，我們將比較兩個最常使用的架構－－主從式（Client－Server）架構以及純主式（Server－Only）架構；我們會對兩個架構作詳細的描述，包括兩個架構的優缺點比較。接著我們將對主從式架構中，兩個造成辨識率下降的原因進行探討：第一個原因是為了降低傳輸速率而對資料進行壓縮所帶來的量化錯誤；第二個原因是無線傳輸中通道的不完善所造成的錯誤。針對第一個原因，在報告中我們以ETSI(the European Telecommunications Standards Institute)的壓縮方式為基礎【1】，提出類似的以中文為基礎的量化方法以減少量化錯誤，而我們的實驗也證明了，在中文大字彙連續語音的辨識中，若以音節做為辨識的單位時，因量化錯誤所造成的正確率下降是非常的少。

而對於第二個原因，在報告中我們分別用兩種情形－－隨機錯誤（Random Error）和群集錯誤（Burst Error）去模擬通道錯誤的情形，試著去找出，到底通訊系統必須把位元錯誤率（Bit Error Rate）降到那一個程度時，語音的辨識正確率才不會有嚴重的退步。另外，我們也討論三種可行的錯誤補償方式，其中兩種－－消去法、外插法是參考前人的作法【２】，另一種是假設在每個傳輸的音框（Frame）中分別對其中的一小部分做錯誤控制編碼，如此一來便可以只對錯誤的部分做修正而保留正確的部分。在我們的實驗中，發現在做了錯誤補償後的確可以改善錯誤的現象，甚至趨近沒有通道錯誤發生時的辨識正確率。在深入的探討主從式架構後，我們也繼續討論另一種熱門的架構－－純主式架構。在這個主題上，我們以GSM（Global System for Mobile Communications）新一代的聲碼器－－調適型多速率聲碼器（Adaptive Multi-Rate Speech Vocoder）為基礎，首先比較利用調適型多速率聲碼器輸出的合成語音（Synthesized Speech）做辨識以及利用原始語音做辨識在正確率上的差異，並以此論證用合成語音做辨識之不可行。接著介紹如何從調適型多速率聲碼器輸出的合成語音當中，找出可供辨識的資訊，並從實驗中去看這個方法的效果如何。

1．5 章節大要

本報告的章節大要如下：在第二章的部分，將詳細介紹分散式語音辨識系統，並再次強調本報告所要解決的問題。第三章主要介紹在報告中所使用的實驗環境架構，包括語音資料庫，語音訊號的特徵抽取，聲學模型架構以及衡量辨識效果的方法，最後介紹整個分散式語音辨識系統的架構。第四章討論在主從式架構下向量量化所造成的影響。第五章討論在主從式架構下通道環境對語音辨識的影響。第六章討論純主式架構的作法。最後在第七章做結論，以及對未來的展望。

第二章　分散式語音辨識系統的介紹

2．1　分散式語音辨識系統在實行上的分類

在第一章中，我們了說明了分散式語音辨識技術所要面對的幾個問題：

一．語音辨識的工作必須適當的分配給遠方的伺服器以及用戶所持有的手持設備（Hand－Held Devices）：因為手持設備的**記憶體有限**，要完全儲存語音辨識所須要的模型參數是有困難的；且手持設備的**計算能力也有限**，對於語音辨識這種需要龐大計算量的工作也不能負荷。

二．由於語音辨識的工作是架構在無線網路的環境中，**資料的傳輸受到頻寬的限制**，因此如何擷取出對辨識有幫助的資訊，同時有效的表示這些資訊，也是分散式語音辨識系統中所須要考慮的。

三．**在無線通道的環境中，傳輸的信號會受到許多雜訊的干擾**。例如，在電子儀器中無可避免的熱雜訊（Thermal Noise）的干擾、無線通道中通道本身的改變、同頻干擾（Co–Channel Interference）或鄰頻干擾（Inter-Channel Interference），因為多路徑衰減所造

成信號的忽強忽弱等，這些會使信號產生錯誤，同樣也影響語音辨識的結果。

根據以上所述，發展出以下三種分散式語音辨識的模型，我們將一一的來做討論。

２．１．１ 純從式架構（Client-Only Model）

如圖２．１所示，純從式架構的基本模型如下：

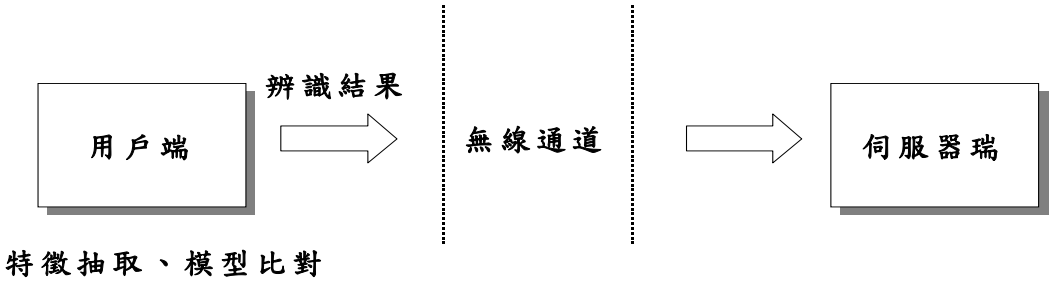


圖2．1 純從式架構

純從式架構的構念，是把所有的語音辨識工作，都在用戶端執行；等到用戶端有了辨識結果，再把辨識結果經由無線通道，傳輸到伺

服器端，伺服器再根據辨識結果回應用戶端的要求。這樣的安排最大的優點在於：語音辨識的工作整個都在用戶端進行，因為沒有分散配置的緣故，只要辨識結果在傳輸的過程中不出錯，那麼語音辨識的正確率便不會受到無線傳輸所帶來的雜訊干擾。不過，這樣安排的缺點也是顯而易見的：想要在用戶端完全的執行整個語音辨識的工作，那麼用戶端的設備也必須具有相當的計算能力與儲存空間。因此，在我們這篇報告中，主要討論的用戶端——手持設備，就不適合這樣的架構；比較合適的用戶端，會是筆記型電腦這樣的設備，除非所要進行的語音辨識工作十分簡單，例如很小的字彙。除了用戶端受限的這個缺點以外，純從式架構在應用上還有另一項不足，那就是這樣的架構，並不適用在一些需要認證的服務上面。就比如說，使用者在請求服務的時候，可能被要求必須通過身份認證，這時候或許可以通過語者辨識的技術來進行認證；但是若採行純從式架構，認證的可信度就會降低許多。基於上述理由，純從式架構的討論我們將就此打住，將研究的焦點集中在接下來要討論的兩個架構上。

2・1・2 主從式架構（Client-Server Model）【3】

如前一節所述，若沒有將整個語音辨識的工作做適當的配置，用戶端所能使用的裝置將受到限制；在純從式架構中，伺服器只是用來回應使用者的服務要求，卻忽略了伺服器同樣可以幫我們處理繁重的語音辨識工作。那麼，要怎樣發展一個新的架構，使得辨識的工

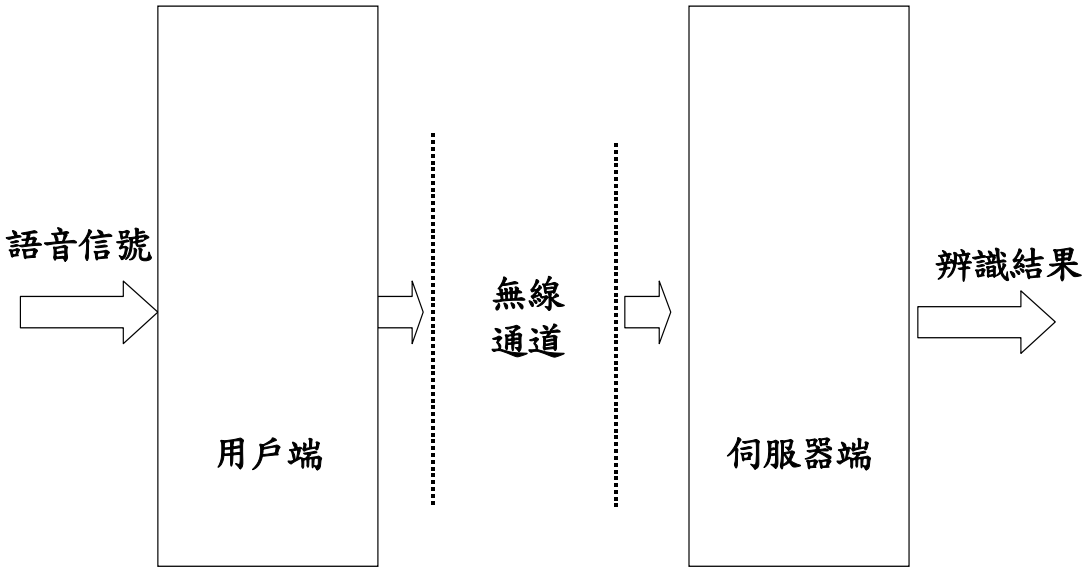


圖2．2　主從式架構

作可以讓伺服器部份負擔，進而使用戶端裝置不會受限呢？

在第一章中，我們曾經介紹了語音辨識的基本架構——主要有兩個工作：特徵抽取以及模型辨識。特徵抽取的目的，是要抽出含有語音資訊的參數，所需的計算量並不會太大；同時，也不會耗費很多

記憶體。而模型辨識的目的，則是利用先前訓練好的統計模型，對抽取出的特徵去作比對；相對於特徵抽取來說，模型辨識因為牽扯到統計上的估算，所需的計算量就大很多；另外，統計模型的參數必須被儲存，所以在模型辨識上也同樣要有大量的記憶空間。綜合以上所述，我們可以很合理的把語音辨識的工作分散在兩部分：在用戶端所持的行動電話上抽取出辨識所需的參數，並經無線通道傳輸這些參數，在伺服器端執行模型比對的工作。如圖２．２所示。

我們再從另外兩個分散式語音辨識系統所要克服的問題去做討論。在傳輸的過程當中，所能使用的頻寬受到限制；因此我們必須從語音信號中抽取出對語音辨識有用的資訊——而這也正是特徵抽取的部分所要做的。就傳統最常使用的梅爾倒頻譜係數而言，其目的是要取出和語音信號內容關係最密切的聲道（Vocal Tract）的特徵轉換函數（Characteristic Transfer Function），這樣的作法本身就含有壓縮的觀念，也是通訊上為了節省頻寬所必要的。因此，若在用戶端抽取出梅爾倒頻譜係數，我們就可以達到利用有限的頻寬傳輸足夠的資訊的要求。

另一方面，在傳送的過程中，因為會受到無線通道中雜訊的干擾，

所以我們必須選擇在有雜訊情況下較強健的特徵參數。而許多的實驗都顯示，若使用梅爾倒頻譜係數，在強健性方面有不錯的表現。另外值得一提的是，在語音辨識中，每一項參數都是設成浮點數，因此在傳送的過程當中，我們必須另外再做資料壓縮的動作，以進一步減少頻寬的使用。因此，當我們利用無線傳送辨識用的參數時，實際上傳送出去的，只是一個代碼（Codeword）－－這個代碼所代表的參數，是所有可傳輸的參數中和原來想傳的參數最類似的。這樣的作法也帶來一個好處：我們可以在這些代碼上做安排，使愈不相似的參數各被不相似的代碼所代表，如此一來就可以防止無線通道中可能遇到的雜訊干擾了。

主從式的架構安排似乎是解決了分散式語音辨識的所面對的三大問題，然而它還是有三個缺點：

一．應用中，在伺服器端需要知道被傳輸的語音信號，而不光只是取出的特徵參數。這類的應用，例如，一些線上的交易，就會需要根據原來聲音的內容來認證語音的工作；另外，在系統初始化的時候，為了效能的最佳化，如果能夠知道原來聲音的內容，那麼就可以對辨識系統進行微調。當然，使用者難免也需要用手持設備直接打電話給別人，而不是光是作有關語音辨識

的工作；此時對方也要能得到原來的語音。但是，當我們使用主從式架構時，傳送到伺服器端的只是抽出的梅爾倒頻譜係數，並非是原來的語音，而想要將梅爾倒頻譜係數轉回原來的語音是有困難的。

二．行動電話對於語音信號的處理，是先將語音信號經過語音編碼器（Speech Encoder）的處理，再經由無線通道來傳輸。如果我們要在行動電話上面抽取出辨識所需的梅爾倒頻譜係數，那麼勢必要在手持設備上另外加入一個用來抽取特徵參數的功能元件。也就是說，現有的手持設備不能見容於主從式的架構當中。此外，因為多了特徵係數的傳輸，通訊協定也要做修正。

三．在主從式架構中，要傳送的除了編碼語音外，還有特徵係數，如不能建立一個有效的通訊協定，則兩者必須同時傳送，會造成頻寬上的浪費。

雖然主從式的架構有上述三個缺點，但是它克服了我們一再提到的分散式語音辨識系統所面對的問題；目前 ETSI（the European Telecommunications Standards Institute），已經把這樣的架構制定標準，雖然未來這個標準是否會廣為使用仍有待時間的考驗，但在本報告中，我們將詳細討論這個架構。

２・１・３ 純主式架構（Server-Only Model）

●純主式架構之一

在上一節當中，我們討論到主從式架構的兩個缺點，其中一個是無法在伺服器端得到原來的語音信號，另外一個是我們必須在手持設備中加入抽取特徵參數——梅爾倒頻譜係數的功能元件，才可以完成主從式架構。那麼，我們是否有辦法利用原有的手持設備來執行語音辨認的服務呢？

手持設備對於語音信號的傳輸，是先將語音信號做編碼後，再傳送到無線通道中；而在接收端，是將編碼語音（Coded Speech）經由語音解碼器（Speech Decoder），轉成合成語音（Synthesized Speech）以供人耳收聽。那麼，如果我們想要利用現有的手持設備架構，一個立即可行的方法就是利用語音解碼器輸出的合成語音來做辨識，如圖２・３所示。像這樣的架構，因為所有辨識的工作都在伺服器端完成，於是稱之為純主式架構。這樣的架構的確克服了上述主從式架構所面對的缺點；另外，我們再考慮之前所提到分散式語音辨識所要面對的三大問題：第一，因為在用戶端所使用的手持設備，

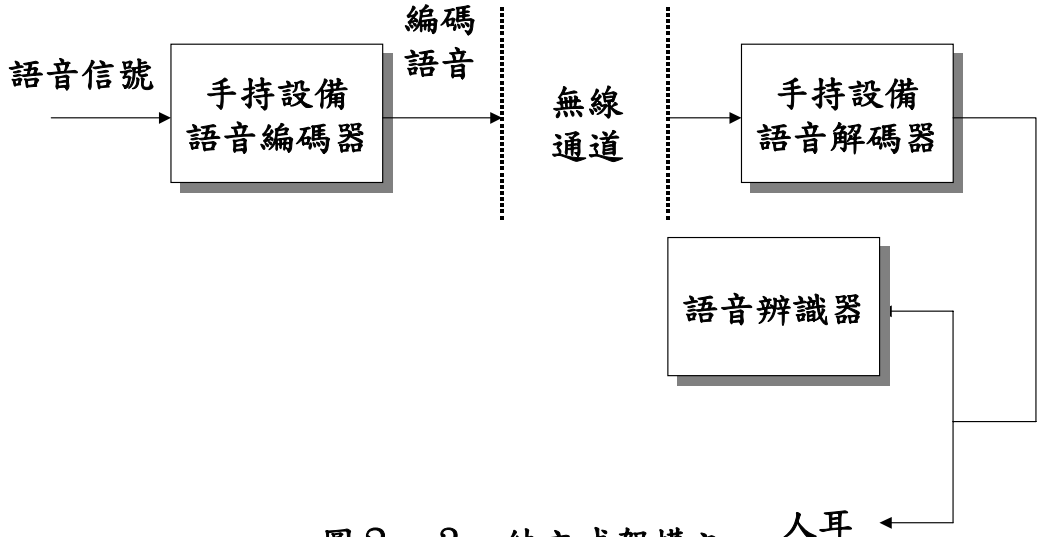


圖２、３　純主式架構之一

只有做原來就有的語音編碼的工作，把整個辨識的工作都交給伺服器去進行，所以並不會有手持設備計算能力或是記憶空間不足的情形；第二，手持設備的語音編碼器有效的壓縮語音，傳送出去的只是編碼後的語音，自然不會有頻寬的問題（這也克服了主從式架構的第三個缺點)；第三，由於純主式架構相容於現有的通訊系統，而在現有的通訊系統當中就有做對於無線通訊中的錯誤保護，因此可以克服雜訊的干擾。

不過純主式架構所遇到的最大問題是：語音編碼器的工作是對語音做有效的壓縮，而語音解碼器的工作是要得到人耳可以聽得懂、品

質不會太差的合成語音——語音編碼的目的是耳朵好聽、而不是為了語音辨識的用途設計的。所以在伺服器端所收到的合成語音，對於人耳來說是可聽、可辨而且和原來語音的品質相去不遠的；但是，對於語音辨識中重要的特徵參數的部分，經過語音編碼後卻改變了許多。因此在一些相關的報告【４】中有提到，如果我們用原始語音去訓練辨識用的統計模型，但用合成語音去做辨識，這樣就造成一個不匹配（Mis-Matched）的情況，辨識效果將會下降；而就算是用語音解碼器輸出的合成語音去訓諫模型，也用合成語音去做辨識，還是會比純粹使用原始語音做訓練及辨識的效果來得差（正確率約下降三個百分點）。

純主式架構由於觀念上相當直接，所以這樣的架構，是分散式語音辨識系統中，最早被提出來討論的（除了純從式架構外，純從式架構有用戶端的限制，所以不討論）。但是由於聲碼器帶給語音信號的破壞，因此在這個架構被討論之後，許多改進的方向也被考慮。其中，第一個被討論的方向，是針對語音的調適：有些研究【５】～【８】致力於語音信號對於無線環境的語音調適，試著從這個方向提高在伺服器端的語音辨識率。另外，架構上的調整也是被考慮的方向：上一節中所提到的主從式架構，以及下面即將提到對純主

式架構的改進，都是朝這個方向改良所得的結果。為了讓改良前後的純主式架構有所區隔，我們以下均稱改良前為純主式架構之一，改良後為純主式架構之二。

●純主式架構之二

因為語音信號會遭受到聲碼器進行壓縮時的破壞，因此若想保有上述純主式架構的優點，也想兼顧語音辨識的正確率，那麼就不能使用解碼後的合成語音來做辨識；我們必須想辦法在語音編碼器的輸出中，找出對辨識有幫助的資訊。

大多數的語音編碼器都是採用線性預估分析（Linear Prediction Analysis），取出線性預估參數（Linear Prediction Coding Coefficient，LPC）以及殘餘信號（Residual Signal），再分別對這些信號作壓縮、編碼。對語音信號做線性預估分析，其實是把發聲的過程模擬成一個激發（Excitation）信號通過一個濾波器；其中，對於無聲子音而言，激發信號是類似白色雜訊（White Noise）的信號，對於有聲子音以及母音而言，激發信號是一連串週期性的脈衝（Impulse）；

另外，濾波器是用來模擬聲道的。而線性預估參數是濾波器的轉換函數（Transfer Function）的係數，而殘餘信號則是激發（Excitation）信號。在前面我們曾經提到過：對於辨識有幫助的資訊，大部分在代

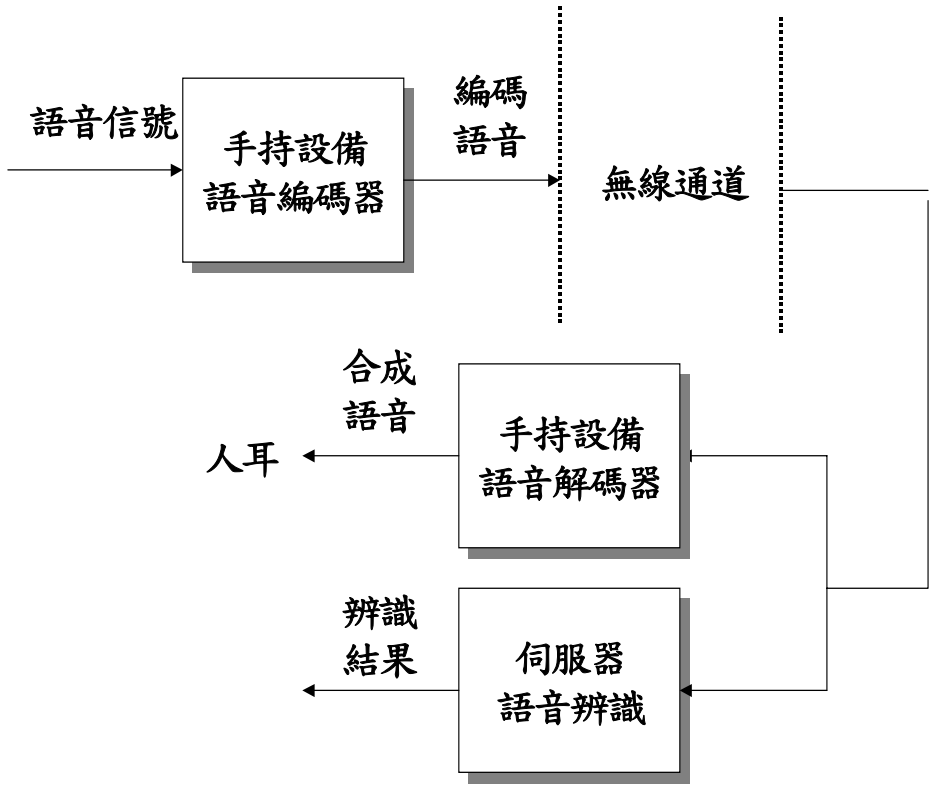


圖２・４　純主式架構之二

表聲道的轉換函數中——這部分的資訊恰好是語音編碼器所取出的線性預估係數所代表的。因此，改良後的純主式架構之二已經呼之欲出了：在用戶端所持的手持設備仍然只做語音編碼的工作，然後將編碼後的語音送入無線通道中；較大的變動發生在伺服器端，如果要做語音辨識的工作，伺服器直接從接收到的編碼語音抽出辨識相關的資訊；如果要讓人耳收聽，就把編碼語音送入語音解碼器得

到合成語音。如上所述，如圖２．４所示。

由於語音編碼器的種類眾多，所以抽取特徵參數的方式也不一致。在第六章的部分，我們將使用 GSM 新一代的聲碼器－－調適型多速率聲碼（Adaptive Multi-Rate Vocoder）來進行純主式架構之二的實驗，屆時對特徵的抽取會有更詳細的描述。

２．２　分散式語音辨識系統的應用

當伺服器端執行完語音辨識的工作後，伺服器可以根據辨識的結果對使用者提供相當多的服務，也因此分散式語音處理的後端可以加入許多語音的相關應用，我們列舉如下：

一．關鍵詞擷取（Keyword Spotting）

為了提供服務，在伺服器端必須知道使用者所做的請求為何。所以，我們可以拿辨識的結果去做關鍵詞抽取，找出想要的資訊。

二．口語對話系統（Spoken Dialogue System）

為了提供服務，伺服器常常需要跟使用者做許多的確認；另外

，伺服器也需要另外跟使用者對談，告知可以提供的服務有那些。所以，在分散式語音辨識系統的後端，也可以加入口語對話系統的應用。

三．資訊檢索（Information Retrieval）

在目前，利用網際網路搜尋資訊已成一股風潮。而可預期的，使用行動電話來搜尋網路上的資訊的風氣，也會慢慢的興盛起來。所以我們也可以在分散式語音辨識系統的後端，加入資訊檢索的功能。

四．多媒體應用（Multi-Media Application）

近來在行動電話上，已經可以看到許多酷炫的聲光效果，因此可以預期的，未來在無線通訊上會有愈來愈多的多媒體應用；如第一章的緒論就提到的，分散式語音辨識系統可以提供我們一個使用這些多媒體應用的方便介面，也就是說，在分散式語音辨識系統的後端，我們可以加入多媒體的應用。

五．聽寫、轉譯、語音合成（Dictation、Transcription、Speech Synthesis）

最後的辨識結果可能需要被記錄下來，因此可以在

系統的後端加入聽寫的應用；辨識結果也可能須要被翻譯成另一種語言，所以也可以加入轉譯的功能；另外，前面有提到某

些服務需要合成語音來做認證或是微調，所以我們也可以加入合成語音的功能。

分散式語音辨認系統的後端還可以加入其他的功能，但本文對這些不多做討論。報告的重點將放在如何克服架構在無線環境所帶來的種種雜訊干擾，以及探討不同的架構的差別。

第三章　實驗背景與基礎架構

3・1　實驗基礎架構

3・1・1　語音資料庫（Database）

本報告所使用的語音資料，是由聲碩公司所錄製的麥克風語料，這是一套中文大字彙連續語音（Large Vocabulary Continuous Speech）的語音資料。我們用表３・１作簡單的介紹：

錄製方式	麥克風
取樣頻率	１６　KHz
訓練語料內容	男性語者兩百人，女性語者兩百人， 每個人兩百句，共四萬句
測試語料內容	男性語者五人，女性語者五人， 每個人三百二十七句，共三仟二佰七十句。
每句話的長度	平均三秒鐘，且平均一句話有中 文的音節（Syllable）約十一個。

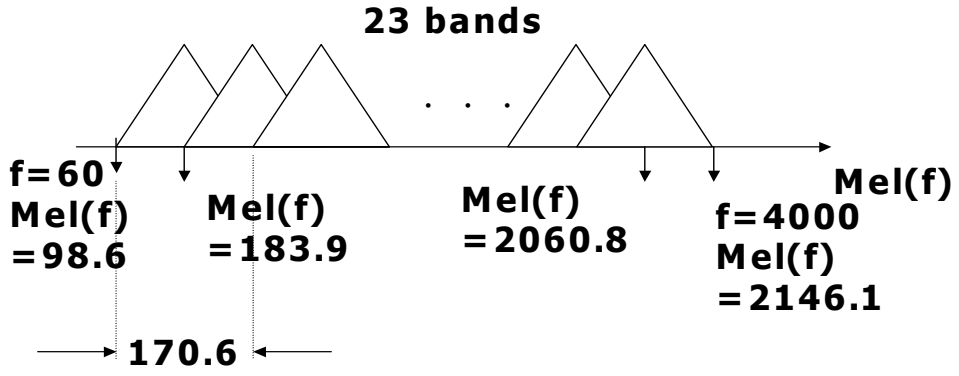
表３・１　語音資料庫說明

３・１・２　語音信號的特徵參數抽取

取樣頻率	８　KHz
快速傅利葉（FFT）轉換點數	２５６
音框長度（Frame Size）	２５　ms
音框平移（Frame Shift）	１０　ms
濾波器組（Filter Bank）	梅爾刻度三角形濾波器 濾波器個數＝２３
特徵向量（聲、韻母模型部分）	１２維的梅爾倒頻譜係數（MFCC）＋１維的能量頻譜係數（Energy），１２維的一次回歸梅爾倒頻譜係數（ΔMFCC）＋１維的一次回歸能量頻譜係數（ΔEnergy），１２維的二次回歸梅爾倒頻譜係數（Δ^2MFCC）＋１維的二次回歸能量頻譜係數（Δ^2Energy），共３９維

表３・２　特徵參數的抽取方式

如前一章所述，在這篇報告中，我們會使用到兩種的特徵參數－－主從式架構中的梅爾倒頻譜係數以及純主式架構之二中的線性預估係數。在純主式架構之二中，線性預估係數是從語音編碼器輸出的編碼語音中得到的，因為方法特殊，所以在這邊先不做描述，在第六章介紹純主式架構之二的部分會另做說明；在這個地方，我們說明的是傳統語音辨識以及主從式架構中梅爾倒頻譜係數的抽取。詳細情形如表３・２所示。



$$Mel\ (f) = 2595.0 * \log(1 + \frac{f}{700.0})$$

圖３・１　梅爾濾波器組

表３・２中，有幾點是需要另外說明的：

一・在第二章主從式架構的介紹當中，我們提到了 ETSI

（the European Telecommunications Standards Institute）將主從式的架構列為標準【１】，而這個標準，也詳細的規定了特徵抽取的方式。為了相容於這個標準，我們把語音資料的取樣率由原來的１６KHz 轉為８KHz；另外，上表所列舉出的各項也是參照這個標準來設定。

二．圖３．１是梅爾濾波器的設計，包括如何將頻率換成梅爾刻度的頻率，以及三角形濾波器在梅爾刻度的頻率上的安排。

３．１．３ 聲學模型架構

本報告所採用的是連續密度隱藏式馬可夫模型（Continuous Density Hidden Markov Model，CDHMM）。如圖３．２所示，模型的架構是採以左到右（Left-To-Right）的形式，總共有五個狀態（State）：狀態０和狀態４分別代表進入（Entry）和跳出（Exit）狀態，進入狀態代表進入了這個模型，此時只能往下一個狀態前進，而跳出狀態代表離開了這個模型；其餘三個狀態的轉移方式是只允許從一個狀態跳至鄰近的下一個狀態，或是留在原狀態上。狀態和狀態間的轉移是由一組機率來描述，稱之為轉移機率（Transition

