一種適用於大量連續語料的語音文句校準方法

簡世杰　　張信常

工業技術研究院 資訊與通訊工業研究所
新竹縣竹東鎮中興路四段 195 號 51 館
{ShihChiehChien and Piosn}@itri.org.tw

摘要

為了使維特比演算法（Viterbi Algorithm）能適用於大量連續語料的語音文句校準，以部分語音文句校準循序進行處裡，是一種較有效率的作法，但如何確保整體搜尋空間的最佳路徑落在部份語音和部分文句所形成的部份搜尋空間集合，以及，如何決定落在部份搜尋空間裡的部分最佳路徑，並且該部分最佳路徑與整體搜尋空間的最佳路徑是重疊的，是實施的關鍵。因此，本文提出一種可靠路徑估測方法估測存在於部分搜尋空間裡的可靠路徑，並藉由可靠路徑估測結果調整部分搜尋空間，以防止最佳路徑可能超出部分搜尋空間的情況。實驗顯示本文方法不但可適用在一般無背景噪音的大量連續語料校準，在高 SNR 背景音樂的情況下也能獲得不錯的結果。

1. 前言

在語音信號處理裡，語音文句校準是常見的前處理工作，其目的在取得語音信號與文句內容之間的對應關係，以進行像是語音辨識的聲學模型訓練或是作為語音合成的合成單元使用。一般而言，這類應用所使用的語料通常都是事先依照需要設計的，並且也常以人工方式進行預處理，以使這些經過設計處理過的語料容易以傳統的維特比演算法（Viterbi Algorithm）進行語音文句校準。不過，對於常見的教學錄音帶或是光碟音軌，這些動輒 5 分鐘以上的連續語料以傳統的維特比演算法來進行語音文句校準，記憶空間和運算時間的耗費是相當大的，並且，當連續語料超過一定長度時，傳統的維特比演算法也就不見得能夠適用了。因此，過去對於這種大量連續語料的處理，通常我們會先採用人工分段，再使用傳統的維特比方法進行細部的校準，但這樣也僅能適用在資料量不大的時候，當資料量大時，譬如要對過去傳統的音訊素材全面的進行數位化和再利用，這時候提供一種適用於大量連續語料的語音文句校準方法，取代人工作業，就是一件相當重要的工作了。

對於大量連續語料的語音文句校準，過去的文獻是設法於連續語料裡取得可信賴的錨點（anchor）以分割語料，將大量連續語料分割成較小的語音片段，並再次取得存在於語音片段裡的錨點，直到這些語音片段得以使用傳統方法進行處理為止[2,3]。其中，幾個重要的模組是這種錨點偵測(Anchor Detection)做法所必備的，包括一個語音辨識器以辨識出可能的文句、一個動態規劃（Dynamic Programming）模組比對識別文句與原始文句以取得一致性的文句、以及一個錨點偵測模組配合一些準則自一致的文句內容裡選出錨點。其中，語音識別器的識別能力和錨點的選擇是影響錨點偵測效果的關鍵所在。對於增強語音識別能力，事前可使用一個文句剖析器

依據給定的文句設定識別器使用的識別詞彙和訓練語言模型，以縮小識別範圍和限定前後文接續關係來提昇識別效果。為使錨點選擇具有可靠性，識別文句與原始文句匹配長度達到一定門檻值的錨點選擇準則是常見的作法。然而，當錨點與錨點間的的語音長度小於文句預估的長度；譬如，語音的音框數小於文句的狀態數，即無法順利完成這些錨點之間的語音文句校準。再者，當重複文句出現，識別文句與原始文句匹配就很容易出現問題，這種情形又特別容易出現在語言教學類型的語料裡，也是以這種錨點偵測方式不易克服的地方。另外一個問題是，不同音訊素材的背景環境或收錄所使用的設備可能是不相同的，在這樣的狀況下，相當於是要以固定訓練環境的聲學模型對不同環境的語料進行語音識別，搭配模型調適或者強健式語音識別技術就是錨點偵測做法所要考慮的，其複雜度和難度可見一斑。

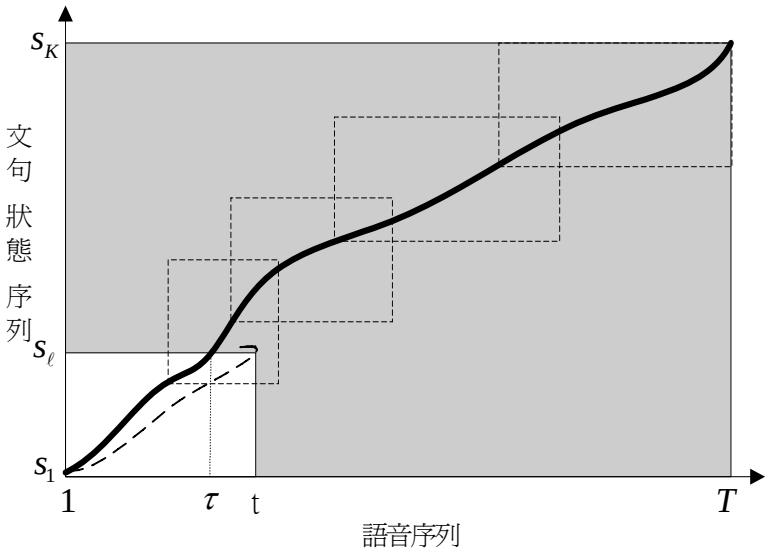
對於已知文句內容的情況下，採用傳統的維特比演算法進行語音文句校準，即便輸入的語音環境可能不同於當時聲學模型的訓練環境，其校準結果通常都仍能具有相當高的準確性。然而，如前所述，傳統的維特比演算法對於大量連續語料的校準有運算速度和記憶體運作上的問題；必須等到所有可能路徑決定以後才能取得存在於整體搜尋空間裡的最佳路徑，難以循序的以部份搜尋結果來進行處理，最大的癥結就在無法確定部分搜尋空間的部份最佳路徑與整體搜尋空間最佳路徑的一致性。因此，本文提出一種可靠路徑估測方法，以找出可能落於部分搜尋空間裡的部分最佳路徑，以部份語音文句校準循序的定出落於部分搜尋空間裡的部分最佳路徑，以使維特比演算法仍能使用於大量連續語料的語音文句校準。實驗顯示該可靠路徑估測方法配合部分語音文句校準，不但可適用在一般無背景噪音的大量連續語料的校準，在高 SNR 背景音