Probability)。要進入每個模型，有一組起始機率（Initial Probability）另外，在狀態１～狀態３中，每個狀態可能出現的特徵參數向量的

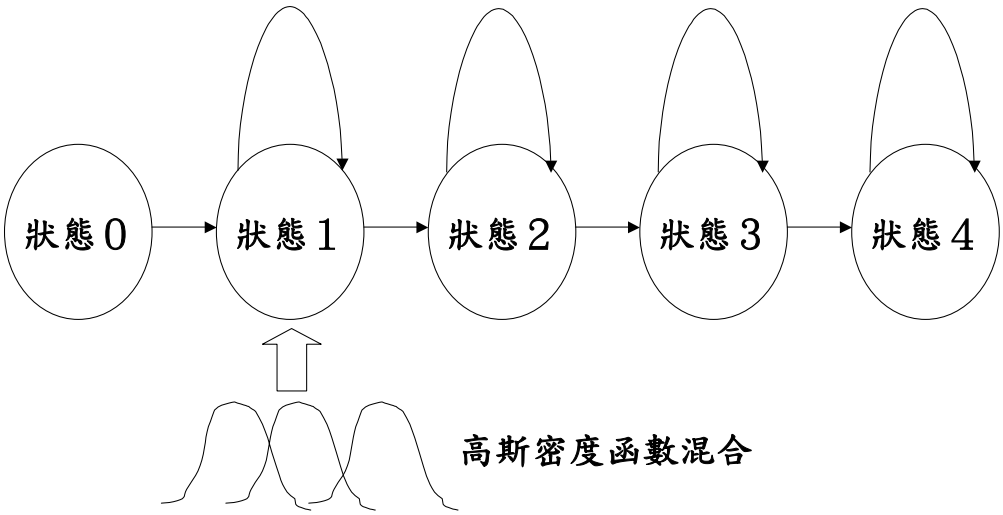


圖３・２　隱藏式馬可夫模型表示法

機率是用高斯密度函數來描述；而為了能廣泛的描述所有的語者的聲音，我們使用了二個，四個或八個高斯密度函數混合（Gaussian Density Function Mixture）作為該狀態下特徵向量出現的機率密度函數，稱之為狀態輸出分布函數（State Output Distribution）。模型的建立，就是利用訓練語料來求出起始機率、轉移機率及狀態輸出函數。

另外，在對音節的描述上，我們把每一個中文音節分成聲母加上介

音及韻母兩個次音節單位（Sub-Syllable Unit），而每一個統計模型，都是用來辨識這些拆解後的次音節單位。例如，「錯」這個字是拼成「ㄘㄨㄛˋ」，在我們所使用的音節描述法中應是「ㄘㄨ」加「ㄛ」。這個音節描述法稱為「聲母加上界音、韻母」模型（Pretoneme）另外，為了進一步描述音節拆解之後，各個次音節單位之間的關係，我們採用音節內右相關模型（Intra-Syllable Right-Context Dependent Model，Intra-RCD）【９】。這個模型是假設在每個音節中的次音節單位才有相關性，不同音節的次音節單位沒有相關性；而每音節中的第一個次音節單位（聲母加上介音的部分），會受到第二個次音節單位（韻母的部分）的影響，所以對第一個次音節單位而言，若後面所接的第二個次音節單位不同會有不同的模型描述。除了上述音節拆解所得到的次音節單位會用來建立模型

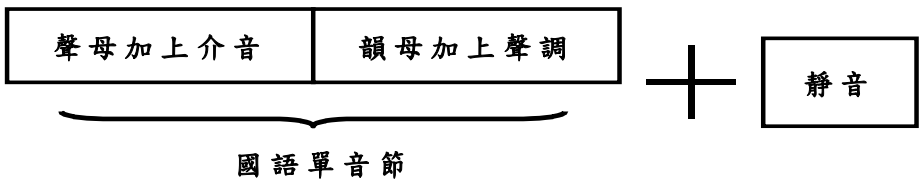


圖３．３　所有的聲學模型分類

以外，因為語料中聲音有短暫靜音的效果，所以我們使用一個靜音模型（Silence Model）來描述這個現象。全部的聲學模型總共有４２９個。圖３．３是所有的聲學模型分類。

３．１．４　辨識效果的衡量方法

在上一節當中，我們說明了辨識模型是建立在對次音節單位的描述，但是在衡量辨識效果上，我們是以音節（Syllable）做為辨識的基本單位；也就是說，我們將經由隱藏式馬可夫模型所辨識出來的次音節單位再結合成音節，由音節去衡量辨識效果。在本報告中，我們使用音節辨識正確率（Syllable Recognition Accuracy Rate）做為辨識效果的衡量標準，辨識正確率越高越好。其計算方法如下：

辨識結果共分成四種，正確、插入（Insertion)、取代（Substitution)、消去（Deletion)，

音節辨識正確率（Syllable Recognition Accuracy Rate）

＝（音節正確數－音節插入數）／（總音節數）

在以下章節的實驗當中，我們都以音節辨識正確率來衡量辨識的效果。

3．2 基礎實驗

在上一節中，我們提到用音節辨識正確率來作為衡量語音辨識的標準；而在第二章的開頭，我們曾討論分散式語音辨識系統遇到的幾個問題，這些問題，都可能造成辨識正確率的下降。在這一節當中，我們將建立一個基礎實驗：這個基礎實驗沒有模擬無線傳輸的環境，而是直接作傳統的語音辨識。因為無線傳輸會帶來錯誤，預期將會帶來辨識正確率的下降，所以基礎實驗所得到的結果，將是在無線環境下語音辨識效果的比較標準。

如表３．３所示是基礎實驗所得到的結果。基礎實驗是把語音資料庫訓諫語料中所有的兩百位語者、共四萬句話拿來訓練一個語者不限定（Speaker Independent）的模型，而用測試語料中所有的十位語者去做測試。在結果中我們可以看到的是，隨著高斯混合數的增

	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8
音節辨識正確率	**42.64**	**49.20**	**55.31**

表３．３　基礎實驗的辨識正確率

加，辨識正確率也隨之上升，而高斯混合數由２到４所得到的進步幅度，會比高斯混合數由４到８所得的進步幅度來得大。雖然從數據上可以知道高斯混合數是愈多俞好，但在實驗當中，我們只討論２～８個高斯混合。這是因為考量到在辨識時若高斯混合數愈大，則所花的計算量會愈大的緣故。

第四章　主從式架構（一）：特徵向量量化後對語音辨識結果的影響

在第二章中，我們已經詳細的介紹了主從式架構的工作配置方式：如圖４．１，用戶所持的手持設備只做特徵抽取的工作，之後就把特徵參數經過無線通道送到遠方的伺服器，將需要大量記憶空間及繁雜計算能力的辨識工作交給伺服器來做。由於受到無限通道頻寬的限制，所以特徵參數必須先經過壓縮，才可以送進無線通道做傳輸。在本章中，我們將探討傳送特徵參數的資料壓縮方式，並且透過辨識的結果來找出特徵向量的壓縮對語音辨識的影響。

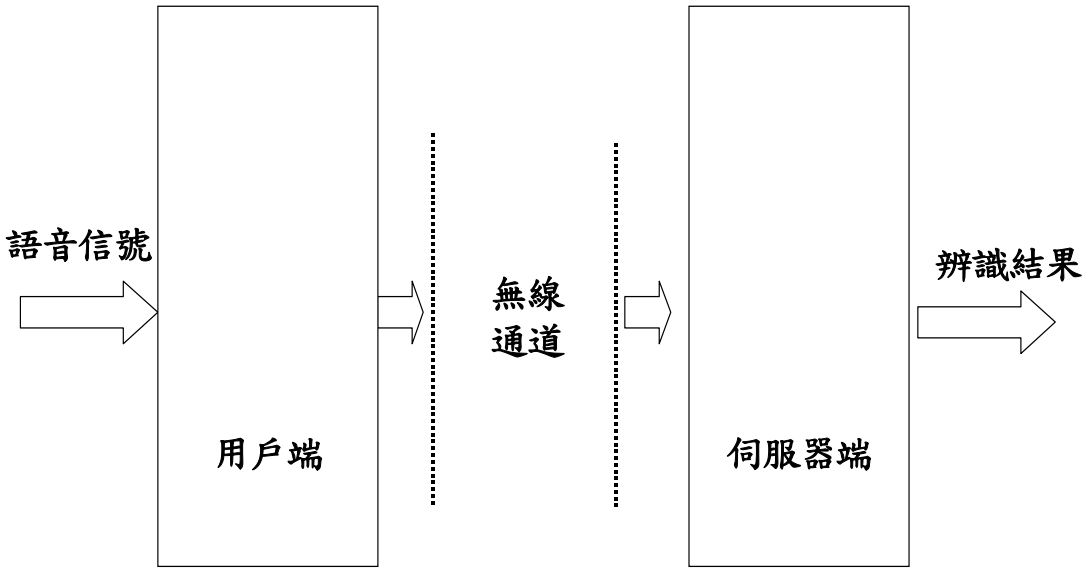


圖４．１　主從式架構

４．１　特徵向量需要壓縮的原因

語音辨識的過程，是取出語音信號中對辨識有幫助的部分做為特徵參數，再藉由訓練過程中所得到的統計模型去和特徵參數做比對，找出最有可能的語音信號內容。在第三章中，我們提到了特徵參數的表示方式：特徵參數是個３９維的向量，前面的１２維，代表每個語音的音框所取出的梅爾倒頻譜係數中的前１２個（在轉換到倒頻譜頻域時是採用 IDCT，其中第０維代表的直流成份並不在上述的１２維當中），第１３維是語音音框的能量頻譜係數（Energy），在此處取其對數值，故以 Log Energy 表示之；接下來１３維的係數，是前面１３維係數的一次回歸；再接下來１３維的係數，又是上一個１３維係數的二次回歸。在我們的辨識實驗中，使用的是浮點數的運算，我們可以計算，如果直接傳送上述的特徵向量，需要多大的傳輸速度：

∵１個浮點數（Floating Number）＝ ４個位元組（Byte）

＝ ３２個位元（Bit），

又特徵向量有３９維，共３９個浮點數，

∴每個特徵向量共需３９ × ３２ ＝ １２４８個位元去描述。

又我們是對每一個音框去取出特徵向量，音框產生的速率是每秒鐘

１００個，

∴所需傳輸特徵向量的速率

＝ １２４８位元／音框 × １００音框／秒

＝ １２４８００位元／秒

＝ １２４．８仟位元／秒（Kbps，Kbits／sec）

這個傳輸速率是非常的高，會佔去相當大的頻寬，在無線通訊上是不可行的！因此對於特徵向量，必須先做壓縮以減少資料量，然後再傳送到無線通道中。

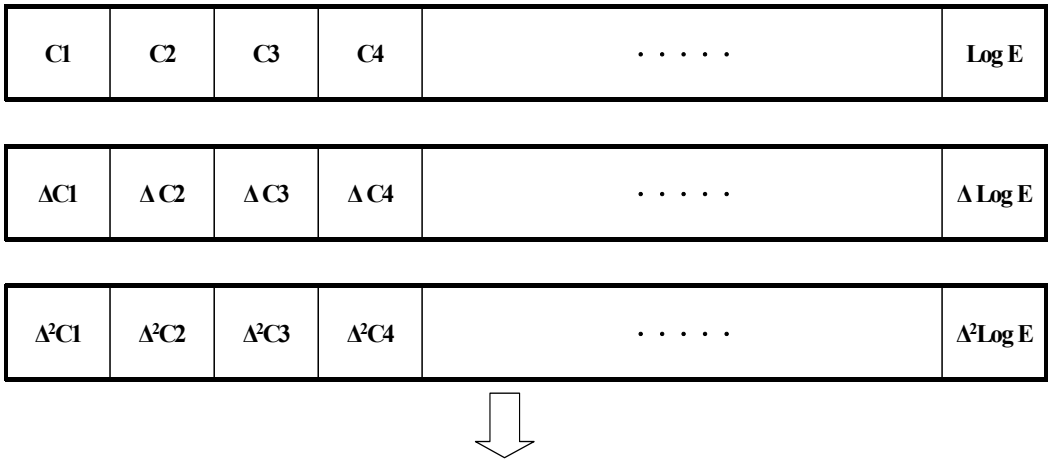


圖４．２　特徵向量及其資料量大小、所需傳輸速度。其中，Ck代表第k個梅爾倒頻譜係數，Log E代表能量頻譜係數取對數值，Δ代表回歸

４．２　特徵向量壓縮的方法－－分割式向量量化（SVQ，Split Vector Quantization）

４．２．１　向量量化的觀念及作法介紹

在前一小節中我們提到了特徵向量做資料壓縮的必要性。那麼，什麼樣的壓縮方式是我們所要的呢？有兩件事情是我們的需求：第一，要能確實壓縮資料內容，使得傳輸速度不大，不會耗去太大的頻寬；第二，在資料壓縮過後，不會造成辨識結果有很大的差異；若能夠進一步提昇辨識正確率又更好。

有幾個基本觀察是值得注意的：

一．在特徵向量中，前十三維的係數是代表音框所轉成的前十二個梅爾倒頻譜係數、以及能量頻譜係數；下一個十三維係數是前面十三維係數的一次回歸；最後的十三維係數又是上一個十三維係數的一次回歸。如果我們可以知道前面十三維的係數，那麼就可以推知另外的二十六維係數；所以，我們可以只傳輸有關前面十三維係數的資訊，在伺服器端再計算出整個的特徵向量；如此一來，我們便可以把傳輸速度降低到原先的三分之一

（４１．６kbps）。若假設在傳輸的過程中沒有發生錯誤，那麼這樣的做法，是不會影響到語音辨識的結果的。（注意！為了方便說明，以下所提到的特徵向量均是十三維）

二． 上述的作法雖然可以降低傳輸速度到原先的三分之一，但是所得到的傳輸速度還是太高（４１．６kbps）。對於特徵向量的前十三個係數，還要另外做向量量化（Vector Quantization）以減少資料量。

在這個小節中，我們先介紹一下，何謂向量量化。假設我們觀察的是N－維的向量，那麼向量量化的目的，就是利用一組從０到M－１的整數（M通常是２的冪次方），來表示所有被觀察到的向量；每一個整數代表某一個特定的N－維向量，換句話說，我們用一些特定的向量，來替代所有被觀察的向量。其中，這些特定的向量叫做碼本向量（Codebook Vector），所對應到的整數值稱做代碼（Codeword），把碼本向量和代碼之間的對應關係收集起來，稱做碼本（Codebook）。就幾何空間的概念上來說，每一個N－維的向量，都算是N－維空間中的一點；假設有一些事先已經選定好、具有代表性且編碼過的點（也就是前述的碼本向量），那麼對某一個向量做向量量化，就相當於選擇一個與該點在幾何空間中最“接近”的

代表點，以代表點的代碼值來代表該點，就如圖４．３所述。要注意的是，這裡所謂的“接近”，牽扯到如何去定義這Ｎ－維向量空間中任兩點的距離；在本報告中，是採用最常用的距離定義方式－－歐氏距離（Euclidean Distance）。當然，還有許多的距離定義方式

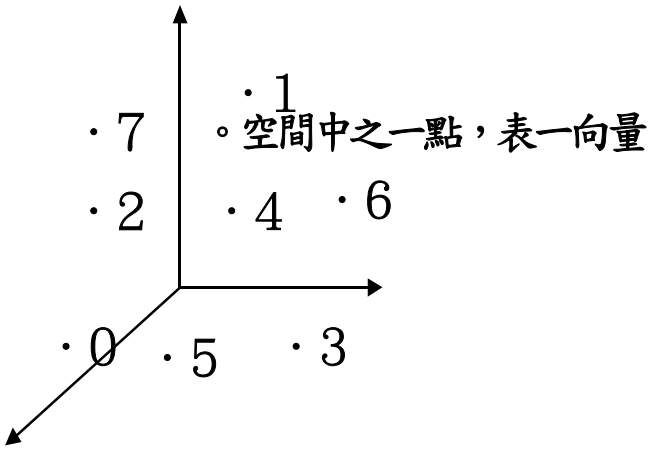


圖４．３　向量的量化。上面的例子是３－維的向量，用８個編碼過的代表點去描述。在上圖中可以看出待量化的向量距離代碼是１的點最近，故在向量量化後，以１來表示這個點。

，在這裡不多做贅述。

在前面的地方曾說過，在向量壓縮過後，我們不希望語音辨識的結果有很大的差異。直覺上來說，如果在經過向量量化過後所得到的

代表向量若和原向量的差異越小，那麼得到的辨識結果也會愈像；也就是說，選出來的碼本向量要越有代表性越好。我們希望提昇向量量化的精確度（Resolution），因此，如何訓練出一組精確度高的碼本是非常重要的。在本報告中，我們使用 K 中心（K－Means）的方式來訓練碼本。以下列出 K 中心方法的要點：（假設觀察的是 N－維的向量，用 M 個點去描述這個向量）

一．定 M 個 N－維向量做為初始的碼本向量，將成這 M 個向量各自獨立作一群，然後分別以０～M－１間的整數來代表之。

二．將所有用來訓練碼本的 N－維向量分成 M 群。分群的標準，是找出被觀察的向量和所有的碼本向量的距離中最小者，將此向量併入該碼本向量所在的那一群。

三．算出每一群的中心（Centroid），並以此中心做為該群新的碼本向量。另外再計算每一群中的每一個向量和該群中的碼本向量的距離和，最後再把每一群的距離和再加總起來，當成是向量量化後的失真。

四．比較本次的失真和前一次的失真：若本次的失真不再比前次的失真來得小、或是小得太少，那麼停止向量量化的動作，每一群的中心即是最後的碼本向量；反之，重複二～三直到本次失真不再比上次失真來得小為止。

４・２・２　分割向量的原因及方法

根據上一節所述，當我們選定碼本的大小 M 之後，使用 K 中心的方法可以幫助我們找到代表１３維特徵向量的碼本。但是，如果想直接對這個１３維的特徵向量作向量量化，會遇到以下的因難：

一・當我們將碼本大小設成一般的大小時，例如 M＝２５６，因為特徵向量多達１３維，因此這樣的向量量化會造成最後的精確度不佳；也就是說，所找出來的碼本向量代表性不足、不“像”原先的向量。

二・若想讓精確度變大，最直接的方式就是加大碼本的大小。不過，因為向量的維度多達１３維，加大碼本的大小會使得訓練碼本的工作太過繁雜，因此加大碼本的大小實際上並不可行。

在上面兩點的討論中，說明了不能直接對１３維的特徵向量做量化的原因，乃是在於其維度太大之故；如果是維度較小的向量，不用太大的碼本大小，就可以做到不錯的準確度。所以為了避免上面所提到的問題。所以我們應該設法減小向量維度的大小。一個可行的方向是：我們可以把１３維的特徵向量切成數個維度比較小的向量，然後再對這些比較小的向量分別作向量量化，最後再把量化後所

得到的小向量結合起來，還原成１３維的向量。 這樣的作法有別於原來的直接作向量量化，因此，我們給這樣的作法一個新的名稱，叫做**分割式向量量化**(Split Vector Quantization)。

那麼我們該怎麼樣分割這１３維的特徵向量呢？從對這１３維特徵向量的基本了解可以給我們一些提示：

一． １３維的特徵向量包括１２維的梅爾倒頻譜係數，另外再加上能量頻譜係數。**由於能量頻譜係數取對數值的關係，所以數值的大小並不如其他１２維來得大**，在做向量量化的時候，應該單獨對這項去做量化。

二． 梅爾倒頻譜係數的特性：**每個音框所求出的２３個梅爾倒頻譜係數，任兩個係數若愈靠近，兩者之間的相關性**(Correlation)**會愈大；反之，若愈遠離，兩者之間的相關性會愈小**。例如：假設我們把所取出的第 K 個梅爾倒頻譜係數記做 C_K，那麼，C_K 和 C_{K-1} 或是 C_K 和 C_{K+1} 的相關性會比 C_K 和其他的梅爾倒頻譜係數的相關性來得高。因此，我們可以把１３維特徵向量中較靠近的梅爾倒頻譜係數取出，做成一個維度較小的特徵向量。

由以上的基本了解，並參照 ETSI 所制定的標準，我們決定了１３維特徵向量的分割方式：將１３維特徵向量分割成七個子特徵向量（Sub－Vector）：第一個子特徵向量，是由１３維特徵向量的第一維和第二維所組成，分別是第一個和第二個梅爾倒頻譜係數（C_1和C_2）；而第二個子特徵向量，是由１３維特徵向量的第三維和第四維所組成，分別是第三個和第四個梅爾倒頻譜係數（C_3和C_4）・・・。以此類推至第六個子特徵向量，是由１３維特徵向量的第十一維和第十二維所組成，分別是第十一個和第十二個梅爾倒頻譜係數（C_{11}和C_{12}）。最後一個子特徵向量，是由能量頻譜係數單獨形成。以上所述如圖４・４所示。

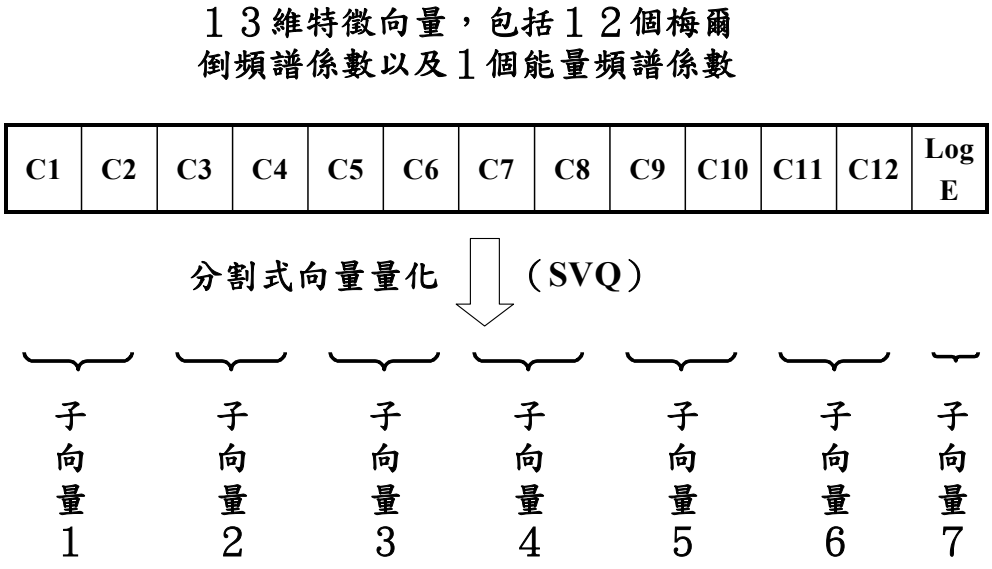


圖４・４　分割式向量量化示意圖。其中C_K是第ｋ個梅爾倒頻譜係數，Log E表示能量頻譜係數取對數值

１３維的特徵向量還具有另一個特性：**次序比較前面的梅爾倒頻譜係數對於辨識來說比較重要**；也就是說，如果m＞n，那麼 C_m在辨識上的重要性會大過 C_n。因此，對於次序比較小的子特徵向量，我們可以設法突顯其重要性以取得較佳的辨識效果。這樣的重要性其實可以反應在給每一個子特徵向量的碼本大小上，我們可以給次序比較小先的子特徵向量一個較大的碼本，而給次序比較後的子特徵向量一個較小的碼本：因為如前面所提到的，加大碼本的大小能使向量量化的精確度提高；若加大某個子特徵向量的碼本大小，在向量量化後該子特徵向量在原來１３維特徵向量的部分會愈趨近於未量化前的該部份；若加強的部分恰好是對辨識較重要的部分，那麼可預期的，量化後的辨識效果和量化前的辨識效果比較不會有太大的差別。**給定各個子特徵向量不同的碼本大小，除了上述的好處以外，另外還可以讓無線通訊的頻寬有效的被使用**——這是因為無線通訊的頻寬有限，故每一秒我們所能傳輸的位元數也是有限的；為了維持一定的辨識正確率，我們應該在可用的位元中分配多一些給對辨識較重要的部分，也就是給較重要的子特徵向量較大的碼本。在本章後面實驗的部分，我們會用不同的碼本大小組合進行實驗，並從中論證以上所述。

４．３　實驗結果：向量量化對語音辨識的影響

在這一節當中，我們將對特徵向量進行分割式向量量化，並以量化後所得的結果做辨識，來探討向量量化的影響。

４．３．１ 實驗架構

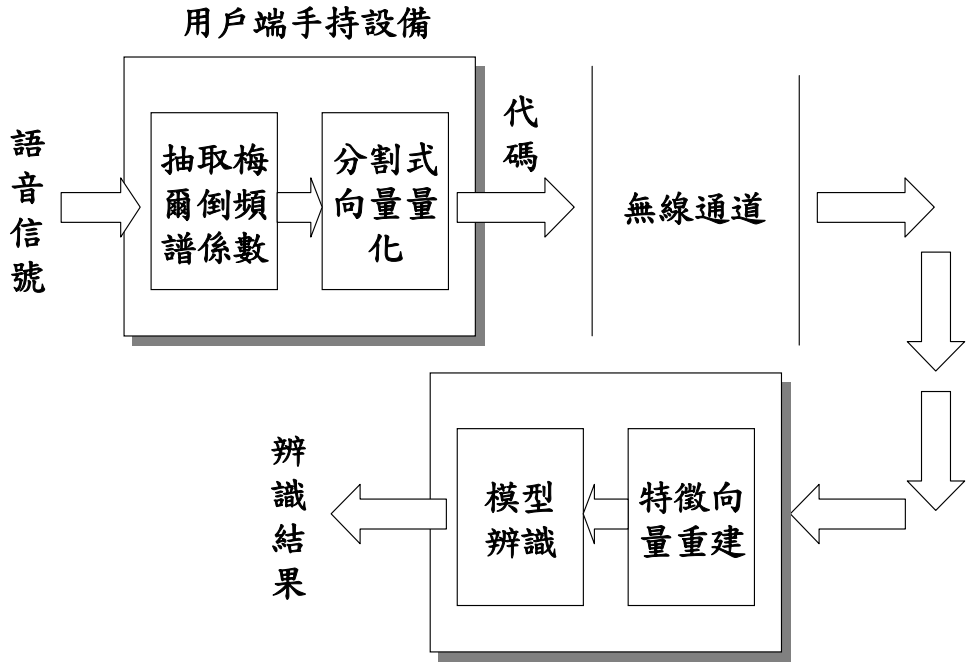


圖４．５　主從式架構之實驗環境配置

如圖４．５所示，是主從式架構的實驗環境配置。在用戶端所持的

手持設備中，所進行的工作是特徵向量的抽取，包括了未量化前的梅爾倒頻譜係數的抽取、以及分割式向量量化的部份；傳送入無線通道的是向量量化後所得的代碼；在伺服器端，會先將接收到的代碼，根據碼本重建回所對應的碼本向量，再以所得到的碼本向量作為特徵向量來辨識。在這一章裡，我們主要想要知道的是向量量化的影響，故我們暫時先假設無線環境中沒有雜訊的干擾，也就是在傳輸的過程當中，不會發生錯誤。

４．３．２　實驗結果及討論

如表４．１所示，這是探討向量量化的影響所得到的中文大字彙連續語音的音節辨識結果。以下的幾個小節，將分別針對在實驗結果中所觀察到的現象進行討論，在這邊我們先對表４．１做說明。在第一欄中所列出的是將１３維特徵向量分割成七個子特徵向量後，所給定七個子特徵向量的碼本大小。例如，在第二列所示“６６６６６６８”，是代表給第一個子特徵向量一個大小為2^6的碼本，給第二個子特徵向量一個大小為2^6的碼本．．．，最後給第七個子特徵向量一個大小為2^8的碼本；第三列所示“７７６６６６６”

碼本大小	是否匹配	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8	傳輸速率（Kbps）
基礎實驗		42.64	49.20	55.31	124.8
6666668	是	39.71	46.92	52.43	4.4
	否	30.33	35.71	40.61	
7766666	是	40.64	48.03	52.88	4.4
	否	30.93	36.59	41.23	
7766668	是	40.14	47.66	52.86	4.6
	否	30.89	36.58	41.24	
8776666	是	40.33	47.74	53.06	4.6
	否	30.87	36.44	41.55	
8777666	是	39.49	47.33	52.92	4.7
	否	31.08	36.65	41.57	
8777766	是	40.28	47.72	53.04	4.8
	否	31.16	36.63	41.67	
8777776	是	40.31	47.59	52.93	4.9
	否	31.21	36.71	41.74	
8877776	是	40.55	47.77	53.50	5.0
	否	31.33	36.83	41.87	

表４・１　實驗結果：向量量化對辨識結果的影響。如表中所示，列出的是在不同的向量量化情況底下，中文大字彙連續語音的音節辨識正確率。

，是代表給第一個子特徵向量一個大小為2^7的碼本，給第二個子特徵向量一個大小為2^7的碼本・・・，最後給第七個子特徵向量一個大小為2^6的碼本。第二欄所示，是訓練語料所取出的特徵向量和測

試語料所取出的特徵向量是否匹配。所謂匹配，是指訓練語料所取出的特徵向量也有經過相同的向量量化的處理，此時測試語料所取出的特徵向量，因為也有經過向量量化，故訓練出來的統計模型可以與之相匹配，反之，不匹配是表示訓練語料有別於測試語料，並未經過向量量化的處理。中間三欄列出不同的高斯混合數。最後一欄則是列出傳送第一欄的碼本組合時所需的傳輸速度；之前我們一再強調無線傳輸時的頻寬限制，因此列出傳輸速度做為衡量效能的另一項依據。最後，在表的第一列所列出的是第三章末曾提到的基礎實驗結果：基礎實驗是所有的特徵向量均沒有經過壓縮，也沒有遭受無線環境雜訊的干擾。因為我們希望經過向量量化之後的辨識效果可以趨近未做向量量化所得到的辨識結果，甚至可以更好，故列出基礎實驗的結果以資比較。

●<u>不匹配情形所造成的影響</u>

在表４．１中第二欄所列，是有關訓練語料及測試語料所取出的特徵向量是否匹配的情形。在第三章中曾提到過，我們使用隱藏式馬可夫模型做為辨識用的統計模型，並使用高斯密度函數做為狀態輸

出函數。由於高斯密度函數是連續機率密度函數，因此雖然測試語料所取出的特徵向量已經經過向量量化的處理，但仍可以用未量化的特徵向量所訓練出來的統計模型來做辨識。如果因向量量化所產生的量化錯誤（Quantization Error）不會太大，理論上不匹配的情形應該不會導致辨識效果變差太多；不過在實際上，因為我們是以中文的大字彙連續語音做為研究對象，所要求的辨識精確度相當高，因一方面我們也可以假設量化錯誤是一個平均分布的隨機變數；故量化錯誤的影響也可能很大，此時就可能造成辨識效果的下降。

那麼究竟匹配與否，會造成多大的辨識效果差異呢？在表４．１由第二列起，我們可以觀察到這項差異在辨識效果上的影響：不匹配的情況所得到的辨識正確率比匹配的情況所得到的辨識正確率足足差了大約１０％～１１％之多（例如，４０％降到３０％）！從實驗的結果我們可以推知量化帶來很大的誤差，因此造成辨識正確率的下降。其實從資料的壓縮也可以看出個端倪：我們將原先需要１２４８個位元來描述的音框，壓縮到只用４４個位元來描述，當然量化錯誤會不容小覷。

為了量化誤差的問題，訓練一個匹配的模型是需要的：我們可以使用語音辨識領域中，對抗雜訊情況的強健性（Robustness For Noise）常用的模型分析來說明訓練匹配的模型的必要性：假設未經量化的特徵向量是 $\underline{V}$，$\underline{V}$ 經過特徵向量量化後得到的新特徵向量是 $\underline{V'}$，又量化的誤差向量（Error Vector）記做 $\underline{n}$，

則 $\underline{V} = \underline{V'} + \underline{n}$ ，

即 $\underline{V'} = \underline{V} - \underline{n}$ ，

此時 $\underline{V'}$就好比是乾淨的特徵向量加上一個雜訊 $\underline{n}$，因此若我們想要辨識一個被雜訊干擾的特徵向量，必須也用被雜訊干擾的特徵向量來訓練一個模型，如圖４．６所示。

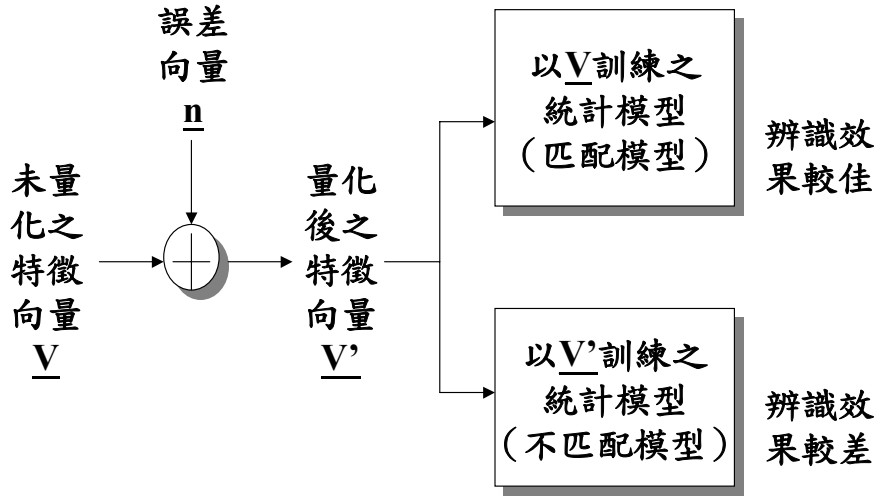


圖４．６　匹配與不匹配情形之等效模型。在本圖中可以看出一個匹配的模型的需求

最後要附帶說明的是，特徵向量在經過向量量化之後已轉成數目有限的碼本向量，因此在實際上是可以改用離散的機率函數來做為隱藏式馬可夫模型的狀態輸出函數；但是考慮下列情況：假設在向量量化後，用來訓練音節 X 的特徵向量組合為：{$\underline{V_1}$，$\underline{V_2}$，．．．．，$\underline{V_N}$}，若測試語料中所抽出了一個特徵向量 $\underline{U}$，其對應的音節為 X，但卻未出現在上述的特徵向量組合中，那麼若使用離散的機率函數做為狀態輸出函數，就必然產生一些誤差。因此我們仍然使用高斯密度函數做為狀態輸出函數。

●**不同的子特徵向量所反應出重要性的不同**

碼本大小	是否匹配	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8	傳輸速率（Kbps）
6666668	是	39.71	46.92	52.43	4.4
	否	30.33	35.71	40.61	
7766666	是	40.64	48.03	52.88	4.4
	否	30.93	36.59	41.23	
7766668	是	40.14	47.66	52.86	4.6
	否	30.89	36.58	41.24	
8776666	是	40.33	47.74	53.06	4.6
	否	30.87	36.44	41.55	

表４．２　實驗結果：不同的子特徵向量所反應出重要性的不同

上表４・２是表４・１的部分擷取。表４・２的第一列以及第二列，分別是碼本大小“６６６６６６８”及“７７６６６６６”所得的辨識結果，第三列及第四列則是碼本大小“７７６６６６８”及“８７７６６６６”所得的辨識結果；以上兩組具有共同的特點，那就是兩組用的傳輸速度都是相同的，第一組是４・４Kbps，第二組則是４・６Kbps。

在４・２・２中，我們曾經討論到不同的子特徵向量，在語音辨識上所表現出來的重要性也不一樣，而表４・２的兩組數據正反應出這樣的事實。在兩組數據中，因為傳輸速度相同，所以對於各組中不同碼本的配置，最後用來描述每個音框的位元數是固定的：第一組是用４４個位元來描述每個音框，第二組是用４６個位元來描述每個音框。在第一組中，”６６６６６６８”是給定前面的六個子特徵向量同樣的碼本大小＝2^6＝６４，最後一個子特徵向量給定碼本大小＝2^8＝２５６，而“７７６６６６６”是給定前面兩個特徵向量碼本大小＝2^7＝１２８，後面五個特徵向量給定碼本大小＝2^6＝６４；也就是說：“６６６６６６８”強調最後一個子特徵向量（能量頻譜係數）的重要性，但“７７６６６６６”強調的是前面二個特徵向量（梅爾倒頻譜係數前四個C_1～C_4）的重要性。由實驗

結果中可以看到，高斯混合數＝２、４、８時，加強前面的子特徵向量可以使辨識結果均得到些微的改進。同樣的在第二組數據（“７７６６６６８”及“８７７６６６６”）中，若把“７７６６６６８”用來描述第七個子特徵向量的８個位元分出兩個，一個給第一個子特徵向量，另一個給第三個子特徵向量，使碼本安排變為“８７７６６６６”，在表４．２中可以看到，除了在高斯混合數＝２、不匹配的情況是稍稍變差，其他的情況都是辨識效果有得到改進。因此，我們經由實驗證實了之前提到的對特徵向量的基本了解：梅爾倒頻譜係數中，次序愈小的係數對於辨識愈重要。

●**傳輸速率和辨識結果的關係**

如圖４．７，我們畫出了在各個傳輸速率下，所得到的辨識正確率的變化；其中圖４．７（Ａ）是在高斯混合數＝２時的變化情形，圖４．７（Ｂ）是在高斯混合數＝４時的變化情形，圖４．７（Ｃ）是在高斯混合數＝８時的變化情形。在表４．１當中有兩個傳輸速度的實驗，有兩種不同的碼本組合，得到兩種不同的結果；於是我們將兩個實驗結果做平均，得到單一個值做為在該傳輸速

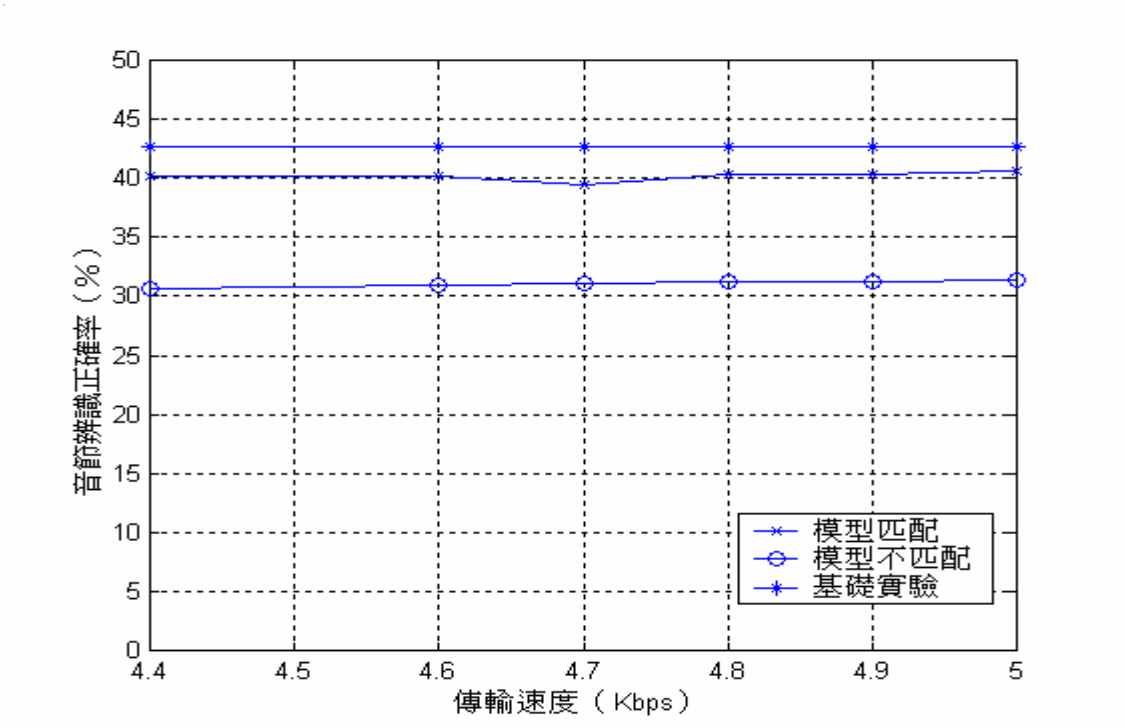


圖４．７（Ａ）　在高斯混合數＝２時，傳輸速度和音節辨識正確率的關係。

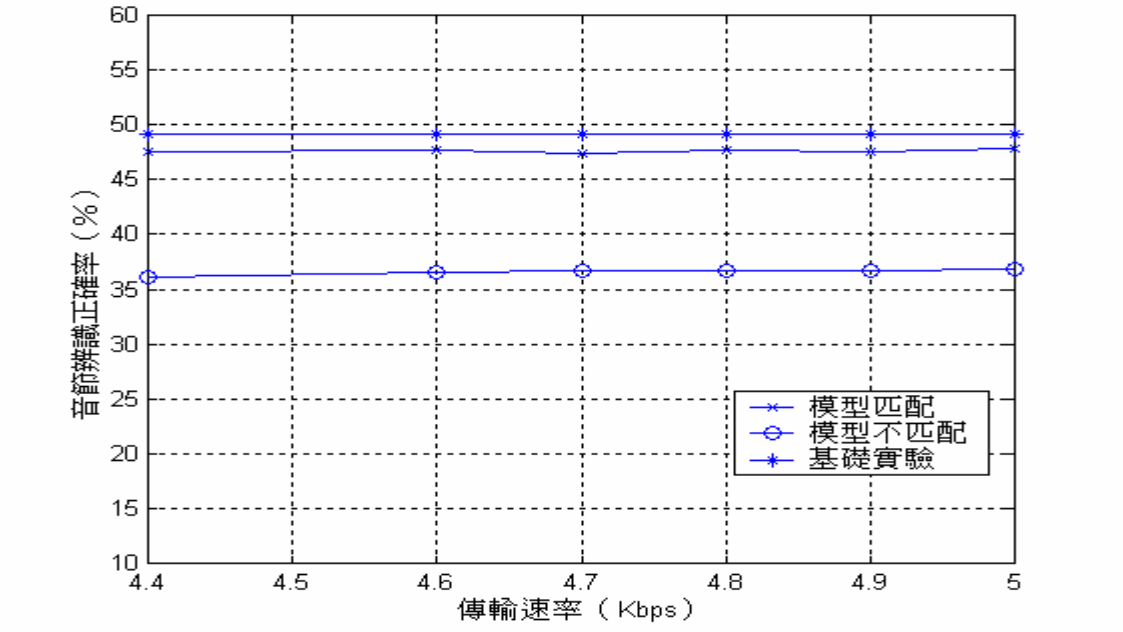


圖４．７（Ｂ）　在高斯混合數＝４時，傳輸速度和音節辨識正確率的關係。

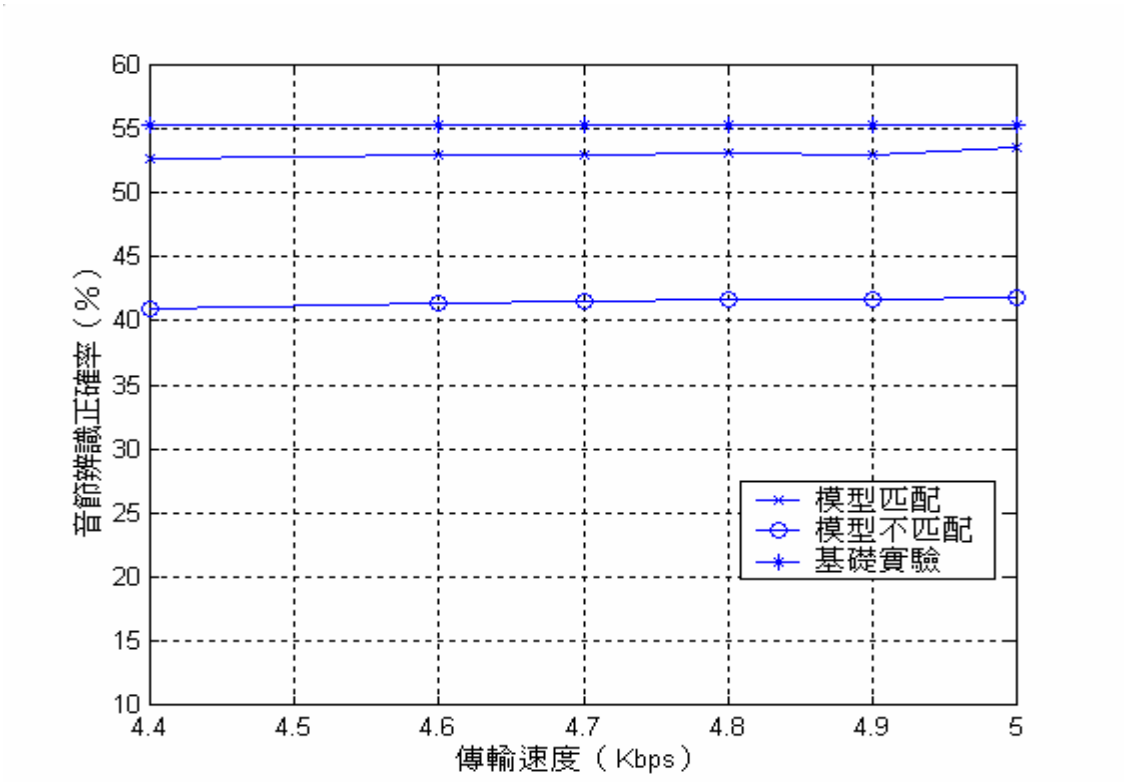


圖４．７（Ｃ）　在高斯混合數＝８時，傳輸速度和音節辨識正確率的關係。

度下的辨識結果。

在三張圖中，我們都可以觀察到一個趨勢：隨著傳輸速度的增加，音節的辨識正確率也隨之微幅的增加。特別是在模型不匹配的情況下，除了在高斯混合數＝４，４．７Kbps～４．８Kbps 間辨識正確曾有少許的下降（３６．６５%變成３６．６３%）以外，所觀察到的都是持續上揚的趨勢。至於模型匹配的情況，在傳輸速率變快的過程中，雖然辨識率會有跳動，但大致上還是呈現上揚的趨勢。

會有這樣的趨勢其實不難理解，因為傳輸速率的增加，意味著用來描述每個音框的位元數增加，故特徵向量量化的精確度也跟著增加，所造成的影響是：

一．對於模型不匹配的情況，因為用來訓練模型的特徵向量未做量化，會讓向量量化的精確度提高，這樣一來量化後的向量和原先的向量差別會愈小，而得到的辨識效果也可以更趨近未做量化前的辨識正確率。

二．對於模型匹配的情況，因為用來訓練模型的特徵向量已做過量化，所以訓練出來的統計模型也不相同，因此我們認為：最後的辨識結果呈現跳動，應該是統計上些許的差異。不過整體來說，提高向量量化的精確度，會讓量化後的特徵向量更具鑑別力，也就是說，各個音節的區別也更加明顯。故隨著傳輸速度增加，辨識正確率仍然大致呈現上升的趨勢。

因為傳輸速度會影響辨識效果，那麼我們可以事先選定兩種不同的傳輸速度（當然，這兩種傳輸速度必須對應到可以接受的辨識正確率），建立一個可適性（Adaptive）的機制，由伺服器所在的接收端評估當時通訊情況是否是良好的，再把評估的結果告知用戶端所持的手持設備；最後，手持設備再根據被告知的內容執行以下兩種

選擇：

一． 如果在通訊狀況良好、允許我們使用較多的頻寬的時候，則手持設備就加大用來做向量量化的碼本、然後傳送較多的代碼來提升辨識正確率。

二． 如果通訊狀況不佳，通訊系統需要較多的編碼以保護資料內容，造成我們可以使用的頻寬較少時，手持設備就選擇一個所需傳輸速率較低且辨識效果可以接受的碼本組合，來進行向量量化的工作。

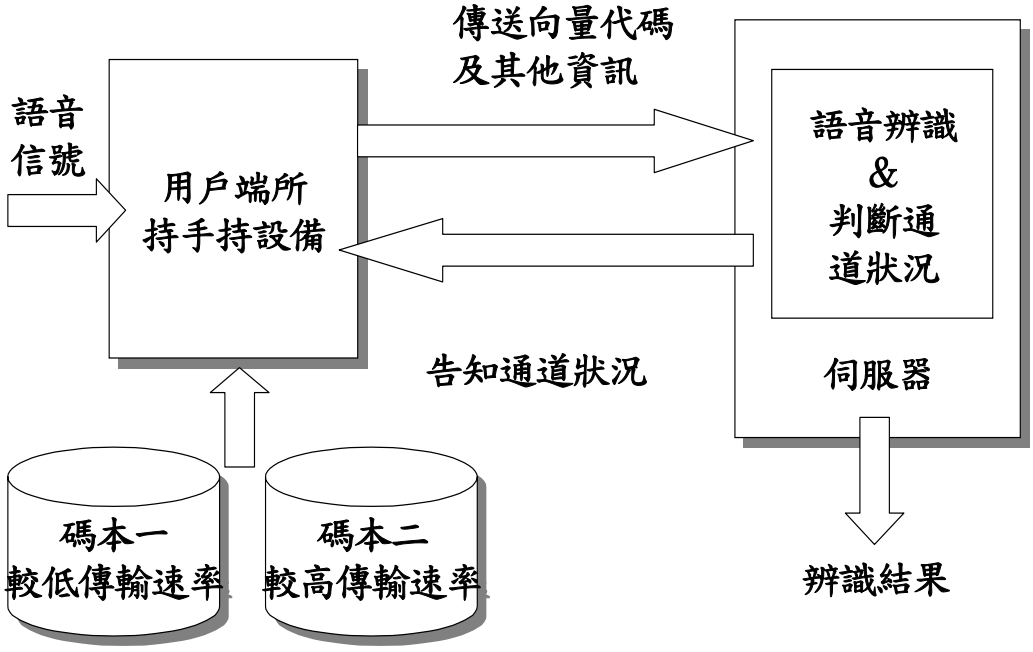


圖４・８ 可適性機制示意圖

以上所述如圖４・８所示。

●**量化後辨識正確率些微的下降**

在圖４・７（Ａ）（Ｂ）（Ｃ）中，另外標示出基礎實驗（未經過向量量化、且假設無線通道中沒有發生錯誤）的結果。在三個圖中，我們可以很清楚的觀察到，在不匹配的情況下，特徵向量經過向量量化後的辨識效果不如基礎實驗所得到的；若是模型匹配，辨識正確率就會有所提升，大約是２～３％的下降。因為我們的模型是用來辨識中文的音節（Syllable），在中文的音節中，有許多的音節，在聲學上的特性相似，所以我們認為，在經過向量量化之後，對於音節和音節間的區別將比較失去鑑別性，如此一來就會造成辨識率的下降。

為了觀察向量量化使那些音素在區分時產生混淆，我們將基礎實驗所得到的辨識結果，與特徵向量經過量化後的辨識結果去做比較，辨識結果有所不同之處，即是容易產生混淆的音素。比對之後我們發現，上述容易混淆的音素包括了“ㄉ”和“ㄍ”、“ㄑ”“ㄒ”、”ㄣ”和“ㄥ”、“ㄧ”和“ㄩ”、“ㄚ”和“ㄠ”、捲舌音跟不捲舌音等等。像上面所描述的這些音素，就算是未做量化直接做對其做辨識，也是容易混淆的。

像這些容易混淆的音素，或許在單音節的辨識當中無法清楚的分別，但是在整個語音辨識中，我們後級尚有語言層次的處理：對於辨識出來的國語單音節，我們可以考慮詞彙及前後文的關係，並且以那些音素容易造成混淆做為背景知識，對發生錯誤的音素部分做修正，以得到合於文法、語意的結果。因此上述的音素混淆，在接下來利用語言知識做處理的過程當中可望獲得修正；也就是說，原先在音節部分的辨識正確率在後級將可以獲得補償。因此，我們可以預期的是，在中文的整體辨識率上，應該和未做量化前的效果相去不遠。基於以上所述理由，我們可以得到一個結論：在本章中所提出的分割式向量量化的方法，在對語音的特徵向量做過處理後，並不會使辨識效果有太大的退步。

4．4 結論

在主從式架構中，用戶所持的手持設備把辨識用的特徵向量抽取出來，經由無線通道後傳到伺服器端做辨識；因為通道的頻寬有限，所以必須要先將特徵向量做向量量化後才能傳送出去。但特徵向量的大小共１３維，若直接對它做向量量化有實際上的困難，所以本

報告使用特殊的向量量化方式－－分割式向量量化，先將特徵向量分割成數個子特徵向量後，再分別對這些向量做量化。最後，我們得到本章的重要結論，那就是我們所提出的壓縮方式，只帶給音節辨識正確率些微的下降；因此，本章所提出的分割式向量量化的方法是可行的。

第五章　主從式架構（二）：無線通道環境對語音辨識結果的影響

在主從式架構中，為了在有限頻寬的無限通道中傳送特徵向量，因此我們必須對特徵向量做向量量化；在第四章中，我們利用中文大字彙連續語音的音節辨識實驗，來找出向量量化對辨識結果的影響。我們的發現是，辨識正確率受到量化錯誤的影響，將無可避免的下降。除了量化所帶來的錯誤，在無線傳輸的過程當中，通道的不完美所帶的雜訊也會造成錯誤。因此在本章中，我們將接著對無線通道中的雜訊所產生的錯誤來進行探討，並找出可行的錯誤補償方式。

5．1　無線通道環境可能發生的錯誤

在無線通道傳輸的過程當中，因為通道的不完美，常常會造成許多雜訊的干擾，使得傳輸出去的資料因為遭遇到這些干擾因而發生錯誤。錯誤發生的原因有很多，舉例來說，有以下這幾項：

一．熱雜訊（Thermal Noise）的干擾

在元件中，電子因為具有能量會自然的擾動，反映在信號上，就成為不規律跳動的雜訊，這樣的雜訊就稱作熱雜訊。熱雜訊常以白色雜訊的型態出現，也就是說，在頻譜上各種頻率的信號都將受到白色雜訊的干擾；同時，因為熱雜訊必然存在在各種電子元件當中，因此在我們所討論的無線傳輸環境，同樣也受到熱雜訊的干擾。

二．其他訊號的干擾

在無線通道的環境中，使用相同頻道傳輸的信號不只有一個，這些信號彼此之間是靠著傳送端的調變（Modulation）使得各個信號呈現正交（Orthogonal）的情況；而在接收端，若要知道每個信號原先傳輸的內容，必須要能和傳輸端所送出的信號同步（Coherent）才能確實的解回每個信號。如果同步的處理作得不好，那麼在解出的信號當中，就會摻雜著其他同頻信號的成分，這種情況就稱作同頻干擾（Co－Channel　Interference）。另外，鄰頻信號如果強度夠大，對於真正想要接收的信號也會產生干擾，這種情況則稱作臨頻干擾（Inter－Channel Interference）。

三．通道本身的不完美

任何通道都有先天上不完美的地方，無線通道也不例外。通道

不完美表現的方式，可能是有頻寬的限制，因此造成在某些頻帶上信號的衰減；也可能是頻道本身的特性轉換函數不完美，造成信號強度（Amplitude）或是相位（Phase）的失真（Distortion）。

四．**多路徑衰減（Multi－Path Fading）**

這是無線通訊中最特別、也是問題最大的部分。在都市裡有許多的建築物，這些建築物對於傳輸資訊的無線電波來說，相當於一個一個的反射面；因此信號從無線通道的過程當中，可能經過許多不同的反射，最後接收端所接收的信號，往往是好幾個經由不同路徑的信號的加總；信號加總的結果，可能是加強信號的強度，因此未嘗不是件好事；但是信號的加總也可能是衰減信號的強度，此時接收端可能就因為信號過於微弱而產生判斷內容上的錯誤，像這樣信號忽強忽弱的現象，我們稱作衰減（Fading）。一般來說，衰減出現時，如果接收端一直處在信號過低的地方，這時候接收到的信號將產生一連串的錯誤，這也就是所謂的群集錯誤（Burst Errors）。如不在都市而在鄉間，傳輸環境不同，問題略不一樣，但衰減現象也一樣會發生。

在我們的實驗架構當中，主要是針對傳輸端所送出的資料內容作處

理。其中，資料內容包括前一章所述在向量量化後所得到的碼本代號（Codeword），資料檔頭（Header），保護用的編碼等等的位元串流（Bit Stream）。因為是對位元串流作處理，所以我們不管在實體層中（Physical Layer）中雜訊的表現方式，我們只管雜訊對位元串流所造成的影響。在上述的幾種雜訊中，我們最關心的是白色雜訊，以及多路徑頻道中衰減所造成的影響：因為白色雜訊的產生是無所不在，而衰減帶來的群集錯誤會使通訊的品質大打折扣。這兩者對位元串流的影響是：白色雜訊帶來隨機錯誤（Random Error），而衰弱帶來的群集錯誤會使得位元串流中一連串的位元出錯。在以下的部分我們將對兩種錯誤進行模擬，同時藉由實驗結果來觀察這兩種錯誤帶給辨識效果的影響。圖５．１是敘述無線通道發生錯誤時所帶來的影響。

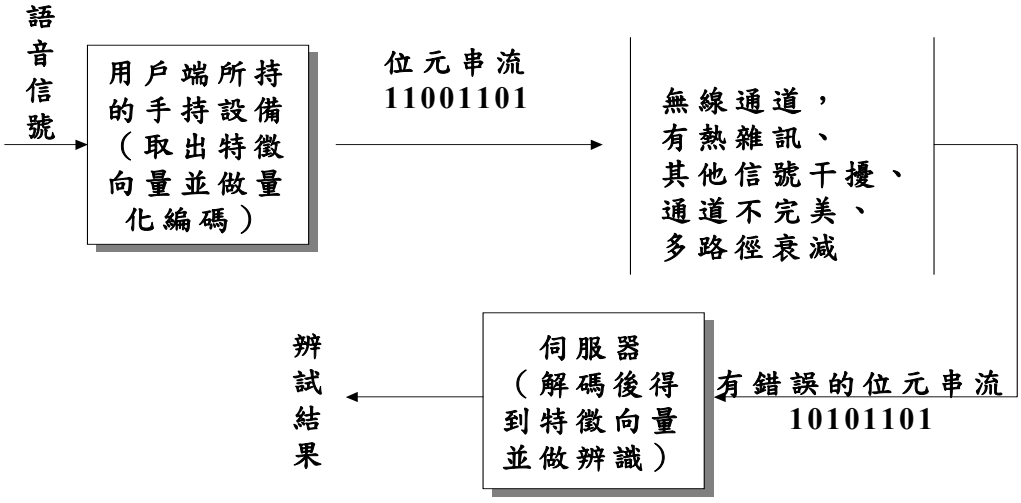


圖５。１　無線通道發生錯誤時反應在位元串流的情形

5．2　無線通道錯誤的模擬(一)－－隨機錯誤（Random Errors）

5．2．1　隨機錯誤的產生原因

如前一節所提到，在無線通道中如果雜訊的來源是白色雜訊時，那麼從用戶端所持的手持設備所傳送出來的位元串流（其中包括我們對特徵向量做向量量化後所得到的代碼（Codeword）、有關資料內容的檔頭（Header）、為了保護資料所做的編碼等等），將會發生隨機錯誤。另外，當通道中發生群集錯誤時，傳輸端如果對傳送出去的位元串流有做交錯分散（Interleaving）的處理，同時在接收端做反交錯分散（De－Interleaving）的處理，那麼群集錯誤會被打散，最後所得到的位元串流就看不到大量發生的錯誤。因此，當我們使用隨機錯誤來模擬無線通道中發生的錯誤情形時，我們是基於以下兩種假設中任一種：

一．通道中雜訊的來源是白色雜訊；

二．在傳輸時有做交錯分散的處理，讓可能發生的群集錯誤被打散了。

在５．２這一節當中，我們將探討：若使用主從式架構的分散式語音辨識系統，當隨機錯誤發生的時候，其辨識效果將受到多大的影響；我們希望知道，不同的位元錯誤率將導致辨識效果的差異有多少。

５．２．２ 隨機錯誤的產生方式

為了模擬接收到的位元串流發生隨機錯誤的情況，我們做了以下的推導：

一．假設每一個位元發生錯誤的機率固定，均設為 x，那麼每個位元發生錯誤的事件都是一個白努利試驗（Bernoulli Trial）。

二．因為每個位元發生錯誤的事件是互相獨立的，而且發生錯誤的機率均相同，故發生錯誤的位元總數會是一個二項式分布（Binomial Distribution）的隨機變數，而這個隨機變數的期望值，是所有位元的數目，乘以每一個位元發生錯誤的機率。若假設在位元串流中，所有的位元數目為 N，則該期望值為 Nx。

三・位元錯誤率（Bit Error Rate）

＝　發生錯誤的位元數／總位元數，

故位元錯誤率 ＝　（Ｎｘ）／Ｎ　＝　ｘ，

則位元錯誤率　＝　每個位元發生錯誤的機率

故由上面的推導可以知道，如果要在位元串流中產生隨機錯誤，那麼只要把每一個位元發生錯誤的事件，當作是獨立的白努利試驗，並且將白努利試驗的發生機率定為最後接收端所看到的位元錯率。

５・２・３　實驗流程

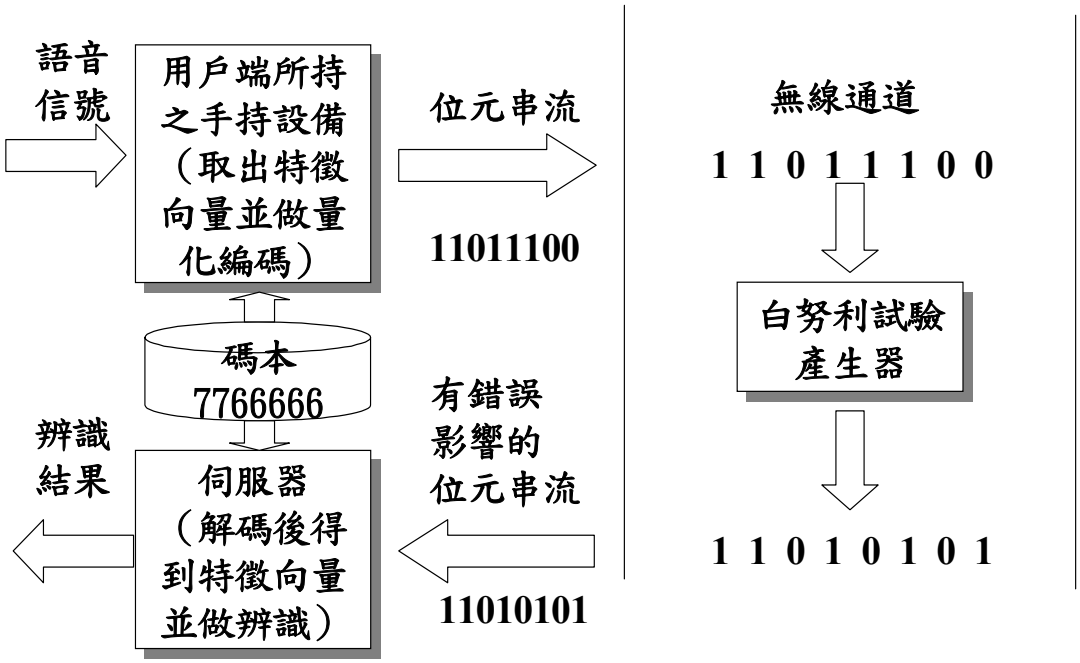


圖５・２　位元串流發生隨機錯誤的實驗流程

整個實驗的流程如圖５．２所示。在主從式的架構下，用戶端所持的手持設備對語音信號取出辨識所需的特徵向量，接著對特徵向量進行第四章所述的分割式向量量化，然後將量化後所得到的位元串流（量化後的代碼）送到接收端；在接收端的前級，我們對接收到的位元串流模擬產生隨機錯誤，然後由接收端的伺服器將挾帶錯誤的位元串流還原成所對應的碼本向量，最後進行辨識。

上述過程有兩點要附帶說明：

一，對於分割式向量量化碼本組合，我們選定的是“７７６６６６６”的碼本大小。選擇的原因是：這樣的碼本組合使用４４個位元來描述一個特徵向量，符合 ETSI 所制定資料傳輸速度４．４Kbps（加上資料的檔頭共４．８Kbps）的標準；並且在第四章探討不同的傳輸速度所對應音節辨識正確率的實驗中，這個碼本組合所表現出的效能也不錯，因此我們選定這個碼本組合做為向量量化的碼本。

二．在第四章中，我們同時也看到了若用來訓練模型的特徵向量和用來做測試的特徵向量不匹配，那麼辨識正確率會大大的衰減；因此在下面的實驗中，我們一律使用匹配的統計模型來做辨識。

根據上述的實驗架構，我們希望找出，在不同的位元錯誤率下，所得到的音節辨識正確率是多少，並藉此評估隨機錯誤對語音辨識的影響。

5・2・4 實驗結果：隨機錯誤對語音辨識的影響

位元錯誤率	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8	附加說明
0	42.64	49.20	55.31	基礎實驗
0	40.64	48.03	52.88	
10^{-5}	40.64	48.03	52.87	
10^{-4}	40.57	47.94	52.84	
10^{-3}	40.21	47.60	52.36	
10^{-2}	35.94	43.05	47.66	

表5・1 實驗結果：隨機錯誤對辨識結果的影響。如表中所示，列出的是在不同的位元錯誤率底下，中文大字彙連續語音的音節辨識正確率。

上表5・1列出在無線通道發生隨機錯誤時，在各種不同的位元錯誤率下，中文大字彙連續語音的音節辨識正確率。在第一欄中所列出的，是各種不同的位元錯誤率：這一項的是所有發生錯誤的位元

數，除以所有的位元總數而得來，我們取１０$^{-2}$～１０$^{-5}$之間的值；中間三欄是在不同的高斯混合數下，所得到的辨識正確率。另外，在表中的第一列所列出的是未經量化的基礎實驗的結果，第二列所列出的是在傳輸端用”７７６６６６６”的碼本組合做向量化、但在無線通道中沒有錯誤干擾時的辨識正確率；列出這兩列是為了要做比較：以第一列做基準可以看出當發生傳輸錯誤時，因無線傳輸所帶來辨識正確率的總和影響（量化錯誤＋傳輸錯誤），由第二列可以看出單單傳輸錯誤所帶來的影響。以下我們先將上面的數據製圖如圖５．３，接下來再討論實驗結果。

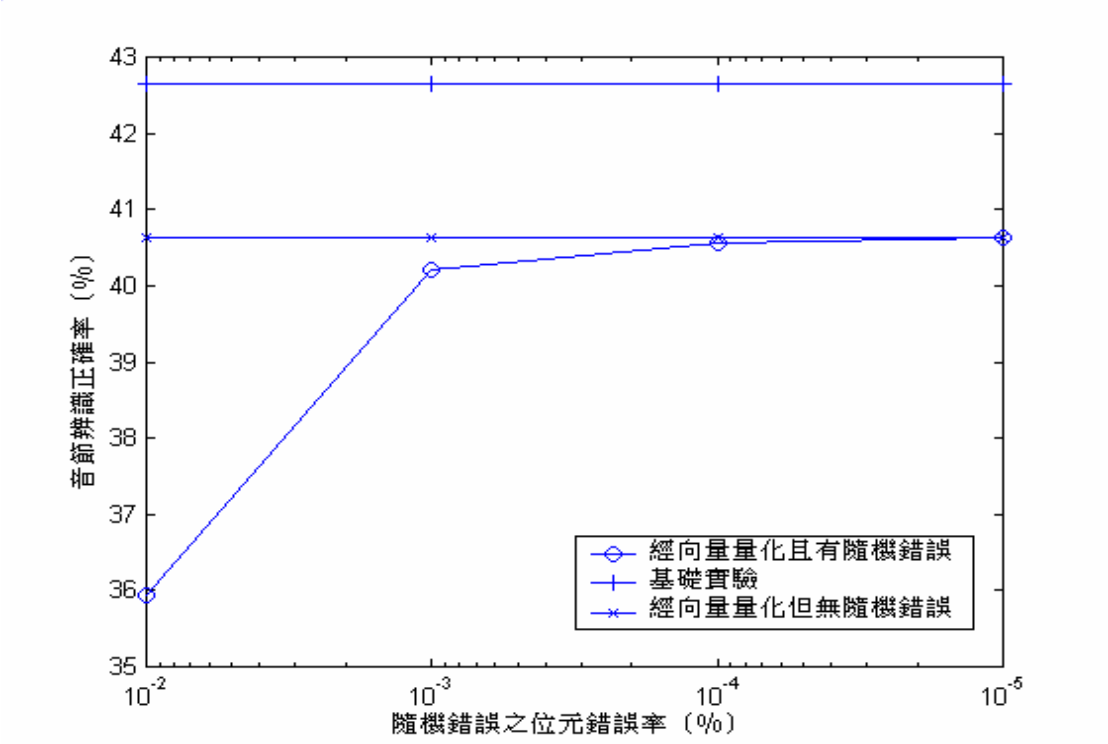


圖５．３（Ａ）　在高斯混合數＝２時，發生隨機錯誤時位元錯誤率和音節辨識正確率的關係。

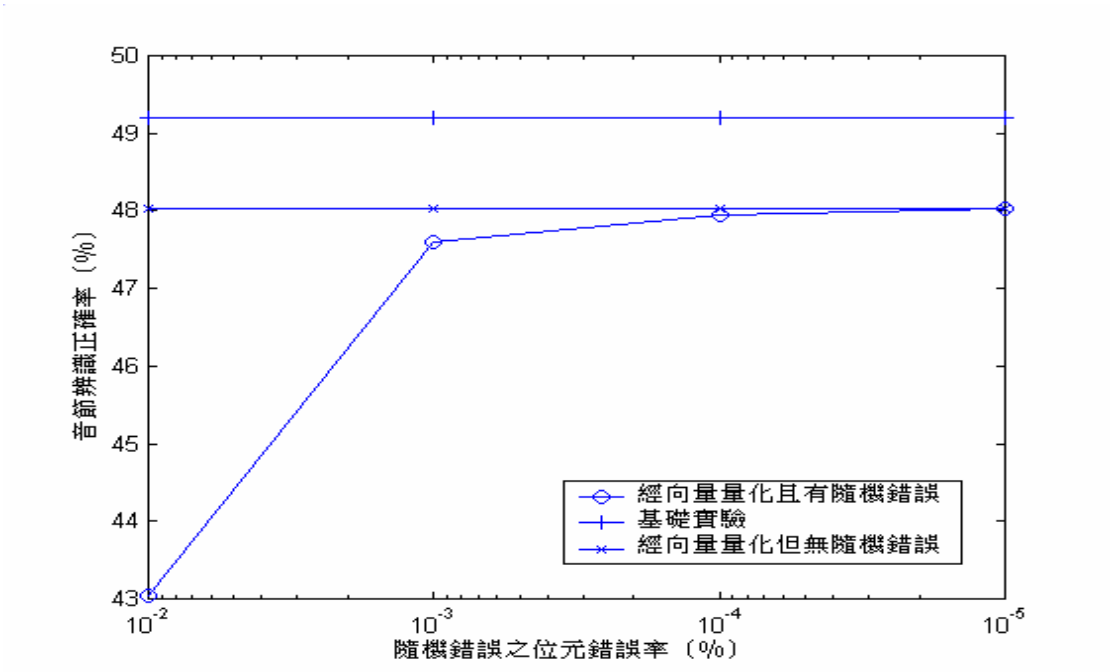


圖５．３（Ｂ）　在高斯混合數＝４時，發生隨機錯誤時位元錯誤率和音節辨識正確率的關係。

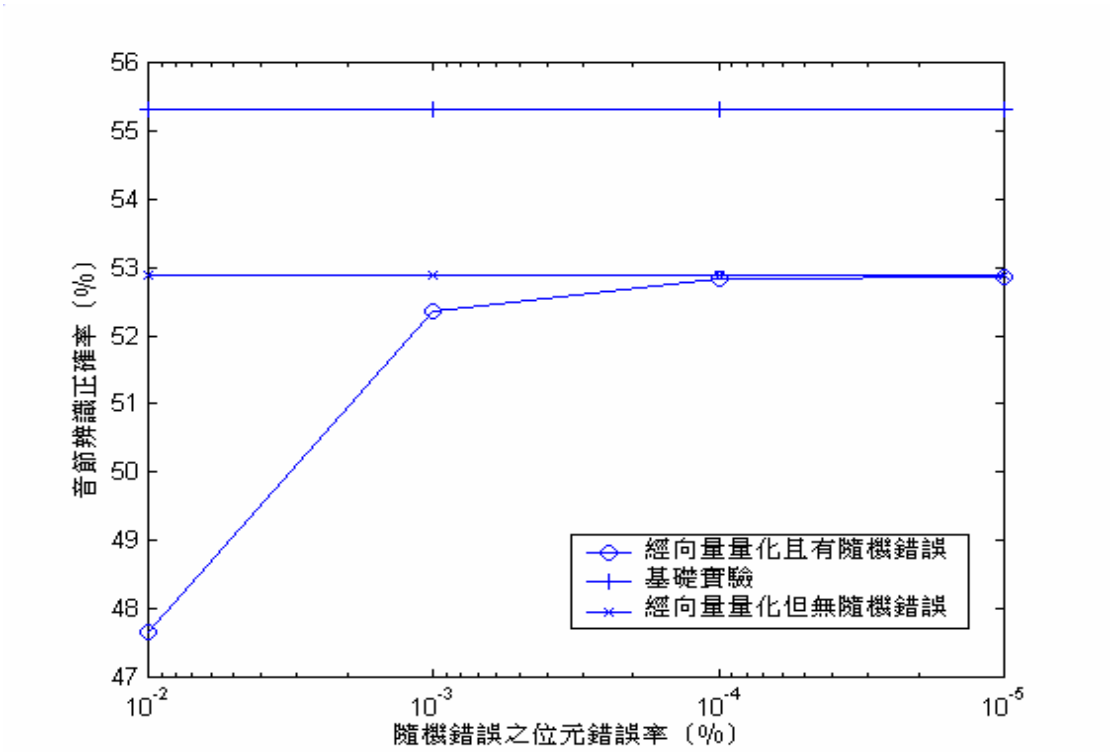


圖５．３（Ｃ）　在高斯混合數＝８時，發生隨機錯誤時位元錯誤率和音節辨識正確率的關係。

●**隨機錯誤之位元錯誤率與辨識正確率的關係**

上面的三張圖中各列出三條線：第一條是在發生隨機錯誤的情況下，不同的位元錯誤率所對應的辨識正確率；第二條是基礎實驗所得的結果；第三條是沒有發生隨機錯誤時的情況。如前所述，列出第二條和第三條是為了比較之用：第一條線和第二條線中的間距，代表發生因無線傳輸所帶來的辨識效能下降；第一條線和第三條線中的間距，則是單獨代表因隨機錯誤所帶來的辨識效能下降。若間距愈大，表示辨識效能所受的影響愈大。三張圖中，觀察第一條線和第三條線中的間距，不難發現：當高位元錯誤率的時候（約10^{-2}），單獨考慮隨機錯誤的因素，辨識正確率大約下降５％。但是當位元錯誤率下降到10^{-3}的時候，此時辨識正確率就只有微幅的、約０．５％的下降，如果位元錯誤率更進一步下降到10^{-4}、或是10^{-5}的低位元錯誤率時，此時辨識正確率就幾乎沒有下降了。

以上所得到的結果，由以下的分析可以得到解釋。

一．在我們所使用的語料中，一個字的平均時間長度大約是０．５秒。在第三章曾提到，對於音節的描述，我們是採用「聲母加上界音、韻母模型」的方式來描述；假設「聲母加上界音」以

及「韻母」兩個次音節單位的時間長度一樣長，那麼上述兩部分音素的平均時間長度就是０．２５ｓ。

二．由語料中所取出的音框，其時間長度是１０ｍｓ（０．０１ｓ），故對於每一個次音節單位，取出的音框平均數大約為

０．２５／０．０１＝２５音框／次音節單位，

又一個音框取出一組特徵向量，故對於每一個次音節單位，取出的特徵向量平均數即為２５特徵向量／次音節單位。

三．而在主從式架構中，對於每一個特徵向量，在向量量化後我們用４４個位元來描述其所得到的量化代碼（使用“７７６６６６”的碼本），

因此，當位元錯誤率＝10^{-2}時，平均每２．２７個特徵向量會發生一個錯誤，在每個次音節單位中有２５／２．２７≒１１個特徵向量會發生１個錯誤；當位元錯誤率＝10^{-3}時，平均每２２．７個特徵向量會發生一個錯誤，在每個次音節單位中有２５／２２．７≒１個特徵向量會發生１個錯誤；當位元錯誤率＜10^{-3}時，則在每個次音節單位中不到１個特徵向量會發生錯誤；

四．當次音節單位所對應的特徵向量發生改變的時候，那麼所得到的辨識結果就有可能發生改變。如圖５．４所示，當發生辨識

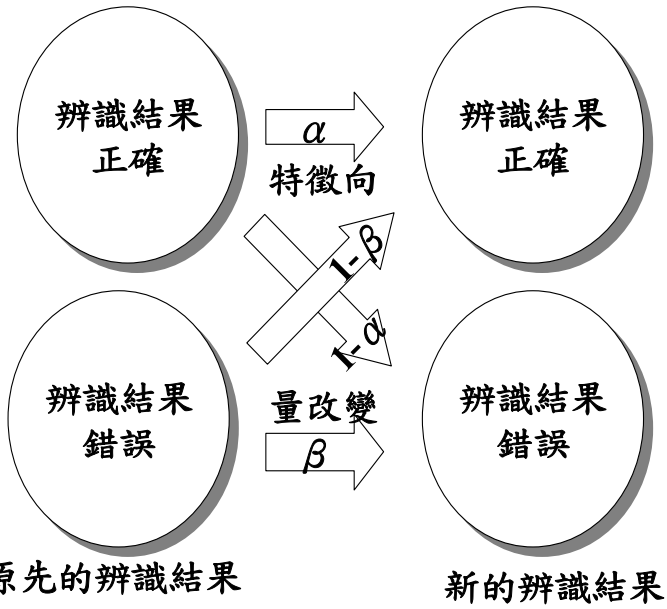


圖５．４　特徵向量改變後辨識結果的變化情形。其中箭頭上的α代表原來辨識正確，新的辨識結果也正確之機率，β代表原來辨識錯誤，新的辨識結果也正確之錯誤

結果改變時，共有下列幾種情形：

第一種情形是，原先的辨識結果是正確的，而新的辨識結果也是正確的，我們假設發生這樣情況的機率是α；

第二種情形是，原先的辨識結果是正確的，而新的辨識結果是錯誤的，發生這樣情況的機率是$1-\alpha$；

第三種情形是，原先的辨識結果是錯誤的，而新的辨識結果是也是錯誤的，我們假設發生這樣情況的機率是β；

第四種情形是，原先的辨識結果是錯誤的，而新的辨識結果是正確的，發生這樣情況的機率是$1-\beta$。

上述的α、β值很明顯的和特徵向量改變的總數有關。在一個

次音節單位所對應的２５個特徵向量中，若特徵向量改變的數目愈多，原先辨識正確的愈有可能辨識錯誤，因此α會愈小，對β的影響也是一樣，因此原先辨識錯誤的愈有可能辨識正確。只不過變動的程度會有所不同：對於一個已知的次音節單位而言，辨識錯誤的可能性總共有４２８種，辨識正確的可能性只有１種（所有次音節單位總數為４２９）。因此α值要比β值大得多，所以我們可以很合理的假設：**在特徵向量改變時，只有α值會小於１，β值可假設為１；而且特徵向量改變得愈多，α值變得愈小**。

五．由第三點以及第四點可以得知，因為在位元錯誤率＝10^{-2}時，每個次音節單位中平均發生錯誤的向量個數多達１１個，α值下降的幅度必然很大；但在位元錯誤率＝10^{-3}時，每個次音節單位中平均發生錯誤的向量個數只有１．１個，α值不會比１小太多，若位元錯誤率更小時，那麼α值會更接近１。因此，當位元錯誤率＝10^{-2}時，因為α變小，造成辨識效能的大量下降；但是在位元錯誤率＝10^{-3}、或是更低的時候，α的改變不大，因此辨識結果跟沒有發生錯誤時，並不會有太大的差別。

●**<u>隨機錯誤對音節辨識的影響小於向量量化所帶來的影響</u>**

在前一小節曾提到：圖５．３中，第一條線和第三條線中的間距，代表因為無線通道中發生隨機錯誤而造成的辨識效果下降。另外，第二條線和第三條線的差距代表因為向量量化的量化誤差所導致的辨識效果下降。若第一條線和第三條線中的間距，比第二條線和第三條線中的間距要來得大，代表隨機錯誤的影響大過量化錯誤的影響；反之，則是量化錯誤的影響大過隨機錯誤的影響。由圖５．３可以知道，在位元錯誤率大於10^{-3}時，隨機錯誤的影響大過量化錯誤的影響，但在位元錯誤率小於10^{-3}時，隨機錯誤的影響就極小，主要的錯誤來自於量化錯誤。在無線通訊中，要將位元錯誤率降到10^{-3}以下其實是並不困難的，也就是說，在一般的狀況底下，通道若發生隨機錯誤，對於整個音節辨識結果影響並不是很大；主要的辨識率下降，還是取決在因向量量化所帶來的量化錯誤——也正因為量化錯誤還是主要的辨識正確率下降的來源，更顯出如何找到一個具有代表性的碼本的重要。由以上所述，在５．２這一節中，我們得到的結論是：**在主從式對於音節的辨識，若致力使整個通訊系統的位元錯誤率維持在10^{-3}以下，那麼我們就可以忽略通道中的隨機錯誤所帶來的影響；此時若想進一步改善辨識正確率，**

主要應從如何使碼本最佳化著手。

5．3　無線通道錯誤的模擬（二）－－群集錯誤（Burst Errors）

群集錯誤的發生，對於行動通訊來說，通話的品質將會大打折扣；但是當群集錯誤發生在傳送語音資訊的位元串流時，那麼對於伺服器端要執行的辨識服務會有怎樣的影響呢？在５．３這一節中，我們將模擬產生群集錯誤，並觀察群集錯誤帶給語音辨識的影響。

5．3．1　群集錯誤的產生方式

在本報告中，我們參考【１０】來產生群集錯誤。根據位元發生錯誤的群集性，位元錯誤可以分成二類（如圖５．５所示）：

一．群集錯誤（Burst Errors）

發生的位元錯誤並非單獨存在，必定是兩個以上。

二．單一錯誤（Single Errors）

發生的位元錯誤是單獨存在，前後並無相鄰的位元錯誤。

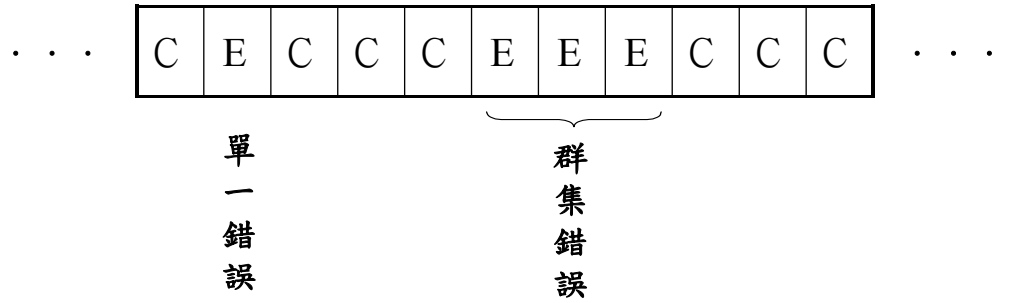


圖５．５　位元錯誤分類示意圖。其中C代表無誤，E代表位元發生錯誤

因此，我們可以把每個接收到的位元其所在的狀態可以分成下列三個狀態：

一．G狀態，代表接收到的位元並沒有發生錯誤；

二．S狀態，代表接收到的位元正發生單一錯誤；

三．B狀態，代表接收到的位元正發生群集錯誤。

如果可以描述從一個狀態跳至另一個狀態的機率，也就是說，若可以描述三個狀態間的轉移機率（Transition Probability），再加上如果可以知道在B狀態中會停留幾個位元（群集錯誤的位元錯誤數目），那麼就可以用這三個狀態所建立的馬可夫鏈（Markov Chain）來模擬產生群集錯誤。

關於這三個狀態間轉移機率，以及在B狀態中所停留的位元數目：

一．假設由G狀態跳到S狀態的機率是ａ，由G狀態跳到B狀態的機率是ｂ，那麼由G狀態回到G狀態的機率就會是１－ｐ－ｑ。

二．在S狀態下，下一個轉移到的狀態必為G狀態；因為如果S狀態回到S狀態或是跳到B狀態的話，那就會產生群集錯誤，而不是S狀態所對應的單一錯誤了。

三．我們規定在B狀態下，下一個轉移到的狀態必為G狀態；也就是發生群集錯誤時，在B狀態下停留的位元數是群集錯誤的位元錯誤數。另外，群集錯誤的位元錯誤數是由泊松（Poisson）隨機變數來描述；也就是每進入一次B狀態即產生一個泊松隨機變數，用該變數值來做為在B狀態停留的位元數。

泊松隨機變數的機率：

$$P_k = P(X = k) = \frac{e^{-\lambda} \times \lambda^k}{k!} \qquad (5-1)$$

其中，X代表泊松隨機變數，λ是泊松隨機變數的平均，k是非負整數。此處需要注意的是，根據上面的定義，在B狀態停留的位元數必是大於等於2的正整數，所以我們必須對泊松隨機變數做調整，這部分在下面的地方會有詳細的說明。

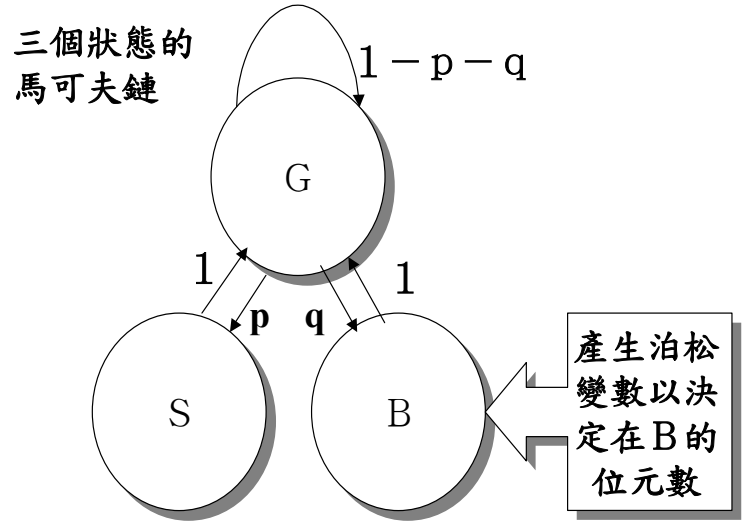


圖５・６　三個狀態的馬可夫鏈。在狀態和狀態間的轉移所標示的是轉移機率。

接下來就是參數的決定了。在上述的模型當中，所需決定的參數值有ｐ、ｑ、以及泊松隨機變數的平均。我們先定義三個相關的機率值：P_G、P_S、P_B，其中P_G代表在G狀態時的機率，其中P_S代表在S狀態時的機率，其中P_B代表在B狀態時的機率。在上述的馬可夫鏈達到平衡時，

$$\begin{aligned} P_S &= p \times P_G \\ P_B &= q \times P_G \end{aligned} \quad , \qquad (5-2)$$

因此，

$$P_G = \frac{1}{1+p+q} \quad , \qquad (5-3)$$

我們先來考慮，當在B狀態出現的群集錯誤之位元數目。在前面已

經提到過，我們使用泊松隨機變數來描述群集錯誤之位元數目，但是因為在Ｂ狀態時出現的群集錯誤數目必大於或等於２，因此泊松隨機變數必須做修正：修正的方式就是忽略變數值＝１或變數值＝０的情形，所得到修正後的泊松隨機變數之機率即是將原先的泊松隨機變數之機率做正規化的動作，則修正後的泊松隨機變數之機率為

$$P_{k'} = P_k \frac{\sum_{k=0}^{\infty} P_k}{\sum_{k=2}^{\infty} P_k} = \frac{1}{1-e^{-\lambda}(1+\lambda)} \frac{\lambda^k e^{-\lambda}}{k!}$$

（５－４）

其中ｋ是大於２或等於２的正整數，且

$$\sum_{k=2}^{\infty} P_{k'} = 1 \text{ 。}$$

（５－５）

我們可以計算修正後的泊松隨機變數期望值λ’，則

$$\lambda' = \frac{\lambda(1-e^{-\lambda})}{1-e^{-\lambda}(1+\lambda)} \text{ ,}$$

（５－６）

λ’的意義，就是在Ｂ狀態時的群集錯誤之位元錯誤數目平均值，因此，若定義P_e為位元錯誤率時，則P_e和P_S、P_B的關係是

$$P_e = P_S + \lambda' \times P_B \text{ ,}$$

（５－７）

由（５－２）知，

$$P_e = p \times P_G + \lambda' \times q \times P_G ,$$ （５－８）

由（５－３）知，

$$P_e = \frac{p + \lambda' q}{1 + p + q} 。$$ （５－９）

如前所述，模型的建立要決定的參數共有ｐ、ｑ、λ（泊松隨機變數的期望值），這些值的決定方式如下：

一．透過對衰減（Fading）的平均時間長度的推導，再將此時間換算成所對應的位元數，就可以得到λ’：λ’決定了，就可以利用（５－６）來計算λ。

二．我們可以事先選定我們想要觀察的位元錯誤率，然後利用（５－９）以及前面得到的λ’，可以得到ｐ、ｑ之間的關係。

三．定義群集程度Ｂ（Burstness）

＝（發生群集錯誤的位元數）／（發生錯誤的位元數），

由（５－８）知，

Ｂ＝（λ’ｑ）／（ｐ＋λ’ｑ）（以百分率來表示），

則利用Ｂ以及在上一步所得到ｐ、ｑ的關係式可以得到ｐ、ｑ的值。

當上述的參數決定了之後，我們就可以利用３個狀態的馬可夫鏈來

模擬產生錯誤密集度不一的群集錯誤了。

5・3・2 實驗流程及參數設定

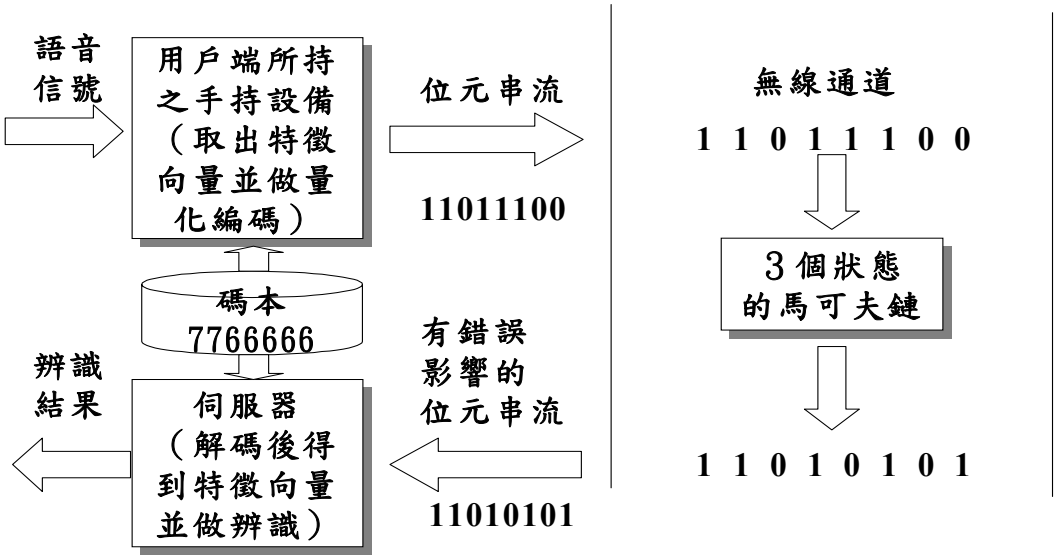


圖5・7 位元串流發生群集錯誤的實驗流程

整個實驗的流程如圖5・7所示，基本架構和位元串流發生隨機錯誤的實驗並沒有太大的差別，唯一的差異是無線通道中模擬產生隨機錯誤的白努力試驗產生器換成了5・2中所介紹的3個狀態的馬可夫鏈。

在上面的實驗流程確定之後，接著我們必須決定3個狀態的馬可夫鏈的三個參數。根據上一節所述，第一個必須要決定的參數是λ，

也就是泊松隨機變數的期望值。我們可以先決定衰弱的平均時間長度。造成衰弱的原因是信號強度太弱，我們可以透過公式【１１】計算，當信號強度是方均根值的某個倍數時，，平均衰弱持續的時間（Average Fade Duration）$\tau_{\bar{r}}$ 是

$$\tau_{\bar{r}} = \frac{e^{\bar{r}^2} - 1}{\bar{r} f_m \sqrt{2\pi}}, \qquad (5-10)$$

其中

$$\bar{r} = \frac{r}{r_{RMS}}, \qquad (5-11)$$

r 是信號的強度，r_{RMS}則是代表信號強度的方均根值，這一項是計算信號強度和其方均根值的比值；

而 f_m是最大都卜勒平移（Maximum Doppler Shift），

$$f_m = f_c \times \frac{v}{c}, \qquad (5-12)$$

其中 f_c是載波頻率，v 是手持設備的移動速度，c 是光速。有關上述各個參數的設定：

一．載波頻率方面，我們使用ＧＳＭ（Global System for Mobile Communication）其中一種的載波頻率９００ＭＨｚ；

二．因為衰減發生的原因是因為信號強度過於弱小，此時如果手持設備的移動速度過於緩慢，容易停在收訊不良的地方而使得衰減的情況特別嚴重。為了強調衰減的效果，我們假設手持設備

是以人行走的速度移動，大約是５公里／小時；

三．假設信號強度是其方均根的１／１００，即２０分貝（ｄＢ）的衰減。

由以上所述得到的平均衰減持續的時間＝９．５８ｍｓ。因為負責傳輸信號的位元串流是由０和１所組成，即使在衰減期間也可能有一半的機會可以“猜”中信號內容，因此我們將要換算成群集錯誤的平均值時，我們先將上面所求出來的值除以２，得４．７９ｍｓ。接著我們將４．７９ｍｓ換算成群集錯誤的平均值：因為我們使用４４個位元來描述１個特徵向量，而１個特徵向量是由１個音框所取出，１秒鐘共產生１００個音框，因此４．７９ｍｓ就會等於２１．０７６個位元所佔的時間長度。這個就是λ’的值。再利用（５－６），得到λ的值也是２１．０７６。

接著我們決定ｐ、ｑ之值。ｐ、ｑ值的決定需要上面所求出的λ’、位元錯誤率（P_e）、以及群集程度（Ｂ）。在下面的實驗當中，我們想知道：在不算太小的位元錯誤率（10^{-2}或10^{-3}）時，群集錯誤的五種群集程度（Ｂ）０％、２５％、５０％、７５％、１００％對辨識效果的影響，並以此作圖觀察其中趨勢。因此根據上述情況，表５．２列出在不同的位元錯誤率、不同的群集程度下所對應

的p、q值。

	群集程度 ＝0％	群集程度 ＝25％	群集程度 ＝50％	群集程度 ＝75％	群集程度 ＝100％
位元錯誤率 $=10^{-2}$	p＝0.0101 q＝0	p＝0.0076 q＝0.0001	p＝0.005 q＝0.0002	p＝0.0025 q＝0.0004	p＝0 q＝0.4747
位元錯誤率 $=10^{-3}$	p＝0.001 q＝0	p＝0.7506 $\times10^{-3}$ q＝0.0119	p＝0.5003 $\times10^{-3}$ q＝0.0237	p＝0.2501 $\times10^{-3}$ q＝0.0356	p＝0 q＝0.4745 $\times10^{-4}$

表5・2　不同的位元錯誤率、不同的群集程度所對應的p、q（三個狀態的馬可夫鏈之轉移機率，p是G到S的轉移機率，q是G到B的轉移機率）值。

5・3・3　實驗結果：群集錯誤對語音辨識的影響

下表5・2是群集錯誤對語音辨識之影響的實驗結果。表5・2（A）是位元錯誤率＝10^{-2}的情形，5・2（B）則是位元錯誤率＝10^{-3}的情形。在兩張表的第二欄列出的是群集程度，第三、四、五欄則是列出在不同的高斯混合數下，中文大字彙連續語音的音節辨識正確率。在表中的第一、二列同時列出基礎實驗及只經過向量量化但沒有通道錯誤的實驗結果；列出這兩列的目的是為了比

較之用。

位元錯誤率（%）	群集程度 B（%）	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8
0	基礎實驗	42.64	49.20	55.31
0	0	40.64	48.03	52.88
10^{-2}	0	33.15	39.72	45.04
10^{-2}	25	34.25	41.13	46.12
10^{-2}	50	34.96	42.34	47.00
10^{-2}	75	35.17	41.90	46.91
10^{-2}	100	35.62	42.66	47.39

表５・２（Ａ）　位元錯誤率＝１０$^{-2}$時，在各種不同的群集程度下，中文大字彙連續語音的音節辨識正確率。

位元錯誤率（%）	群集程度 B（%）	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8
0	基礎實驗	42.64	49.20	55.31
0	0	40.64	48.03	52.88
10^{-3}	0	39.92	47.33	52.24
10^{-3}	25	40.08	47.55	52.35
10^{-3}	50	39.97	47.44	52.36
10^{-3}	75	39.95	47.17	52.06
10^{-3}	100	39.64	46.95	51.77

表５・２（Ｂ）　位元錯誤率＝１０$^{-3}$時，在各種不同的群集程度下，中文大字彙連續語音的音節辨識正確率。

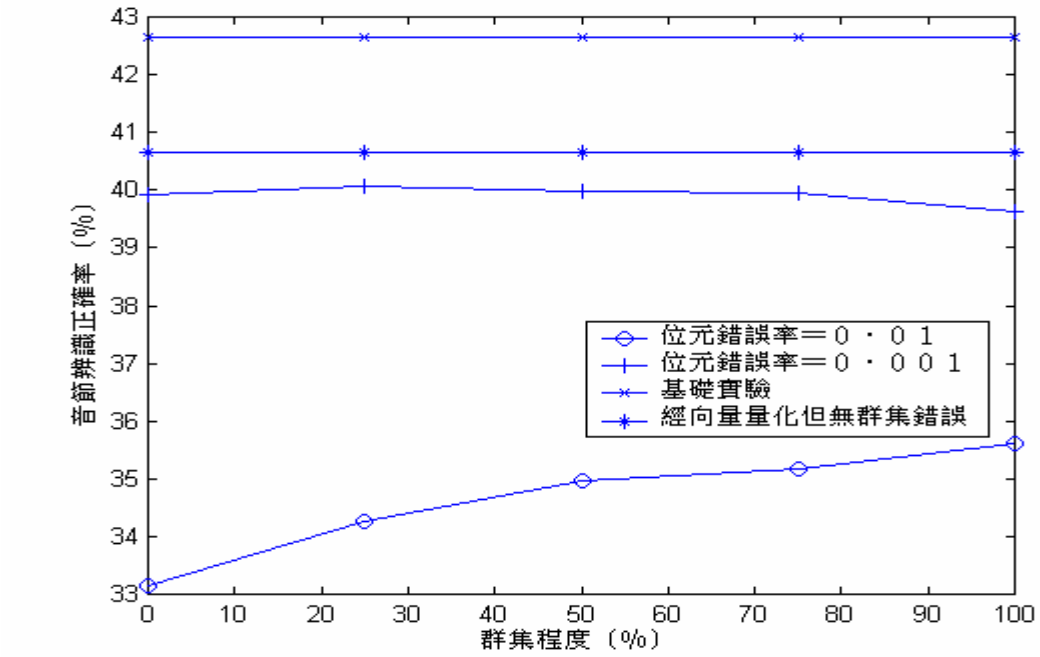


圖５・８（Ａ）　在高斯混合數＝２時，發生群集錯誤時群集程度和音節辨識正確率的關係。

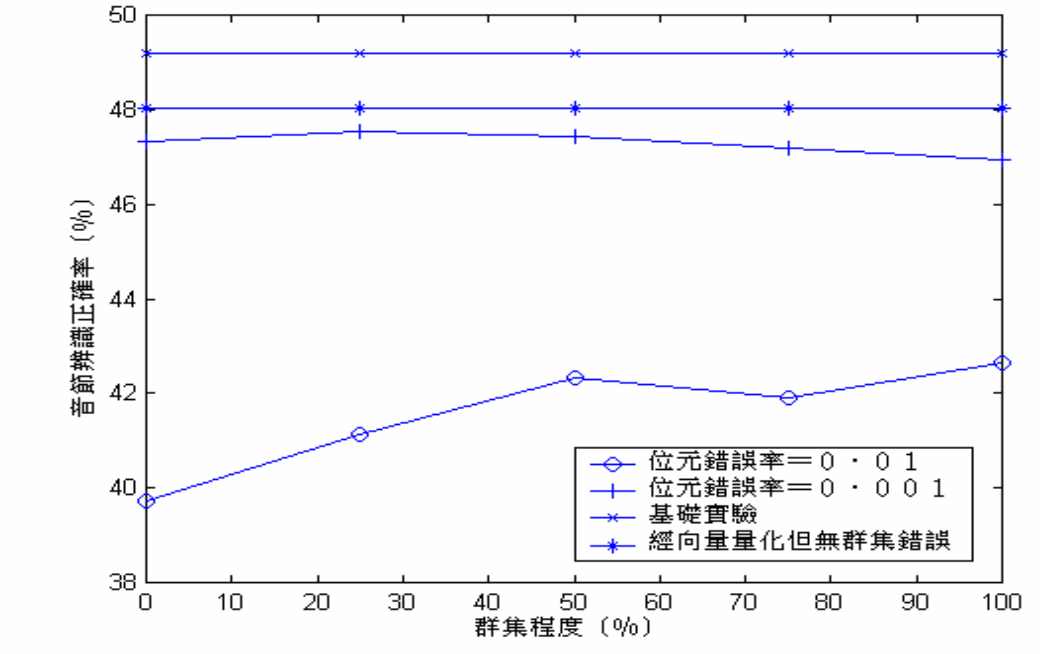


圖５・８（Ｂ）　在高斯混合數＝４時，發生群集錯誤時群集程度和辨識正確率的關係

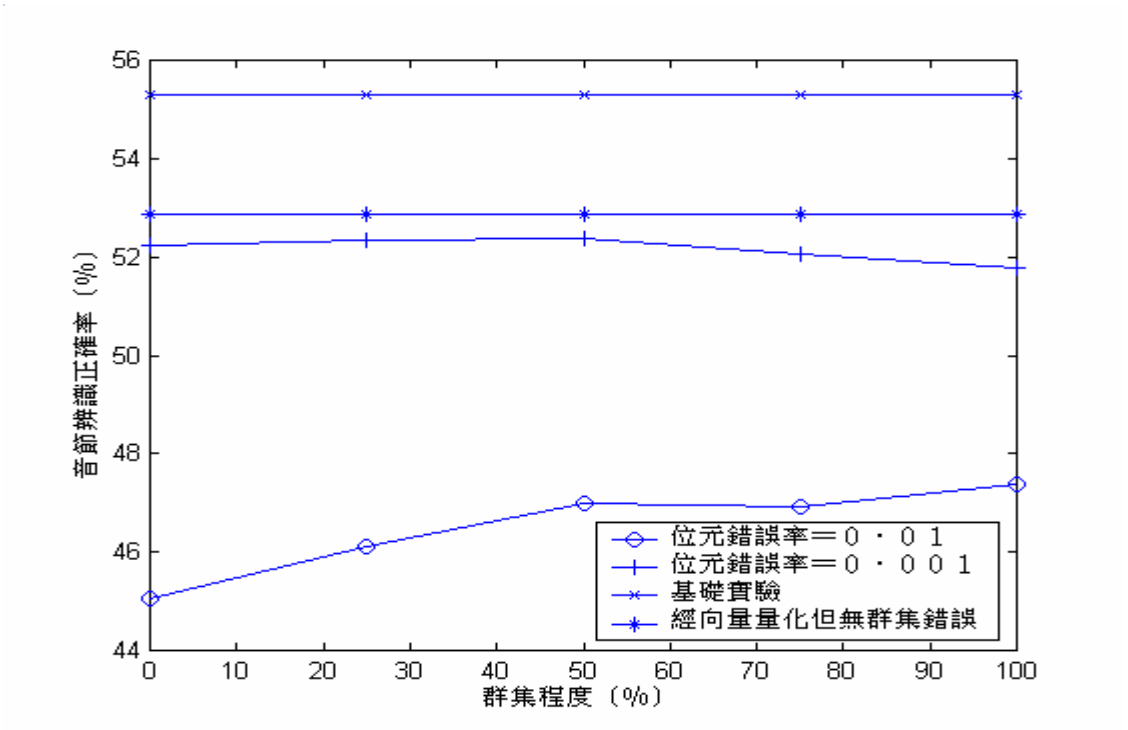


圖５・８（Ｃ）　在高斯混合數＝８時，發生群集錯誤時群集程度和音節辨識率的關係

我們將先對表５・２所得到的實驗結果製圖（圖５・８），然後我們再針對實驗數據、以及圖５・８中所看到的趨勢來進行討論。在上面的三張圖中共四條折線。標示位元錯誤率＝10^{-2}及10^{-3}，是代表在上述兩個位元錯誤率時不同的群集程度所對應的音節辨識正確率；另外標示基礎實驗以及經向量量化但無群集錯誤的，是用來和上述兩條線做比較用。以下是對於實驗結果的討論。

●隨機錯誤和群集錯誤模擬環境的不同

在上一節討論隨機錯誤對辨識結果影響的實驗中，結果顯示當隨機錯誤的位元錯誤率＝10^{-2}時，高斯混合數＝２、４、８時的辨識正確率分別是３５．９４％、４３．０５％、４７．６６％；當隨機錯誤的位元錯誤率＝10^{-3}時，高斯混合數＝２、４、８時的辨識正確率分別是４０．２１％、４７．６０％、５２．３６％。以上的這些數據，若對照圖５．８中相對應的位元錯誤率折線，可以發現除了５２．３６％這個數據在其所對應的位元錯誤率＝10^{-3}、群集程度＝５０％處有出現過以外，其他的５個數據均未在所對應的折線上出現過。這顯示**隨機錯誤的實驗環境和群集錯誤的模擬環境並不相同**，尤其是對於隨機錯誤的模擬及群集程度＝０％的模擬，我們很容易將這兩者的情形視為相同，但事實上，群集程度＝０％表示完全沒有位元錯誤群集在一起，所有的位元錯誤都是這一節一開始所定義的單一錯誤，但是隨機錯誤的模擬卻可以看到群集錯誤的發生，顯示這兩種情況是不相同的。而在不同的群集程度下的模擬情況，也不會和隨機錯誤的相同；因為群集錯誤的模擬將平均群集錯誤長度設為定值２１．０７６，但隨機錯誤中發生群集錯誤的平均長度卻不會那麼大（當隨機錯誤的位元錯誤率＝10^{-2}或

10^{-3}時，要使位元串流連續發生２１個錯誤位元的可能性是非常小的）。

●**不同的位元錯誤率下，群集程度對辨識正確率影響的趨勢也不同**

觀察在圖５．８中不同的位元錯誤率所對應的趨勢：在位元錯誤率＝10^{-2}時，隨著群集程度的增加，辨識正確率大致上隨之提升；在位元錯誤率＝10^{-3}時，隨著群集程度的增加，辨識正確率大致上隨之減少。在不同的位元錯誤率下，出現截然不同的趨勢。在無線通訊的情況裡，群集錯誤愈密集，所帶來的是收訊情況愈差，因此我們也預期，若是群集程度愈大，辨識正確率的下降會愈多。但是在實驗結果上卻看到兩種不同的趨勢，那麼造成這種情況的原因何在呢？

回想在上一節中隨機錯誤對辨識結果影響的實驗討論：我們用４４個位元來描述特徵向量所對應的碼本代碼，也就是說，每一個特徵向量是用４４個位元來描述；另外，對於一個次音節單位（我們訓練統計模型的單位）而言，其所佔的時間長度大約對應２５個特徵

向量，因此每個次音節單位大約是用１１００個位元來描述。

一．當位元錯誤率＝10^{-2}時，每個次音節單位所對應的１１００個位元中平均會出現１１個錯誤；

二．當位元錯誤率＝10^{-3}時，每個次音節單位所對應的１１００個位元中平均會出現約１個錯誤。

因此，

一．當位元錯誤率＝10^{-2}時，每個次音節單位中平均有１１個發生錯誤的位元：這１１個位元如果平均的分布在次音節單位所對應的２５個特徵向量中，那麼會造成１１個特徵向量的錯誤；但是如果這１１個位元集中在一起，那麼只會造成１個特徵向量的錯誤（１個特徵向量是由４４個位元來描述，可以容納得下１１個錯誤）。又，在上一節中隨機錯誤對辨識結果影響的實驗討論曾提過，如果在一個次音節單位中的特徵向量發生錯誤的位元愈多，那麼這個次音節單位發生錯誤的機率就愈大。因此，當位元錯誤率＝10^{-2}時，如果群集程度愈大，辨識效果反而會愈好。

二．當位元錯誤率＝10^{-3}時，每個次音節單位中平均只有一個發生錯誤的位元，也就是說，平均只有一個特徵向量會發生錯誤，而這樣的錯誤對音素辨識的影響也不會太大（從實驗數據來

看也驗證這個說法)。此時如果發生群集錯誤，那麼應該是錯誤的位元群集在同一個特徵向量而其他特徵向量沒有發生誤，也因此群集錯誤所在的次音節單位會較容易發生錯誤。從上面的討論我們得到結論是：當位元錯誤率＝10^{-3}時，如果群集程度愈大，辨識效果會變差。

由上面的討論，本節所得到一個重要的結論是：**在不同的位元錯誤率下，群集程度對辨識結果的影響也不同。在高位元錯誤率下（例如10^{-2})，此時若群集程度愈大，對辨識效果愈好；但中等的位元辨識率下（例如10^{-3})，此時若群集程集愈大，對辨識效果愈差。**

5．4　無線通道錯誤的補償（Error Concealment）

在5．2及5．3中，我們討論到無線通道中隨機錯誤以及群集錯誤對語音辨識效能的影響：如果考慮位元錯誤率和音節辨識正確率的關係時，從實驗結果可以看到：在高位元錯誤率（10^{-2}）時，兩種錯誤都使得辨識效果下降許多；當位元錯誤率下降到10^{-3}時

，辨識正確率仍有微幅的下降，位元錯誤率要下降到大約10^{-4}以下，辨識正確率的下降才可忽略。對現有的無線通訊技術而言，要使位元錯誤率下降到10^{-3}以下並非難事；但是我們考慮通訊情況難免也會出現極差的情況，而就算是位元錯誤率維持在10^{-3}左右，也會有微幅的辨識正確率下降。因此，我們必須建立一些較有效的錯誤補償的機制，透過這個機制對位元串流中發生錯誤的部分做修正。有一件事是我們必須要注意的：要做錯誤補償的先決條件是要能偵測到發生錯誤的位元所在。在我們傳輸出去的位元串流中，除了有向量代碼外（代表每一個音框所抽出的特徵向量），還有描述資料類型的檔頭（Header）、對資料進行保護的編碼等。從這些保護用的編碼中，我們可以知道資料是否受到無線通道中雜訊的破壞。在本報告中，我們並不去研究如何安排這些保護用的編碼以達到偵測錯誤的目的；我們所要研究的是，提出有效的錯誤補償的機制，可以為我們帶來多少辨識效能的改進。因此在以下的討論中，我們均假設：用來保護資料的所做的編碼可以完美的偵淵出發生錯誤位元的所在。然後我們就可以對發生錯誤的位置，對該錯誤進行以下即將介紹的三種錯誤補償方式。

５・４・１ 通道錯誤的三種補償方式

在這一節當中，我們將介紹三種錯誤補償的機制。其中前兩種－－基於音框的消去法（Frame-Based Deletion）和基於音框的外插法（Frame-Based Extrapolation），是對音框所取出的整個特徵向量進行補償（因為補償的對象是從音框所取出的特徵向量，所以名曰“基於音框”）【２】，另一種方式－－基於子特徵向量的外插法(Sub-Vector Based Extrapolation)，是對發生錯誤的子特徵向量進行補償（因為補償的對象是子特徵向量，故名曰“基於子特徵向量”）。在本章一開頭我們就說過，假設我們可以正確的偵測出位元串流中發生錯誤的地方，那麼我們就可以知道發生錯誤的位元是位在那一個音框所取出的特徵向量中，或是進一步知道這個位元是在那一個子特徵向量中，此時，系統便可以宣告某個音框、或是某個子特徵向量發生錯誤；接著，我們就可以利用三種錯誤補償的機制，對發生錯誤的特徵向量、或是發生錯誤的子特徵向量進行補償，最後再以補償後產生的新特徵向量來進行語音辨識的工作。以上所述的實驗過程就如圖５・９所述。

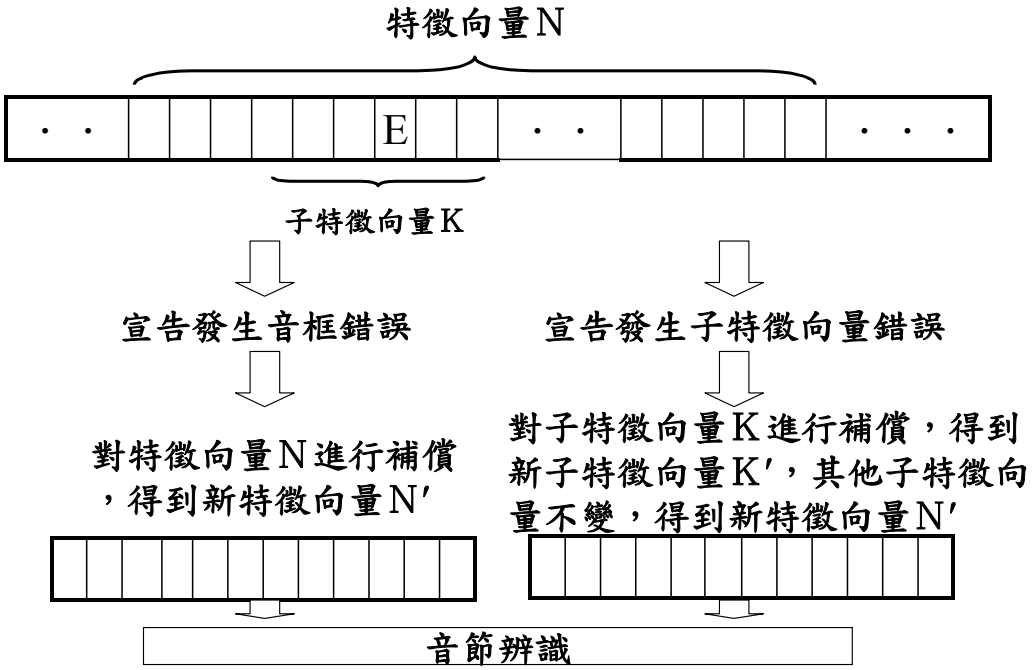


圖５．９　通道錯誤之補償機制

●基於音框的消去法（Frame-Based Deletion）

簡單來說，消去法就是：對於受到通道錯誤影響的音框，忽略不看，只看沒有通道錯誤的音框。其理論基礎是根據模糊特徵理論（Missing Feature Theory）【１２】【１３】。以下是相關的推導：

一．在第三章的地方有提到，對於隱藏式馬可夫模型而言，必須要建立三組參數：起始機率、轉移機率、狀態輸出函數。若我們以λ表示隱藏式馬可夫模型，以Ａ表示轉移機率，以Ｂ表示狀態輸出函數，以π表示起始機率，那麼給定一個隱藏式馬可夫

模型，觀察到一連串的特徵向量O＝｛o_1，o_2，．．．，o_N｝的機率是，

$$P(O|\lambda)=\sum_{q1,...qN}\pi_{q1}b_{q1}(o_1)a_{q1q2}b_{q2}(o_2)\cdots a_{q(N-1)qN}b_{qN}(o_N)$$

（５－１３）

其中

N是O中所有的特徵向量的數目，q_1，．．．，q_N是狀態列（State Sequence），π_{q1}是起始機率。而$b_i(o_n)$是在第i個中狀態中觀察到第n個特徵向量的狀態輸出函數值，

二．假設O中特徵向量o_1發生錯誤。我們想利用O中沒有發生錯誤的部分來進行辨識。於是我們首先將O分割成兩群O^c、O^m，其中O^c代表O中沒有錯誤的部分，O^m代表O中有發生錯誤的部分，o_1屬於O^m，則

$$O=(O^c,O^m)\ ,\qquad（５－１４）$$

因為是根據O中沒有錯誤的地方來做辨識，也就是說，我們要求的機率是：給定一個隱藏式馬可夫模型，觀察到一連串特徵向量O^c的機率：$P(O^c|\lambda)$。

三．根據模糊特徵理論，上述機率

$$P(O^c|\lambda)=\int P(O^c,O^m|\lambda)dO^m\ ,\qquad（５－１５）$$

又

$$\int b_i(o_l)^d o_l = 1 \text{ ,} \qquad (5-16)$$

將（５－１３）、（５－１６）代入（５－１５），所得為

$$P(O^c|\lambda) = \sum_{q1,\ldots qN} \pi_{q1} b_{q1}(o_1) a_{q1q2} b_{q2}(o_2) \cdots a_{ql-1ql} a_{qlql+1} b_{ql+1}(o_{ql+1}) \cdots a_{qN-1qN} b_{qN}(o_N) \qquad (5-17)$$

由於在維特比搜尋（Viterbi Search）中，轉移機率的重要性遠比狀態輸出函數來得小，因此我們可以令（５－１７）中的 $a_{q(l-1)ql}=1$，那麼（５－１７）所得到的結果，就只是將發生錯誤的 o_l 刪去，用其他正確的向量去做辨識罷了。

●基於音框的外插法（Frame-Based Extrapolation）

觀察從每個音框中所取出的特徵向量，可以發現有一個特性：相鄰的幾個特徵向量通常變化不大；這代表相鄰的幾個向量具有相關性。因此，我們可以利用這個性質，對有錯誤的特徵向量進行補償。我們選擇的方式是外插法：外插法是利用之前所接收到的特徵向量，經過線性組合後得到新的特徵向量，然後用此新的特徵向量代替原先有錯誤的特徵向量來做辨識。至於我們所使用的線性組合方式

是：

一．先計算出前五個收到的特徵向量的平均向量;

二．決定一個遺忘常數（Forgetting Factor）δ，

其中$0 \leqq \delta \leqq 1$，

則新的特徵向量

＝ 前五個特徵向量的平均 × （$1-\delta$）

＋ 前一個特徵向量的平均 × δ

由於前一個特徵向量與現在接收到的特徵向量相關性較大，因此我們將δ設為０．９。

●基於子特徵向量的外插法（Sub－Vector Based Extrapolation）

當無線通道有雜訊干擾時，位元串流中發生錯誤的地方可能只是特徵向量中的某一部分，其他的部分仍然是沒有出錯的，如果我們對整個特徵向量做補償，那麼我們會修改到沒有發生錯誤的地方。因為沒有發生錯誤的地方是可信的部分，或許我們可以想辦法保留這個部分，只對有出錯的地方做補償。

在５．４一開頭我們就有提到，我們假設保護位元用的編碼可以正確的偵測出錯誤，因此我們也可以知道出錯的位元是用來描述那一個子特徵向量，而在做補償時，我們只須對出錯的子特徵向量做外插的補償，對於其他沒有出錯的子特徵向量不去做改變－－這就是本節中所謂的基於子特徵向量的外插法。最後要注意的是，基於子特徵向量的外插法和前一節所述基於音框的外插法使用相同的線性組合方式，只不過補償的單位由音框所對應的特徵向量，換成是子特徵向量。

５．４．２ 實驗流程

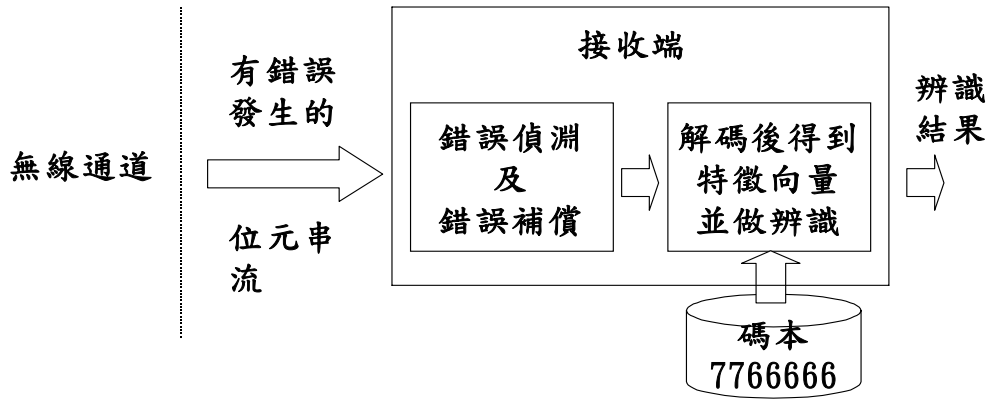


圖５．１０ 對通道中發生錯誤的位元串流進行錯誤補償。

在這一節中，我們簡述：無線通道發生錯誤時，對接收到的位元串流進行錯誤補償的實驗流程。如圖５．１０所示，對於所接收到的位元串流，接收端會先做錯誤偵測，將錯誤位元找出並尋找這個位元所在的特徵向量、或是所在的子特徵向量。接著再對有錯誤的特徵向量、或是子特徵向量進行錯誤補償。在錯誤補償後，再以新的特徵向量來做辨識。

５．４．３ 實驗結果：無線通道發生隨機錯誤時，補償通道錯誤對辨識結果的影響

在這一節及下一節中，我們將探討前述三種錯誤補償對辨識效果的幫助。在這一節中我們首先探討的是發生隨機錯誤時的錯誤補償情況。下表５．３所列出的是通道發生隨機錯誤時，補償通道錯誤對辨識結果的影響的實驗結果。第一欄列出的是各種不同的位元錯誤率；第二欄列出的是錯誤補償的方式，其中標示“無”代表沒有做錯誤補償，標示“消去法”表示基於音框的消去法，標示“外插法（音）”代表基於音框的外插法，標示“外插法（子）”代表基於子特徵向量的外插法；第三、四、五欄列出的是在高斯混合數＝２

位元錯誤率	錯誤補償方式	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8	備註
0	無	42.64	49.20	55.31	基礎實驗未經量化
0	無	40.64	48.03	52.88	有向量量化
10^{-5}	無	40.64	48.03	52.87	
10^{-5}	消去法	40.63	48.02	52.85	
10^{-5}	外插法（音）	40.64	48.03	52.87	
10^{-5}	外插法（子）	40.64	48.02	52.88	
10^{-4}	無	40.57	47.94	52.84	
10^{-4}	消去法	40.72/40.65	48.08/48.00	52.79/52.70	
10^{-4}	外插法（音）	40.67	48.04	52.86	
10^{-4}	外插法（子）	40.64	48.04	52.89	
10^{-3}	無	40.21	47.60	52.36	
10^{-3}	消去法	40.76/40.03	47.88/47.02	52.77/51.82	
10^{-3}	外插法（音）	40.62	48.02	52.85	
10^{-3}	外插法（子）	40.65	48.03	52.85	
10^{-2}	無	35.94	43.05	47.66	
10^{-2}	消去法	29.42/6.85	36.62/8.61	35.90/8.45	
10^{-2}	外插法（音）	40.12	47.67	52.29	
10^{-2}	外插法（子）	40.49	47.86	52.70	

表５・３　實驗結果：通道發生隨機錯誤時，補償通道錯誤對辨識結果的影響。如表中所示，列出的是在不同的位元錯誤率底下，補償錯誤後中文大字彙連續語音的音節辨識正確率。

、４、８時的辨識正確率；第六欄則是備註。在表中的第一、二列同時列出基礎實驗及只經過向量量化但無隨機錯誤的實驗結果。以下我們先對實驗結果製圖。

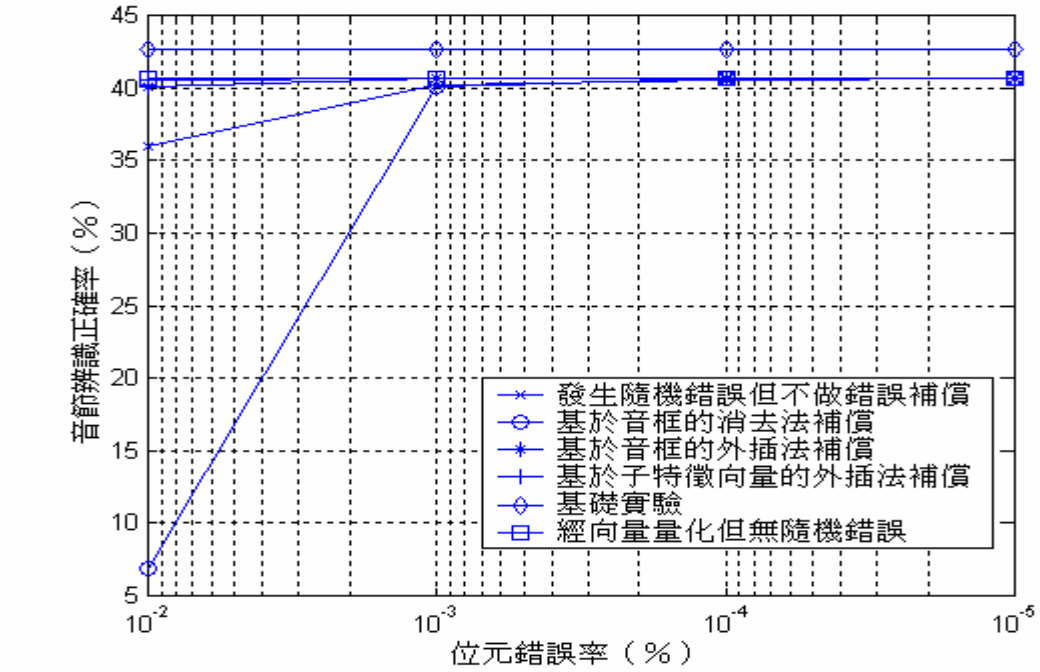


圖５・１１（Ａ）　高斯混合數＝２，發生隨機錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

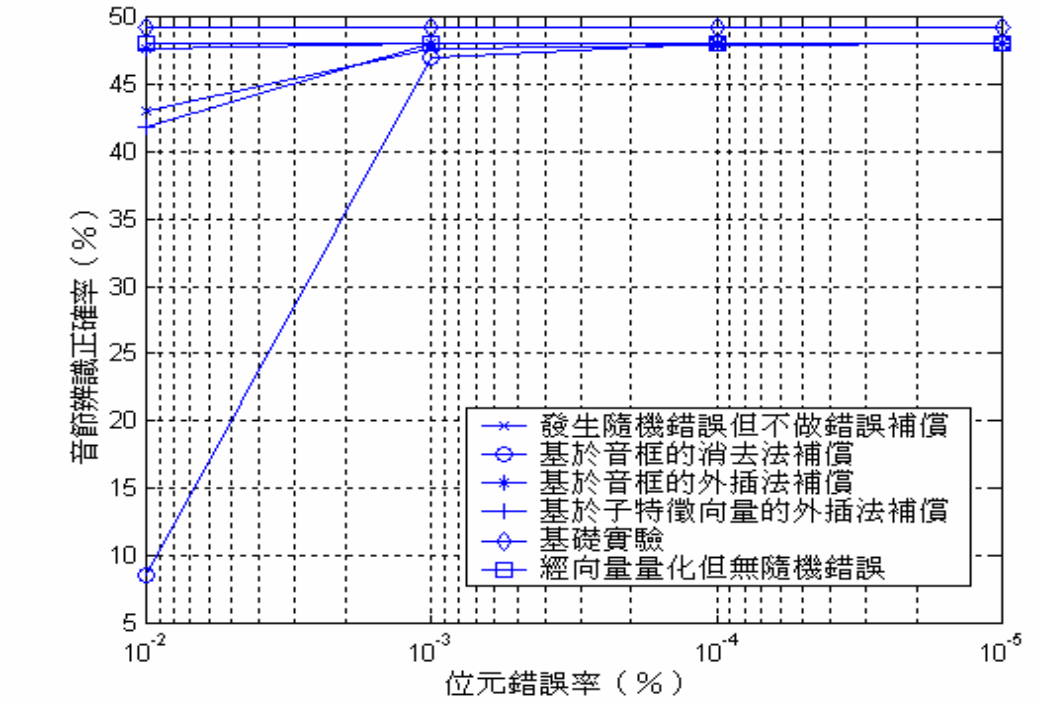


圖５・１１（Ｂ）　高斯混合數＝４，發生隨機錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

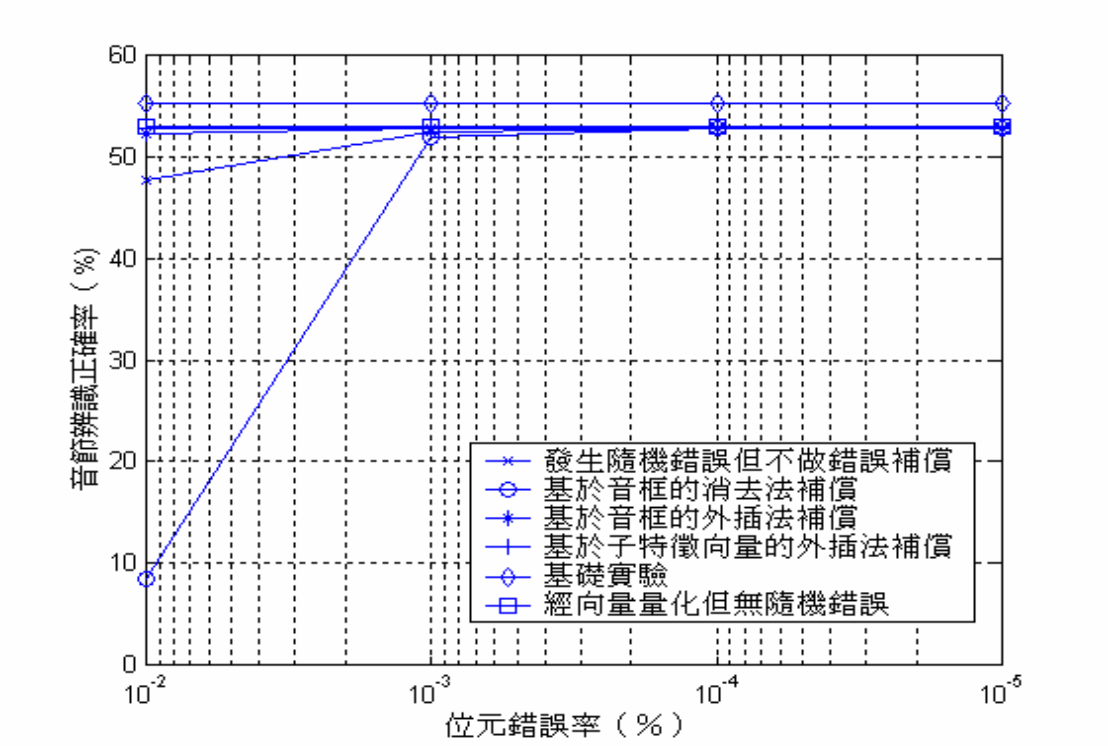


圖５．１１（Ｃ）　高斯混合數＝８，發生隨機錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

以上三張圖共有六條折線：第一條是發生隨機錯誤、但不做錯誤補償所得的辨識正確率；第二條是做基於音框的消去法補償所得的辨識正確率；第三條是做基於音框的外插法補償所得的辨識正確率；第四條是做基於子特徵向量的外插法補償所得的辨識正確率，以上三條折線必須要在第一條折線之上，才能顯出補償的效果；第五條是基礎實驗的辨識正確率；第六條是經過向量量化但沒有隨機錯誤所得的辨識正確率，因為向量量化後會帶來量化誤差，造成辨識效果的下降，所以這條折線被我們視為錯誤補償後辨識正確率的上限

。以下是我們對實驗結果的討論。

●基於音框的消去法對錯誤補償沒有幫助

基於音框的消去法之作法是：在４４個用來描述某一特徵向量的位元中，只要有一個位元發生了錯誤，就把整個特徵向量視為不可靠的辨識資訊而略去不用。那麼，在不同的位元錯誤率下，略去不用的特徵向量會有多少個呢？根據之前的討論，

一．在位元錯誤率＝10^{-2}時，用來描述１個次音節單位的１１００個位元中平均會有１１個發生錯誤；視群集程度的不同，最多可以有１１個特徵向量出錯誤，也就是說，最多會略去１１個特徵向量。

二．在位元錯誤率＞＝10^{-3}時，用來描述１個次音節單位的１１００個位元中平均不到一個會發生錯誤，也就是說，在每個次音節單位所對應的特徵向量可能頂多１個被略去不看。

前面提過，在進行錯誤補償後，我們希望辨識效果能趨近只做向量量化但沒有隨機錯誤影響時的情況（我們所認定的上限）。觀察使

用消去法時調整過後的的辨識正確率：

一．在位元錯誤率＝10^{-2}時只剩不到１０％，這印證了之前所提到的，每個次音節單位中最多將有１１個特徵向量被略去，造成辨識資訊不足的情形；

二．在位元錯誤率小於或等於10^{-3}時，使用消去法還是沒有趨近我們認定的上限。這是因為在位元錯誤率小的時候，可以看到位元錯誤已經很少，而這個錯誤的位元影響只及於所在的子特徵向量，其他的子特徵向量在辨識上依然是可信的——但是略去整個特徵向量，同樣也略去另外的仍然可信的子特徵向量，使得這些可信的子特徵向量無法在辨識上造成幫助。

因為**消去法不只略去較不可信的資訊，同時也略去可信的資訊，所以使用消去法所得到的辨識正確率不僅較我們認定的上限來得差，同時也比有隨機錯誤干擾、但沒有做錯誤補償的情況（雖有不可信的資訊，但也保留了可信的資訊）來得糟**，尤其是高位元錯誤率的時候。因為對錯誤的補償主要就是要改進在高位元錯誤率時的辨識效果，因此，在這裡我們可以下一個結論，那就是：消去法對錯誤補償並沒有幫助！

●**兩種外插法可以補償隨機錯誤所帶來的辨識效`果下降**

由圖５．１１可以看到，使用兩種外插法所得到的折線都在未做錯誤補償的折線上，顯示這兩種錯誤補償的方式達到了效果；不僅如此，這兩條折線均可以逼近我們所認定的上限－－經向量量化但沒有隨機錯誤的干擾。仔細觀察在不同位元錯誤率時，兩種外插法所得到的辨識正確率，則可以發現高位元錯誤率（10^{-2}）時，基於子特徵向量的外插法較基於音框的外插法還要好一些，而在位元錯誤率小於10^{-3}時，則兩種外插法的辨識效果幾乎已經沒有差距。回想在上一個小節中，基於音框的消去法其辨識之效能較未做錯誤補償之效能還要差，原因在於：基於音框的消去法將不可信的資訊略去，但同時也略去可信的資訊，因此造成資料量不足，於是辨識效果比不做錯誤補償來的差（不做錯誤補償：同時保留可信的、不可信的資訊）。**如果要辨識效果比不做錯誤補償來得好，那麼應該要能保留可信的資訊，同時也修正不可信的資訊**。就兩種外插法而言，基於子特徵向量的外插法較基於音框的外插法更能達到上述的要求：因為基於子特徵向量的外插法只有修正有錯誤的子特徵向量部分，對於其他沒有錯誤的子特徵向量並不做任何修改；但是基於音框的外插法不只修改了有錯誤的子特徵向量，對於其他沒有錯誤的

子特徵向量也做了修改。所以在位元錯誤率高的地方，基於子特徵向量的外插法表現得就會比基於音框的外插法要好一些；至於在位元錯誤率低的地方，位元串流中出錯的位元本來就少，因此在上述兩種錯誤補償的機制下辨識效果並沒有太大的差異。

基於音框的外插法，既然對整個特徵向量都做了修改，那麼為何在辨識效果上仍可以勝過消去法、勝過沒有做錯誤補償的、甚至逼近我們所認定的上限呢？因為基於音框的外插法沒有略去任何的資訊，不會造成無法辨識的情形，因此辨識效果要比消去法來得好；基於音框的消去法雖然同時修正了發生錯誤的子特徵向量、以及沒有發生錯誤的子特徵向量，但是因為音框與音框保有一定的相關性（外插法使用的原理），使得原先沒有發生錯誤的子特徵向量即使作了修正，也仍然近似於作修正之前的樣子，因此基於音框的外插法也滿足了上述「要能保留可信的資訊，同時也修正不可信的資訊」的要求，所以基於音框的外插法不但可以比沒有做錯誤補償時的辨識結果來得好，同時也會趨近我們所認定的上限。

最後，從本小節的實驗結果我們可以得到一個重要的結論：對於隨機錯誤所造成的影響，**如果可以建立有效的錯誤偵測機制，那麼在**

這一節當中所討論的兩種外插法幾乎可以完全補償因隨機錯誤所帶來的辨識效果下降；也就是說，整體的音節辨識結果將完全決定在前章所述的向量量化。

５．４．４ 實驗結果：無線通道發生群集錯誤時，補償通道錯誤對辨識結果的影響

在這一節中，我們將無線通道中的錯誤改為群集錯誤，再利用前述的錯誤補償方式對位元串流做補償。在上一節中，我們發現基於音框的消去法並沒有為辨識效能帶來幫助，因此在這一節中，我們只使用兩種外插法來進行錯誤補償。下表５．４列出的是實驗結果：第一欄列出的是各個不同的群集程度；第二欄列出的是使用的錯誤補償類型，“無”代表受到群集錯誤影響但未做錯誤補償，“外插法（音）”代表做基於音框的外插法補償，“外插法（子）”代表做基於子特徵向量的外插法補償；第三、四、五欄列出在高斯混合數＝２、４、８時的音節辨識正確率；第六欄則是列出備註。表中的第一、二列也列出基礎實驗和只經向量量化但沒有群集錯誤影響的實驗結果以供比較。我們先將下表的實驗結果製圖，如下圖５

・１３所示。

群集程度（％）	錯誤補償	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8	備註
	無	42.64	49.20	55.31	基礎實驗
	無	40.64	48.03	52.88	經向量量化
0	無	33.15	39.72	45.04	
0	外插法(音)	33.16	40.46	44.75	
0	外插法(子)	40.27	47.82	52.48	
25	無	34.25	41.13	46.12	
25	外插法(音)	35.78	42.88	47.58	
25	外插法(子)	40.46	47.73	52.71	
50	無	34.96	42.34	47.00	
50	外插法(音)	38.01	45.38	49.87	
50	外插法(子)	40.46	47.72	52.58	
75	無	35.17	41.90	46.91	
75	外插法(音)	39.14	46.60	51.03	
75	外插法(子)	40.61	47.89	52.74	
100	無	35.62	42.66	47.39	
100	外插法(音)	39.90	47.35	52.29	
100	外插法(子)	40.27	47.82	52.48	

表５・４（Ａ）　實驗結果：通道發生群集錯誤時，補償通道錯誤對辨識結果的影響。如表中所示，列出的是位元錯誤率＝10^{-2}時，在不同的群集程度底下，補償錯誤後中文大字彙連續語音的音節辨識正確率。

群集程度（%）	錯誤補償	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8	備註
	無	42.64	49.20	55.31	基礎實驗
	無	40.64	48.03	52.88	經向量量化
0	無	39.92	47.33	52.24	
0	外插法(音)	40.35	47.77	52.47	
0	外插法(子)	40.62	48.08	52.77	
25	無	40.08	47.55	52.35	
25	外插法(音)	40.42	47.91	52.54	
25	外插法(子)	40.64	48.12	52.84	
50	無	39.97	47.44	52.36	
50	外插法(音)	40.45	47.84	52.68	
50	外插法(子)	40.60	48.04	52.91	
75	無	39.95	47.17	52.06	
75	外插法(音)	40.56	47.94	52.69	
75	外插法(子)	40.64	48.09	52.91	
100	無	39.64	46.95	51.77	
100	外插法(音)	40.60	47.93	52.87	
100	外插法(子)	40.66	48.03	52.92	

表５・４（Ｂ）　實驗結果：通道發生群集錯誤時，補償通道錯誤對辨識結果的影響。如表中所示，列出的是位元錯誤率＝10^{-3}時，在不同的群集程度底下，補償錯誤後中文大字彙連續語音的音節辨識正確率。

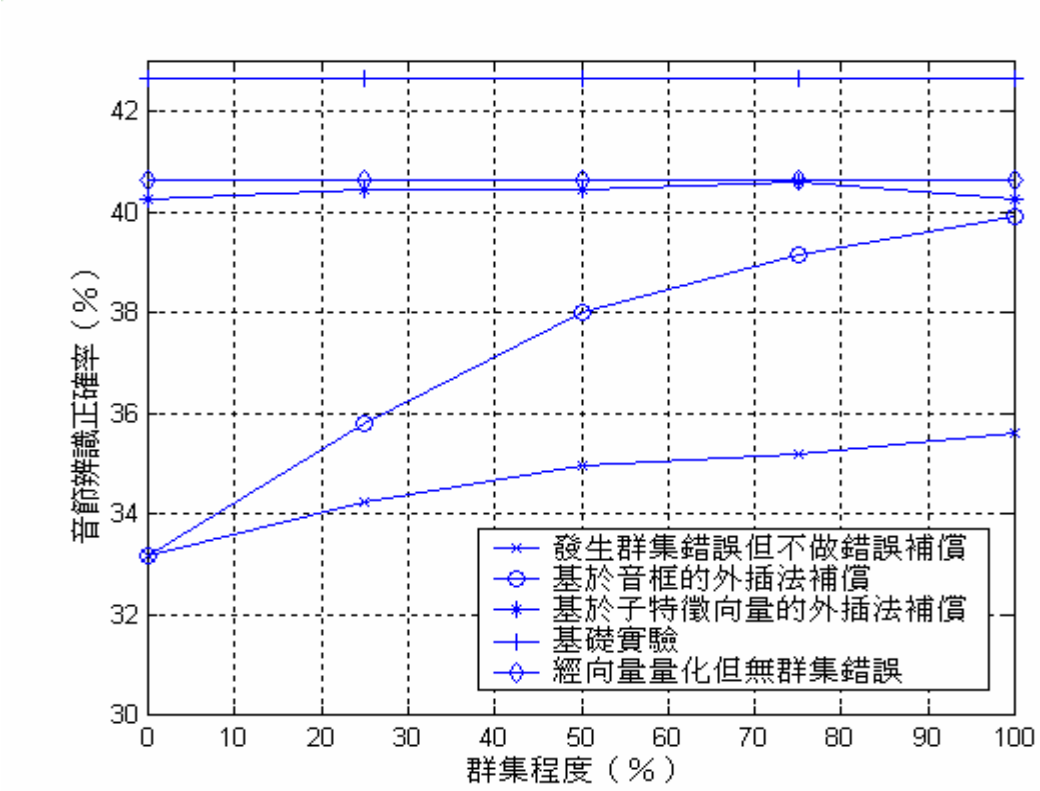


圖５・１３（Ａ）　位元錯誤率＝10^{-2}時，高斯混合數＝２，發生群集錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

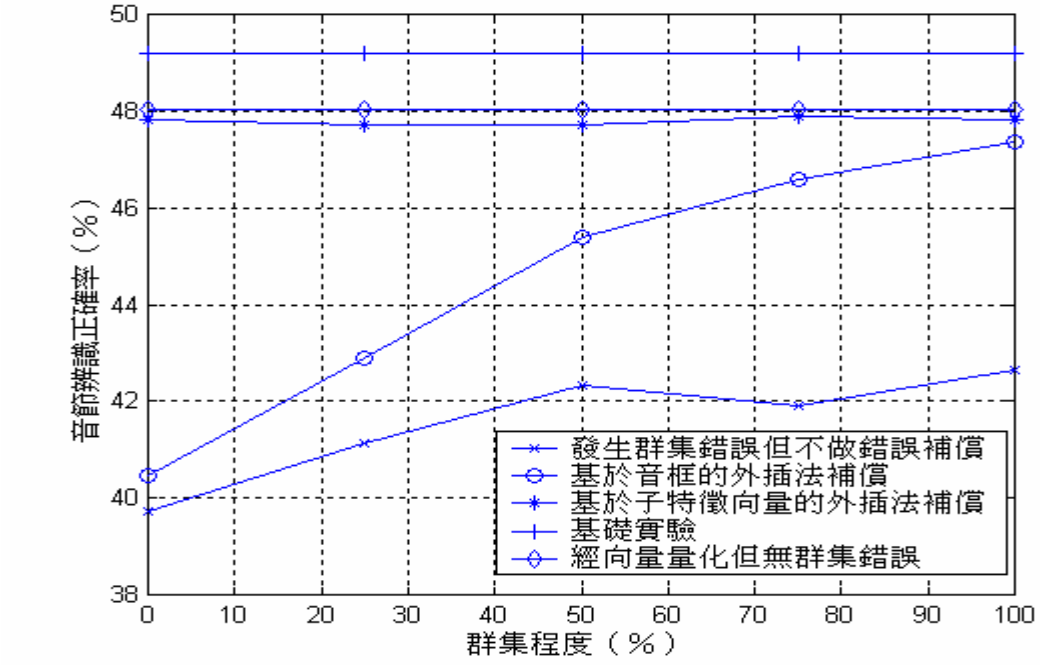


圖５・１３（Ｂ）　位元錯誤率＝10^{-2}時，高斯混合數＝４，發生群集錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

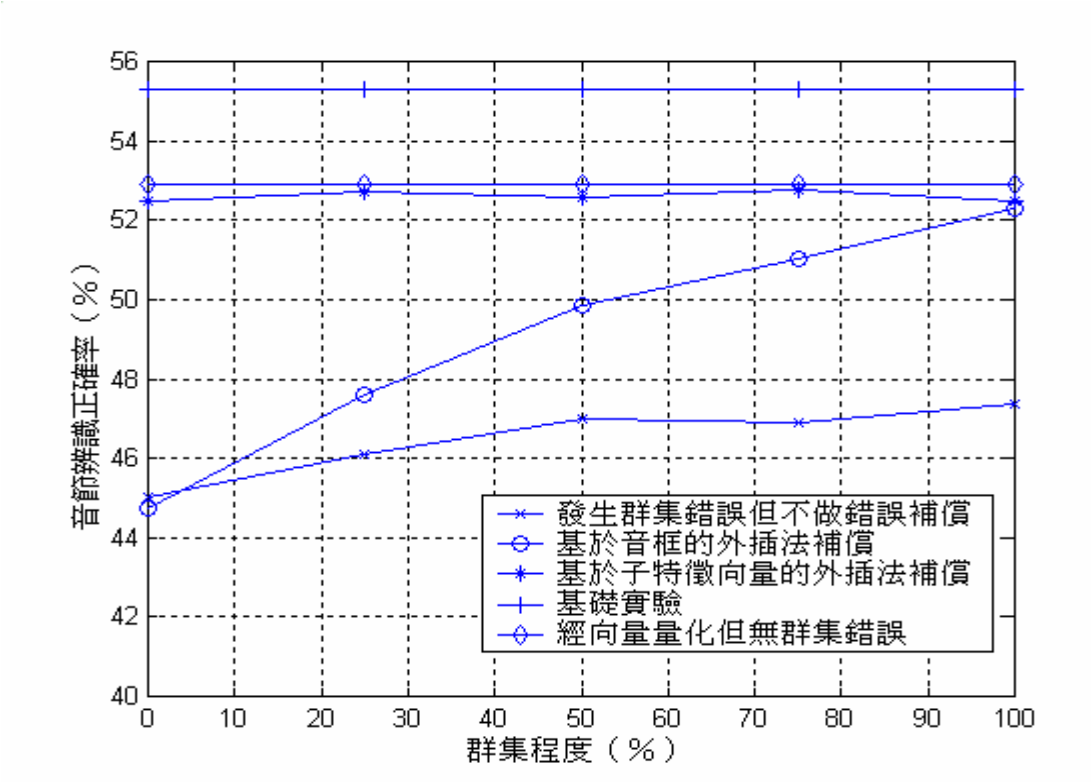


圖５・１３（Ｃ）　位元錯誤率＝10^{-2}時，高斯混合數＝８，發生群集錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

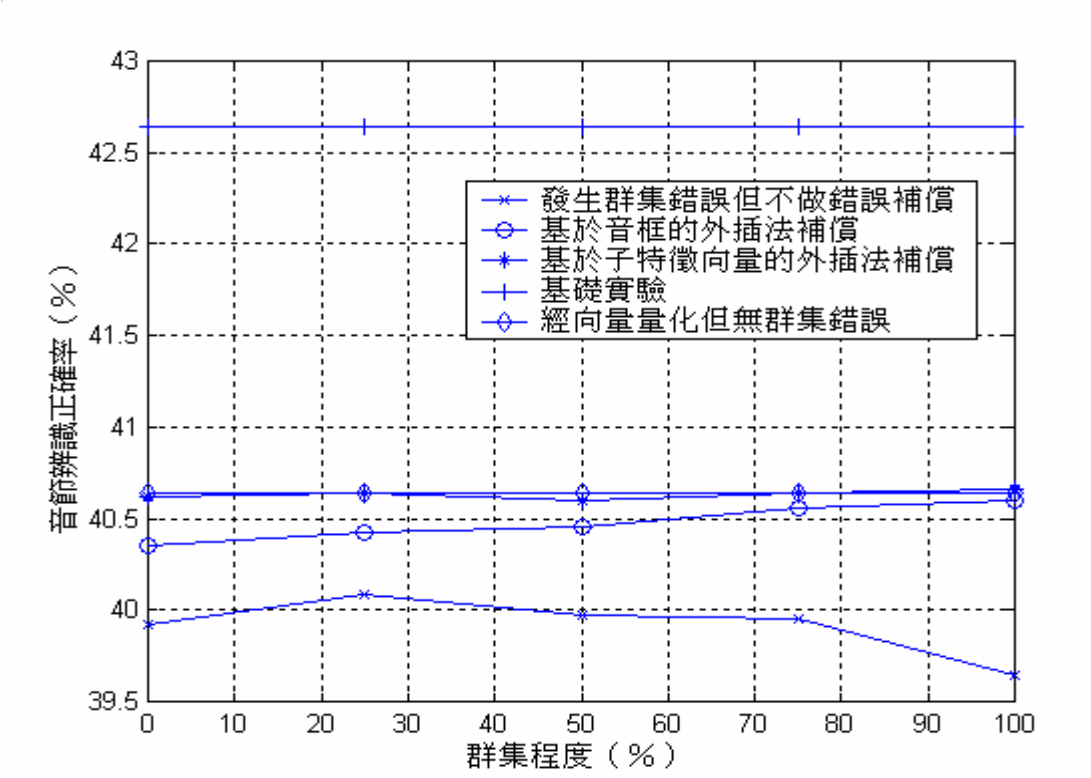


圖５・１３（Ｄ）　位元錯誤率＝10^{-3}時，高斯混合數＝２，發生群集錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

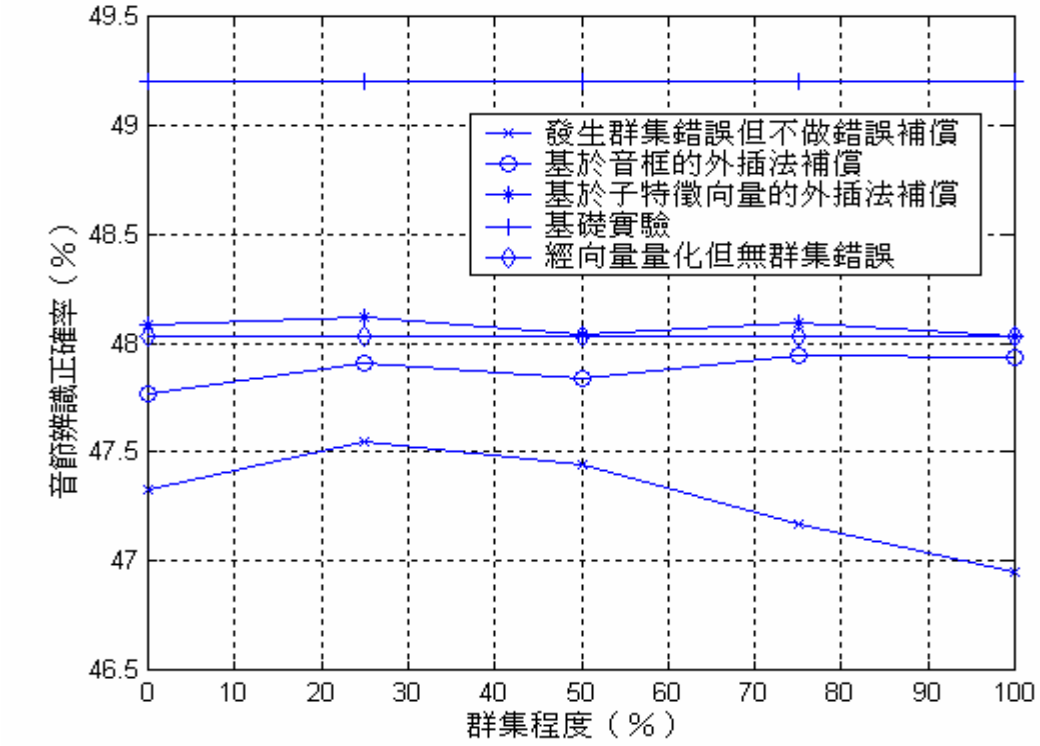


圖５・１３（Ｅ）　位元錯誤率＝10^{-3}時，高斯混合數＝4，發生群集錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

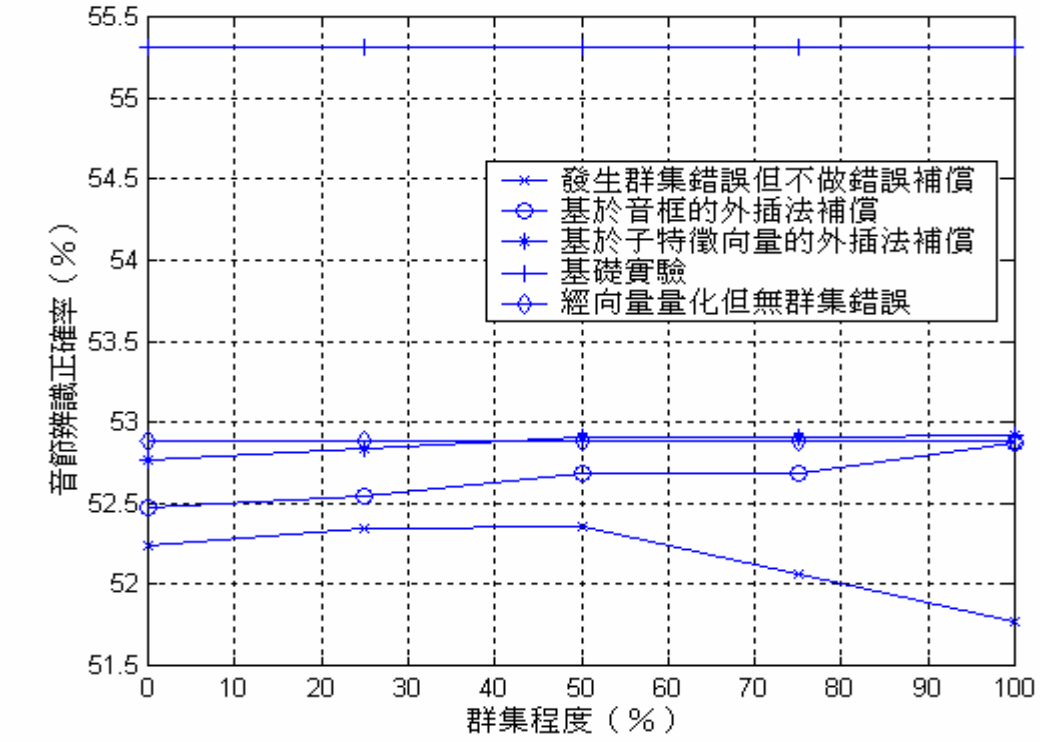


圖５・１３（Ｆ）　位元錯誤率＝10^{-3}時，高斯混合數＝８，發生群集錯誤時對位元串流做錯誤補償所得到的音節辨識正確率。

在以上六張圖中總共列出五條折線：第一條是發生群集錯誤但是沒有做錯誤補償；第二條是對發生群集錯誤的位元串流做基於音框的外插法補償；第三條是對發生群集錯誤的位元串流做基於子特徵向量的外插法補償；第四條是基礎實驗的結果；第五條是經過向量量化但沒有群集錯誤時的實驗結果。同樣的，第五條折線被我們視為錯誤補償後辨識正確率可以上升的最高極限。以下是我們對實驗結果的討論。

●**發生群集錯誤時，利用兩種外插法做錯誤補償仍有產生效果**

在前一節中有提到：錯誤補償的機制若發生效果，所得到的辨識正確率折線將落在發生群集錯誤、但不做錯誤補償所得的折線以上。觀察圖５・１３六張圖中，除了在圖５・１３（Ｃ）中群集程度＝０％時使用基於音框的外插法所得的折線落在未做錯誤補償所得的折線以下，其餘的地方兩種外插法所得到的辨識正確率折線均在未做錯誤補償所得的折線以上。**這代表兩種外插法在群集錯誤的情況下也可以發揮效果。**

●**在不同的群集程度下，基於音框的外插法所得的補償效果不一**

雖然兩種外插法可以對發生群集錯誤的位元串流進行補償，但是基於音框的外插法在不同的群集程度下卻有不太一樣的辨識效果：觀察圖５．１３的六張圖可以發現，**在群集程度低時，基於音框的外插法所能給的錯誤補償不多；隨著群集程度逐漸上升，補償的效果才逐漸明顯，到高群集程度時辨識正確率才可以比較逼近我們認定的上限**——經過向量量化但沒有群集錯誤所得的辨識結果。這樣的現象在位元錯誤率＝10^{-2}時更為明顯。

讓我們再回想一下之前討論過的兩點：

一．基於音框的外插法的補償方式：在描述１個特徵向量的４４個位元中，如果其中有任何一個位元發生錯誤，那麼我們就將此特徵向量視為是錯誤的，然後利用之前所收到的特徵向量做線性組合，拿來取代錯誤的特徵向量，這個方法是基於音框和音框間的相關性；

二．群集程度愈小時，代表錯誤愈分散，所能影響到的特徵向量也愈多，被判斷為錯誤的特徵向量也愈多；反之，若群集程度愈大，代表錯誤愈集中，所能影響到的特徵向量也愈少，被判斷

為錯誤的特徵向量也愈少。

根據上面兩點所述，當群集程度愈小時，被判斷為錯誤的特徵向量會愈多，因此要做比較多次的錯誤補償。因為基於音框的外插法固然會修正錯誤的子特徵向量，但也同時改變沒有發生錯誤的子特徵向量，換句話說，外插法會帶來失真。當基於音框的外插法使用得愈多，帶來的失真也愈多，對於辨識效果所能提供的改進也會愈少。反之，當群集程度愈大時，被判斷為錯誤的特徵向量會愈少，因此錯誤補償所帶來的失真也會愈小；加上此時被判定為錯誤的特徵向量，其中有錯誤的子特徵向量數目會很多，對這個特徵向量做外插法的補償可以修改的錯誤也很多，因此辨識效果受到的幫助也會愈大。

最後要補充的是：從前一節的結果我們可以知道，用基於音框的外插法來補償隨機錯誤的影響得到的效果不錯，但是從本節的結果告訴我們，用這個方式來補償群集錯誤不一定會得到好的效果；因此我們可以說，**對群集錯誤所造成的錯誤作補償，比對隨機錯誤所造成的錯誤作補償來的難。**

●**基於子特徵向量的外插法有很好的效果**

從圖５．１３的六張圖中可以看到，基於子特徵向量的外插法所得到的折線並沒有一致的趨勢變化；我們可以看出，不管在何種的群集程度下，基於子特徵向量的外插法所得到的辨識正確率，一直都是逼近、甚至是超越我們所認定的上限－－經向量量化但沒有群集錯誤的影響。這也代表了發生群集錯誤時，基於子特徵向量的外插法不管在群集程度為何，均保持不錯的補償效果。因此，我們得到兩個結論：

一．因為對音框做外插法補償效果並不一定好，反觀在各種情況下對子特徵向量做補償均發生一定效果，故**我們有必要對每個子特徵向量作編碼保護**，如此一來才可以偵測出那個子特徵向量出了錯，並對其進行補償。

二．結合上節對於隨機錯誤的補償、以及本節的結果：**若能正確的偵測出位元錯誤發生在那個子特徵向量，那麼透過基於子特徵向量的補償，我們可以忽略通道中隨機錯誤或是群集錯誤的影響。**

5．5 結論

在本章中，我們主要討論無線通道中雜訊所造成的隨機錯誤、以及群集錯誤的影響：

一．對於隨機錯誤而言，在高位元錯誤率（10^{-2}）時，辨識正確率會下降許多（5%）；在中位元錯誤率（10^{-3}）以後，辨識正確率下降的幅度會開始小於1%。因此，只要能夠將隨機錯誤的位元錯誤率控制在10^{-3}以下，隨機錯誤的影響就不算太大。

二．對於群集錯誤所造成的影響：

錯誤群集對於辨識來說，未必不是好事；

在高位元錯誤率（10^{-2}）時，辨識正確率下降在5%～9%，此時若群集程度愈大，辨識效果愈佳；

在低位元錯誤率（10^{-3}）時，辨識正確率下降在1%左右，若群集程度愈大，辨識效果愈差。

三．對無線通道造成的錯誤做補償：

若是無線通道發生隨機錯誤，則我們所提出的基於音框的外插法、基於子特徵向量的外插法幾乎能完全補償隨機錯誤所帶來的辨識效能下降；

若是無線通道發生群集錯誤，則我們所提出的基於子特徵向量的外插法幾乎能完全補償隨機錯誤所帶來的辨識效能下降。因此，透過有效的偵測錯誤，我們所建立的錯誤補償機制可以幫助我們消去無線通道中雜訊所帶來的影響。

第六章　純主式架構

我們先回顧一下，在第二章中曾詳細介紹的純主式架構。用戶端所使用的手持設備，其傳送語音信號的方式，是使用語音編碼器（Speech Encoder）將語音信號編碼，然後經過無線通道傳送到受話的另一端，再由另一端的語音解碼器（Speech Decoder）解出合成語音供受話端接聽。純主式架構基於（一）使用現有架構及（二）節省頻寬的設計概念，因此

一．純主式架構之一是使用語音解碼器解出的合成語音作辨識；

二．純主式架構之二是抽出編碼語音中和語音辨識相關的資訊來做辨識。

圖６．１是純主式架構之一以及純主式架構之二的示意圖。

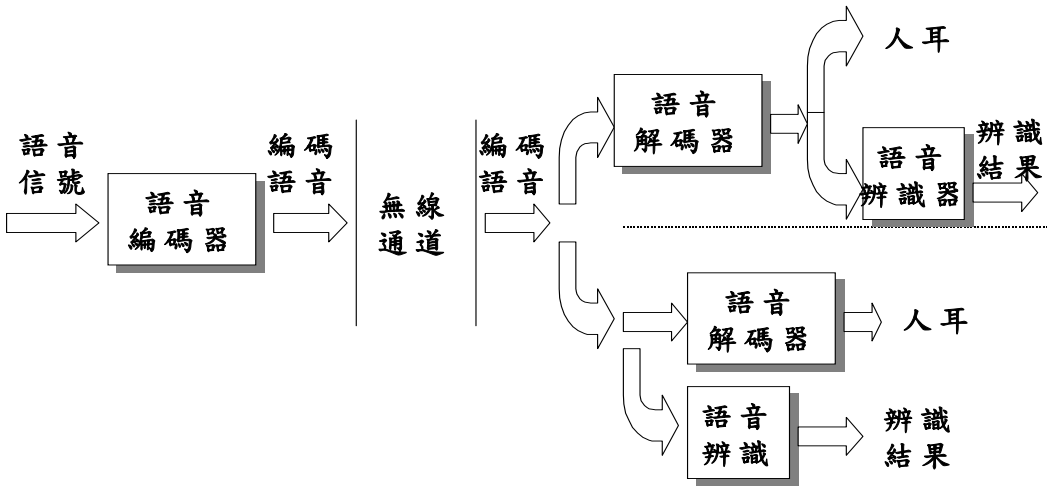


圖６．１　純主式架構。虛線以上是純主式架構之一，虛線以下是純主式架構之二

在過去發表的文獻【４】中均討論到，純主式架構之一所遭遇到最大的問題，就是在語音編碼之後所得到的合成語音，其中所含與辨識相關的資訊，和編碼前的原始語音所含的已經大不相同；因此所得到的辨識結果會比前兩章中所提到基礎實驗的結果來的差。由於純主式架構之一有上述的瓶頸，純主式架構之二的觀念就被提出：因為編碼語音中含有線性預估分析（Linear Predictive Analysis）所得的結果：線性預估分析所得到的線性預估係數（Linear Predictive Coding Coefficients），代表著聲道（Vocal Tract）中的資訊，正是語音辨識相關的資訊，故在接收端，我們可以利用這部分的資料來做辨識；另外，線性預估分析還可以得到殘餘信號（Residual Signal）的資訊，代表著聲道等相關的激發信號；結合線性預估係數以及殘餘信號的資訊，在一些文獻的記載中【２】【１４】【１５】，可以得到和基礎實驗相近的辨識效果。在本章中，我們想要了解的是，在中文大定彙連續語音的音節辨識中：

一．純主式架構之一的合成語音所得到的辨識效果，是否如同文獻中所述，會較原始語音所得的辨識效果差；

二．純主式架構之二中，如何結合線性預估係數以及殘餘信號的資訊以得到最佳的辨識效果。

以下的章節便針對這兩個主題做探討。另外，純主式結構必須使用

聲碼器（Vocoder，語音編碼器＋語音解碼器）傳輸語音；在我們的實驗中，所選定的聲碼器是調適型多速率聲碼器（Adaptive Multi-Rate Vocoder，AMR)；在第三代無線網路（３rd Generation Wireless Networks）中，調適型多速率聲碼器被制定標準的單位３G Partnership Project（３GPP）設為預設的聲碼器，GSM也將之列為其新一代的語音編碼器（使用的傳輸速率１２．２Ｋｂｐｓ)。有關調適型多速率編碼器在６．１有簡單的介紹，詳細的資訊參照【１６】。

６．１　調適型多速率聲碼器的簡介

調適型多速率聲碼器是一種基於線性預估編碼的聲碼器。在第一章中我們曾提到，在線性預估編碼分析中，語音信號被分解成一個線性預估濾波器、以及一個激發信號，稱為殘餘信號（Residual Signal）；所謂的「基於線性預估編碼分析的聲碼器」，即是將上述代表濾波器的線性預估係數、以及激發信號做編碼的聲碼器。

圖６．２是調適型多速率編碼器的功能方塊圖。編碼器以８ｋＨｚ

的取樣頻率對語音信號取樣，並量化成為１３位元的均勻脈衝編碼調變（Uniform PCM）形式，再來編碼器會對語音信號做一些前處理（Pre－Processing）的動作。接著是進行線性預估編碼分析。編碼器以１６０個取樣點為一個音框（２０ｍＳ），對每個音框進行線性預估編碼分析，得到１０階的線性預估係數，然後再將其轉變為線頻譜對（Line Spectrum Pair，LSP）後量化輸出；同時，編碼器以四分之一個音框為一新單位、稱子音框（Sub－Frame），對每個子音框進行線性預估分析得到殘餘信號的部分，然後再對殘餘信號做分析，得到音高（Pitch）、代數碼（Algebraic Code）

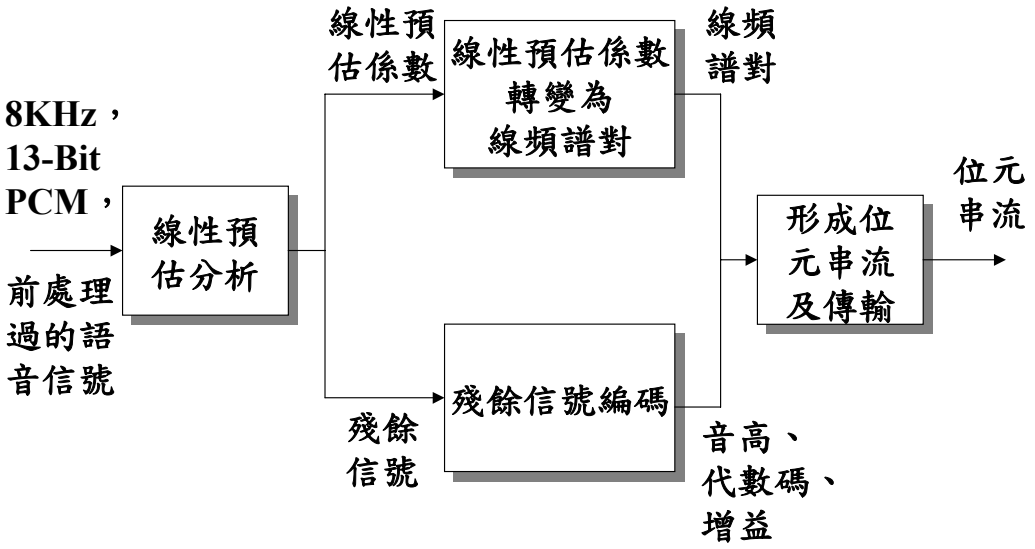


圖６．２　調適型多速率語音編碼器的功能方塊圖

、增益（Gain）等。調適型多速率聲碼器具有多種的位元速率，我們

所採用的是４．７５Ｋｂｐｓ，每個音框對應的輸出位元共９５個，詳細的位元分配如圖６．３所示。

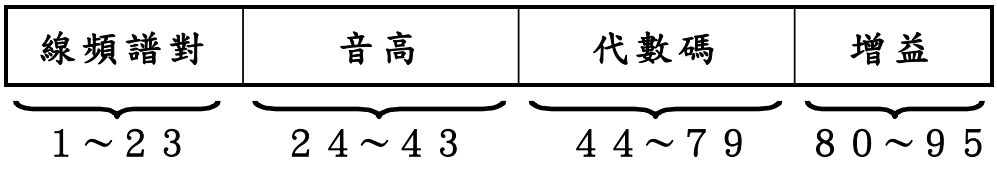


圖６．３　當位元速率＝４．７５ｋｂｐｓ時，調適性多速率語音編碼器所輸出的編碼語音中，每個音框的位元配置格式。

６．２　純主式架構之一－－基於合成語音的語音辨識

６．２．１　實驗流程

在這一節當中，我們進行純主式架構之一的實驗：利用合成語音來做辨識，並將辨識結果與基礎實驗所得之結果比較，以了解語音編碼器對語音辨識的影響。我們所使用的編碼器是調適型多速率編碼器，這是一個具有多種傳輸速率的編碼器，傳輸速率共有８種，由

４．７５～１２．２Ｋｂｐｓ不等。為了和主從式架構所得到的結果進行比較，我們選定傳輸速率是４．７５Ｋｂｐｓ，和前一章中碼本組合“７７６６６６６”的傳輸速率４．４Ｋｂｐｓ相去不多。

實驗的流程如圖６．１中上半部分所示。語音信號經過語音編碼器後得到編碼語音，然後傳輸端將編碼語音送進無線通道中；經過無線通道後編碼語音被接收端所接收，並且解碼得到合成語音，我們就使用得到的合成語音來做辨識。有幾個細節要另外說明：

一．在這一節中我們想要探討的是聲碼器對辨識效果的影響，所以我們假設在無線通道中並沒有發生錯誤；

二．我們使用與前兩章相同的方法，從合成語音中抽取出辨識相關的資訊，亦即：從合成語音中抽取出梅爾倒頻譜係數；

三．合成語音的語音信號已經受到破壞，和原始語音有不匹配的情況，如果用原始語音來訓練統計模型，在接收端取合成語音做辨識，我們預期辨識效果會比用合成語音來訓練統號模型、同時用合成語音來做辨識（匹配）來得差；這一節中我們分別訓練不同模型來觀察上述匹配和不匹配的差別。

６・２・２　實驗結果：純主式架構之一－－匹配和不匹配的情況下，合成語音的辨識結果

模型匹配情況	高斯混合數＝２	高斯混合數＝４	高斯混合數＝８
基礎實驗	42.64	49.20	55.31
主從式架構無通道錯誤	40.64	48.03	52.88
模型匹配	36.37	42.70	47.88
模型不匹配	32.25	38.14	40.28

表６・１　實驗結果：純主式架構之一－－匹配和不匹配的情況下，合成語音的音節辨識正確率。

表６・１列出利用純主式架構之一所得的辨識結果：在第一欄中，列出的是模型匹配情況；在第二、三、四欄中列出的是在高斯混合數＝２、４、８時的音節辨識正確率；在第一列列出基礎實驗的結果，第二列列出主從式架構、沒有無線通道錯誤時的辨識正確率，這兩列是比較用的；第三列列出的是模型匹配所得到的音節辨識正確率；第四列列出的是模型不匹配所得到的音節辨識正確率。從表６・１我們得知：當模型不匹配的時候，若高斯混合數＝２、４時，辨識正確率較基礎實驗下降約１０％，若高斯混合數＝８時，辨識正確率更下降約１５％！而模型匹配時，辨識正確率也下降大約

6～7%。

就上面所得到的實驗結果，我們有幾點的討論：

一．在第四章討論主從式架構時，因為用來訓練模型的特徵向量可以選擇做量化、也可以選擇不做，但是用來做測試的特徵向量必定經過量化，於是有匹配與否的情況；而這裡所陳述的匹配與否是針對是否經過編碼的語料而言。雖然兩種情況不太一樣，但是所呈現的的實驗結果卻是相似，**不匹配的情況的確會影響使得辨識正確率下降。**

二．即便是模型匹配的情況，所得到的辨識正確率也比基礎實驗的結果要來得差。**這代表了使用合成語音做辨識，辨識效果的下降不只是來自於模型的不匹配——經過聲碼器的編碼，語音信號中與辨識相關的資訊已經被改變，致使辨識效果不如未編碼前所得的結果。**

根據實驗結果，我們證實了利用純主式架構之一來實現分散式語音辨識系統是不可行的。因此在下面的章節中，我們將討論如何從聲碼器中的語音編碼器的輸出來取資訊作辨識。

６．３　純主式架構之二－－基於編碼語音的語音辨識

純主式架構之二的概念是：在語音編碼器輸出的編碼語音中，取出辨識相關的資訊，再用這些資訊來做辨識。在這一節中，我們將討論取出這些資訊的方法，並且以實驗結果來探討純主式架構之二的可能性。

６．３．１　實驗流程

在６．１中我們曾經提到：調適型多速率語音編碼器輸出的編碼語音中，有兩部分的資訊：第一部分是關於線性預估編碼分析的部分，以線頻譜對的形式表示；第二部分是關於殘餘信號的部分，以音高、代數碼、增益的形式表示。我們取出資訊的方式是：

一．將位元串流中用來描述線頻譜對的位元轉變成１０個線性預估係數，再將這１０個線性預估係數，轉換到倒頻譜頻域，得到１３個線性預估倒頻譜係數（Linear Prediction Cepstral Coefficients，LPCC）。因為一個音框所佔的時間為２０ｍＳ

，故在此處我們是每２０ｍＳ做一次上述的轉換；

二．編碼語音中用了三種形式來表示殘餘信號。因為無法確定那個部分的資訊對辨識效果最有幫助，所以我們希望可以完整地取出殘餘信號【１５】。因此在語音解碼器處，我們將原語音解碼器中的線性預估分析的線性預估濾波器設為０階，也就是說，除了０階的線性預估係數設為１以外，其他的線性預估係數均設為０。由此作法，就可以得到解回的殘餘信號部分。在得到殘餘信號後，我們對殘餘信號取１３個梅爾倒頻譜係數；為了取出和線性預估倒頻譜係數相同的資料量，此處取梅爾倒頻譜係數所使用的音框平移（Frame Shift）設為２０ｍＳ。

以上所述如圖６．４所示。

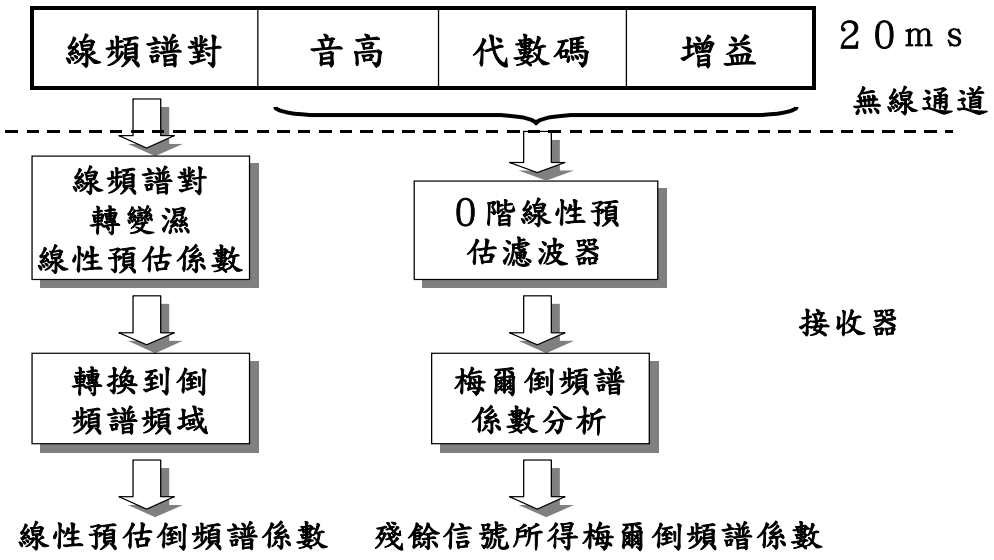


圖６．４　編碼語音中抽出辨識用特徵係數的方法

我們總共取出兩組特徵係數。我們希望結合上述兩項資訊，形成單一個辨識用的特徵向量。有幾個基本觀察：

一．辨識相關的資訊主要在線性預估濾波器中；

二．在倒頻譜頻域中，所求得的倒頻譜係數若序數愈小，在辨識上的重要性也愈大；

三．能量頻譜係數在辨識上是個重要的資訊，我們可以求出殘餘信號的能量頻譜係數做為這部分的資訊。

因此，辨識用的特徵向量其組合方式是：從線性預估倒頻譜係數中取出ｋ個係數，再由殘餘信號所得的梅爾倒頻譜係數中取出（１３－ｋ）個係數。在線性預估倒頻譜係數中，是以序數較小的係數較

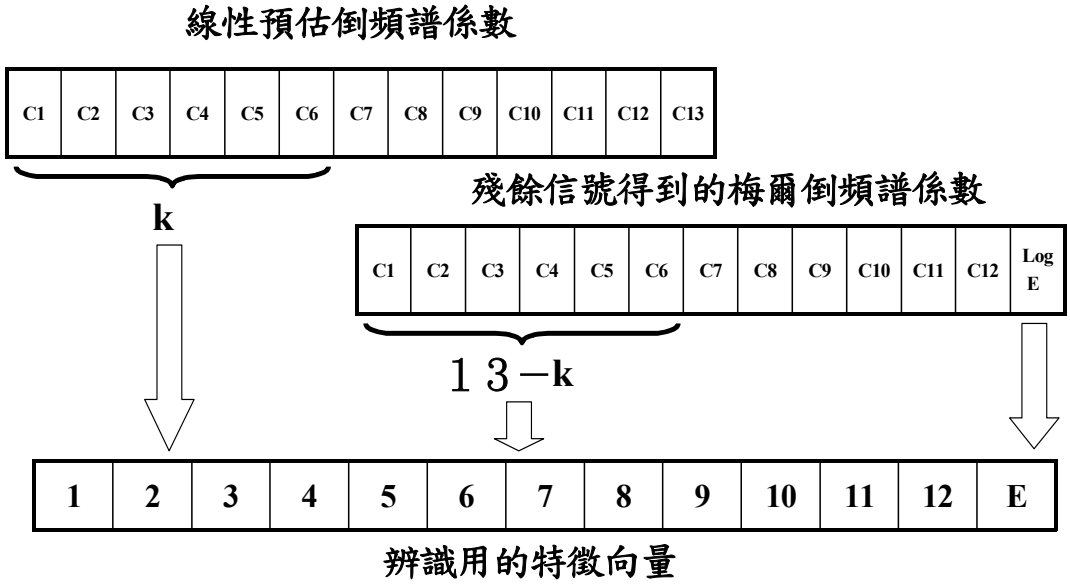


圖６．５　辨識用的特徵向量之組成方式

優先選取；而在梅爾倒頻譜係數中，較優先選取能量頻譜係數，再來才選序數較小的係數。以上如圖６．５所述。取出１３維的特徵向量後，再取這１３維的一次和二次回歸，得到最後３９維的特徵向量。

最後有一點要補充說明：為了符合匹配的條件，不管是用來訓練模型的訓練語料，或是測試用的測試語料，同樣都經過上述過程的處理得到所要的特徵向量；然後用所得到的特徵向量訓練模型、進行語音辨識。

６．３．２　實驗結果：純主式架構之二－－由編碼語音中抽取資訊做辨識所得的辨識效果

表６．２列出的是：線性預估倒頻譜係數和殘餘信號的梅爾倒頻譜係數在各種結合方式下，所對應的音節正確率。第一欄列出的是ｋ值，ｋ代表辨識用的特徵向量中，線性預估倒頻譜係數所佔的個數；第二、三、四欄列出的是在高斯混合數＝２、４、８時所得到的辨識正確率；第五欄列出的是備註事項。另外，第一列列出的是基

礎實驗的結果（原始語音所得到的梅爾倒頻譜係數，未做量化、也沒有發生通道錯誤）；第二列列出的是主從式架構的結果（原始語

k值	高斯混合數＝2	高斯混合數＝4	高斯混合數＝8	備註
	42.64	49.20	55.31	基礎實驗
	40.64	48.03	52.88	主從式架構
	36.37	42.70	47.88	純主式架構之一
0	-3.55	-2.26	-1.52	
1	-3.12	-2.10	-1.63	
2	-2.18	-1.94	-1.79	
3	31.73	34.68	37.65	
4	36.01	40.43	43.42	
5	38.79	43.88	47.23	
6	40.44	45.44	49.85	
7	41.36	46.66	51.02	
8	43.00	48.75	53.00	
9	43.70	49.42	53.43	
１０	43.15	49.25	53.30	
１１	43.41	48.87	52.78	
１２	42.98	48.37	52.73	
１３	34.57	40.27	44.63	

表６・２　實驗結果：純主式架構之二－－線性預估倒頻譜係數和殘餘信號的梅爾倒頻譜係數在各種結合方式下，所對應的音節辨識正確率。k值代表１３維特徵向量中線性預估倒頻譜係數的個數。

音所得到的梅爾倒頻譜係數，做向量量化、但沒有通道錯誤）；第三

列列出的是純主式架構之一、模型匹配的實驗結果（同樣假設沒有無線通道錯誤）；列出這三列是為了比較的用途。以下先對表６．２的結果製圖，如圖６．６。

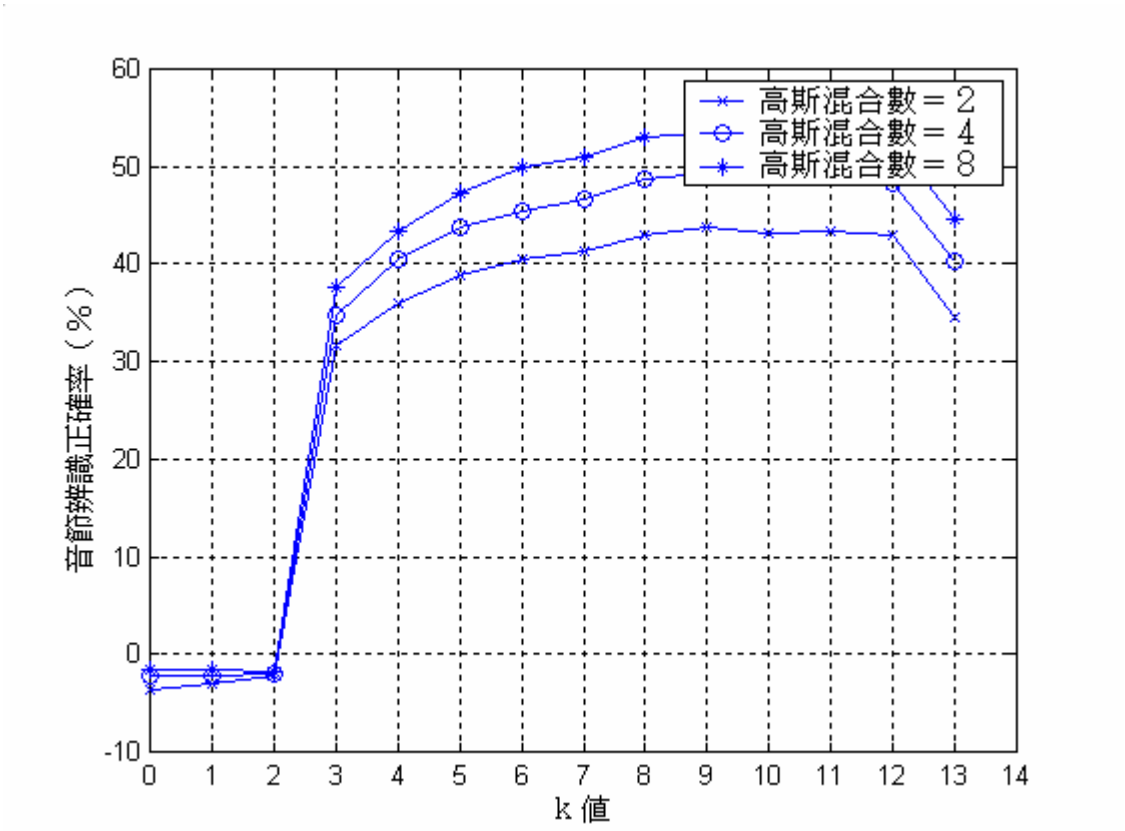


圖６．６　在各種不同的ｋ值時，所對應的音節辨識正確率。

●**殘餘信號可以帶來辨識上的幫助**

表６．２中的最後一列列出的是ｋ＝１３時的辨識正確率。當ｋ＝

１３時，代表１３維的特徵向量全由線性預估倒頻譜係數組成，也就是我們只取線性預估濾波器的資訊來做辨識，所得到的辨識效果大概較純主式架構之一中匹配情形差２％左右；因此只靠這部分的資訊是不足的。另外，在第四列中，我們也評估只使用殘餘信號得到梅爾倒頻譜係數時所得到的辨識效果，也就是ｋ＝０的情況；而由實驗結果知，只靠這部分的資訊根本沒有辨法做辨識（辨識正確率是負值）！但是當這兩部分的資訊結合在同一個特徵向量時，辨識效果開始獲得改進：在圖６．６中我們可以發現，若將殘餘信號的資訊一次一個位元、加入原先只有線性濾波器資訊的特徵向量中，那麼辨識正確率會漸漸提升；到ｋ＝９時，無論高斯混合數為何，可以得到各種組合中的最佳辨識效果，其辨識正確率在高斯混合數＝２、４的時候，甚至可以比基礎實驗所得的結果還要好，而高斯混合數＝８的時候，雖然沒有基礎實驗的辨識效果來得好，但仍然比主從式架構所得到的結果更佳。由此我們可以得到一個結論：如果使用這一節中所提出的方法，**如果以線性濾波器的資訊為主、殘餘信號的資訊為輔，將這兩種資訊以適當的ｋ值結合後，辨識效果就會得到改善。**

●基礎實驗和純主式架構之二的辨識效能比較

表６．２中可以看到，在ｋ＝８、９、１０時，所得到的辨識正確率在某些高斯混合數時可以比基礎實驗來得好。關於這一點是值得探討的：基礎實驗使用的是未做量化的梅爾倒頻譜特徵向量，但是純主式架構之二是使用經過壓縮處理的編碼語音來求特徵向量，按照第四章所觀察到的結果，因為量化錯誤會帶來辨識正確率的下降，我們也可以推論基礎實驗所得到的結果應該要比純主式架構之二要來得好；但事實上並非如此。造成這種情況，除非有其他原因，否則應是純主式架構之二所提出的特徵向量對辨識上的幫助比傳統的梅爾倒頻譜係數來得大，但這個結果有點令人懷疑。為了探索原因，我們進一步來看看基礎實驗以及純主式架構之二中、在ｋ＝９（效果最好）時的辨識結果分析，如表６．３。

表６．３第四欄中音節辨識率，定義為（辨識正確的音節數）／（全部的音節數），而音節辨識正確率＝（辨識正確的音節數－音節插入數）／（全部的音節數），兩者的差別在於插入的錯誤音節數目。

	高斯混合數	音節辨識正確率	音節辨識率
基礎實驗	2	42.64	52.25
基礎實驗	4	49.20	57.49
基礎實驗	8	55.31	62.63
純主式架構之二	2	43.70	45.97
純主式架構之二	4	49.42	51.34
純主式架構之二	8	53.43	55.11

表６．３　基礎實驗及純主式架構之二、ｋ＝９的辨識結果分析。

觀察表６．３，我們可以發現：基礎實驗在音節辨識率上勝過純主式架構之二，但是在音節辨識正確率上卻輸給純主式架構之二。很明顯的，基礎實驗中的統計模型可以辨識出較多的正確音節，但同樣地也辨識出較多額外的音節；也就是說，基礎實驗可以產生較多的辨識結果。讓我們回想：在基礎實驗中所使用的音框平移（Frame Shift），是一般辨識上使用、也是ETSI所制定標準的１０ｍｓ，因此所取出特徵向量的速率是每秒鐘１００個特徵向量；然而在純主式架構之二中，因為調適型多速率聲碼器每２０ｍｓ做一次線性預估分析，所以取出特徵向量的速率是每秒鐘５０個特徵向量。我們可以合理的推論：因為基礎實驗所取出的資料較多，所以可以得到的辨識結果也比較多。為了證實我們的觀點，我們把基礎實驗所使用的音框平移改為２０ｍｓ，所得到的辨識結果分析如表６．４所

示。

	高斯混合數	音節辨識正確率	音節辨識率
基礎實驗（音框平移＝２０ｍｓ）	2	46.98	48.80
基礎實驗（音框平移＝２０ｍｓ）	4	53.33	54.80
基礎實驗（音框平移＝２０ｍｓ）	8	58.51	59.79

表６．４　音框平移＝２０ｍｓ時，所得到的基礎實驗辨識結果分析。

實驗結果恰好驗證我們所提出的論點：當音框平移改為２０ｍｓ時，因為取出的資料量減少，所以辨識出來的結果變少：不僅是音節辨識結果正確的個數減少，連插入的音節數目也減少了。所以在音框平移＝２０ｍｓ的時候，所得到的音節辨識正確率比音框平移＝１０ｍｓ的時候要來得好。另外，我們比較表６．３中的第四、五、六行，以及表６．４，我們發現音框平移＝２０ｍｓ時所做的基礎實驗結果還是比純主式架構之二的結果來的好，這再次說明了量化誤差的確會帶來辨識效能上的下降；同時，這也說明了純主式架構之二的辨識效果會比音框平移＝１０ｍｓ的基礎實驗來得好，不是因為所取出的特徵向量對辨識效果的幫助較大，而是因為資料量不同的緣故。

音框平移＝１０ｍｓ所取出的資料量比音框平移＝２０ｍｓ來得多，理當辨識效果要比較好，可是實驗結果並非如此。可能是我們訓練的統計模型,沒有加入類似持續時間限制(Duration Constraint)的機制，造成資料量變多、卻有較多的插入錯誤。關於這一點，仍待後續研究來論證。

●在沒有通道雜訊干擾的情況下，純主式架構之二與主從式架構的辨識效能比較

在第二章中我們曾提到，目前在分散式語音辨識系統的設計當中，主從式架構以及純主式架構之二都被熱烈的討論著其可行性。實際上該採行何種架構，目前還沒有一致的看法。因此本報告希望能夠從實驗中所觀察到兩個系統的效能來評估何者較為可行。根據表６．２中所示，在很多的情況下純主式架構之二的辨識效能都遠遠勝過主從式架構；但從上一小節的分析中我們知道這個辨識效能上的優越，來自於資料量不同的緣故，因此要比較兩種架構的效能，必須要讓使用相同的資料量才能做比較。從第四章中實驗結果得知，量化的誤差會讓主從式架構的辨識結果較基礎實驗下降２％～３％

左右；如果我們採用音框平移＝２０ｍｓ的基礎實驗為基準，將音節辨識正確率減去３％、當作是以２０ｍｓ為音框平移的主從式架構的音節辨識正確率，則在高斯混合數＝２、４、８時，辨識正確率分別是４３．９８、５０．３３、５５．５１，以上三個數據除了高斯混合數＝８的辨識率較純主式架構之二的結果高出約２％以外，其餘並沒有相差太多。**單從數據上來看似乎主從式架構要好一些些，但是回頭考慮純主式架構之二較主從式架構更節省頻寬，因此我們或許暫時可以說：若使用相同的音框產生率，則兩種架構的效能不會相差太多。**

６．４　結論

在本章中我們主要對兩種純主式架構進行討論；假設無線通道中並沒有錯誤的發生，由實驗所得到的音節辨識正確率來評估兩種純主式架構的可行性。而我們所得到的結論是：

一．關於純主式架構之一，由於受到聲碼器的破壞，合成語音中關於辨識的資訊已經和原始語音大不相同，所以辨識效果會比基礎實驗所得到的結果大約下降６％；

二．關於純主式架構之二，若我們在編碼語音中取出線性預估倒頻譜係數、另外由殘餘信號取出梅爾倒頻譜係數，並且將兩種資訊經過適當結合得到一辨識用的特徵向量，則得到的辨識結果會比基礎實驗（音框平移為１０ｍｓ）的結果要來得好，這顯示純主式架構之二是可行的方式。

第七章　結論與展望

7．1　結論

對於每一個使用無線通信網路的使用者而言，分散式語音辨識系統提供了一個方便使用的介面；透過這個介面，使用者不再被有限的鍵盤所限制，可以下更多、更複雜的指令，如此一來就可以更盡情的存取網路上的資源、享受系統所提供更多的服務。如何在無線通訊的環境中建構分散式語音辨識系統，目前在研究語音的領域中興起廣泛的討論。主要被提出來討論的有兩個架構：主從式架構以及純主式架構之二。在這篇報告當中，我們分別使用這兩種架構、並且模擬無線通道的環境，進行中文大字彙連續語音辨識；然後由實驗的結果討論兩種架構的可行性。在第四章及第五章當中，我們討論主從式架構，在第六章當中，我們則是討論純主式架構。

第四章主要討論在主從式架構下，無線通訊中的量化錯誤所帶給辨識結果的影響。主從式架構是：在用戶端抽出辨識用的３９維梅爾倒頻譜特徵向量，再將這些辨識用的資訊送到伺服器端做辨識。為了節省使用無線通道有限的頻寬，因此在傳輸這些特徵向量前，要

先壓縮這些特徵向量。所以，在主從式架構下，伺服器端所接收到的特徵向量會有量化錯誤的影響。為了將量化錯誤所造成的影響降到最低，因此我們提出的特別的向量量化方式－－分割式向量量化（Split Vector Quantization）：先將特徵向量分割成七個子特徵向量，再分別對每個子特徵向量做向量量化；並且對於辨識上較重要的子特徵向量，給予比較大的碼本大小以強調其重要性。假設無線通道沒有造成任何的錯誤，在這部分的實驗中我們得到幾個結論：

一．必須使用匹配的統計模型，辨識正確率才不會有大幅的下降；也就是說，用來訓練統計模型的特徵向量，同樣也要做向量量化；

二．傳輸資料所需的位元速率（Bit Rate）大約只要４ｋｂｐｓ～５ｋｂｐｓ左右，此時量化誤差所造成的音節辨認正確率下降就算是很小了；

三．傳輸資料的位元速率愈大，辨識正確率會往上昇，量化誤差所造成的影響愈小。

在無線傳輸中，除了上述的量化錯誤外，另外還有無線通道中各種雜訊所造成的錯誤；在第五章中，我們主要討論通道錯誤對辨識結果所產生的影響。我們模擬產生通道中的兩種錯誤－－一種是必然

會發生的隨機錯誤，另一種是無線通訊中最特殊、但造成影響也最嚴重的群集錯誤，並且我們將這兩種錯誤加在由量化代碼所組成的位元串流中；由接收端的伺服器對有錯誤的位元串流進行解碼、得到辨識用的特徵向量後再做辨識。在模擬隨機錯誤影響的實驗當中，我們得到的結論是：當通道發生隨機錯誤時，只有在位元錯誤率高到約10^{-2}左右才會有明顯的辨識正確率下降，若能位元錯誤率控制在10^{-3}以下，那麼隨機錯誤對於辨識結果所造成的影響就可以忽略；此時隨機錯誤所造成的影響會遠小於量化錯誤所造成的影響。而在模擬群集錯誤影響的實驗當中，我們得到的結論是：當群集程度愈大的時候，所帶給辨識效果的影響不一定是愈差的；當位元錯誤率小於或等於10^{-3}的時候，群集程度愈大，辨識的效果會變差；但是當位元錯誤率高到約10^{-2}的時候，群集程度愈大，辨識的效果反而變好。

在第五章中，除了討論兩種錯誤所造成的影響以外，我們也提出了三種錯誤補償的機制，並將這三種機制用在上述兩種通道錯誤干擾的情況，然後由實驗結果討論這三種機制的可行性。假設我們可以正確的偵測到發生錯誤的位元，那麼三種補償機制的作法是：

一．將錯誤位元所在的特徵向量消去，稱基於音框的消去法；

二．將錯誤位元所在的特徵向量移去，以外插法方式產生新的特徵向量，稱基於音框的外插法；

三．將錯誤位元所在的子特徵向量移去，以外插法方式產生新的子特徵向量，稱基於子特徵向量的外插法。

對隨機錯誤做補償的實驗當中，我們發現：除了基於音框的消去法不能補救隨機錯誤所帶來的辨識效果下降外，其他兩種外插法都可以提升辨識正確率，而提升的程度，恰好可以使辨識效果接近沒有受隨機錯誤影響前的辨識效果。因此，如果使用外插法作錯誤補償的機制，那麼隨機錯誤的影響就可以被忽略。而在對群集錯誤做補償的實驗當中，我們使用兩種外插法來評估效果。我們發現，兩種外插法都還是可以產生效果，只不過產生的效果大小不一：基於音框的外插法在群集程度小的時候較不能發揮錯誤補償的功能，隨著群集程度的上升效果才逐漸明顯；但是基於子特徵向量的外插法卻始終保持一定的補償效果，若使用這個方法大致上還是可以逼近未受群集錯誤影響前的辨識正確率。因此，我們了解到群集錯誤所帶給語音辨識比較大的影響，主要不是在於辨識正確率的下降，而是在於錯誤補償機制的難為。綜合以上所言，在第五章中我們得到一個結論：基於子特徵向量的外插法不管是在隨機錯誤、或是群集錯誤發生時，都可以發揮錯誤補償的效果；因此，對每個子特徵向量

進行偵測錯誤的編碼是有必要。

在第六章中，我們對兩個純主式架構進行討論。兩種純主式架構設計的由來，均是希望利用現有的聲碼器來傳輸辨識相關資訊，然後再由這些資訊來進行辨識。在這部分的討論，我們假設無線通道中沒有錯誤的發生。純主式架構之一的觀念，是利用語音解碼器所解得的合成語音來做辨識；而純主式架構之二的觀念，是利用語音編碼器所輸出的編碼語音來做辨識。我們使用 GSM 新一代聲碼器的標準－－調適型多速率聲碼器來進行實驗。在純主式架構之一的部分，我們發現使用合成語音來做辨識，所得到的辨識結果會比用原始語音做辨識的結果來的差，並由此論證聲碼器對語音信號的破壞，以及純主式架構之一之不可行。在純主式架構之二的部分，我們從編碼語音分別抽取兩部份的辨識資訊：線性預估倒頻譜係數、以及殘餘信號所轉換得到的梅爾倒頻譜係數，然後將這兩種資訊各取出一部分，組成１３維的特徵向量。在我們的實驗中發現，當我們以線性預估倒頻係數為主、殘餘信號的梅爾倒頻譜係數為輔時，可以得到比抽取原始信號的梅爾倒頻譜係數相近、甚至更佳的實驗結果。雖然在第六章最後，我們也進一步討論實驗結果更佳的原因，發現這是來自於音框速率的不同，但是純主式架構之二的可行性

已獲得證實。另外，我們也簡略的比較了一下，在無線通道沒有發生錯誤主從式架構以及純主式架構之二的效能：如果使用相同大小的音框來做分析，那麼所得到的辨識效果應該是相差不多。因此我們的結論是：兩種架構所得到的效能相差不多，均是可行的方法。

7．2 展望

對於分散式語音辨識系統，我們所做的研究只是剛開始的一部分，所以接下來的研究方向還是很多。在未來我們計畫完成：

一．純主式架構之二在有無線通道發生錯誤時的辨識效能

在第六章所討論的純主式架構之二，我們是假設無線通道沒有發生錯誤，然後去驗證其可行性。但是我們所研究的分散式語音辨識系統是架構在無線網路的環境，通道中的雜訊干擾對辨識效果所產生的影響是我們所關切的；因此，在接下來的研究中，我們計畫在純主式架構之二中加入第五章模擬的隨機錯誤、以及群集錯誤，再由實驗結果來找出純主式架構之二在無線環境中整體的辨識效能表現及錯誤補償機制，同時也藉由這部分的結果對主從式架構以及純主式架構之二做一個完整的比較

。

二．實際模擬無線通道中各種狀況對辨識效果的影響

在本報告的討論中，無線通道的錯誤是在位元串流中產生，但在無線通道中實際變化所造成的影響我們仍不清楚。或許我們可以藉由無線通道模擬器來模擬這些情況。

三．錯誤補償機制的改進

在本報告中提到使用外插法可以補償無線通道的錯誤所帶來辨識正確率的下降。其實外插法可採的形式很多，我們並沒有做深入研究比較各種方式的效能如何，甚至我們還可以改以內插法的形式來做錯誤補償等等，這些都是可以研究的方向。

四．用戶端的語音強健性

用戶端所持的手持設備常常是在嘈雜的環境中，所以在主從式架構中，用戶端取出的特徵參數都有雜訊的成分存在，這會造成辨識效能的下降。如何消去雜訊、增加信號的強健性也是可以研究的方向。

以上的研究方向已有部分在進行、或是計畫中，期望將來可以有不錯的成果完成。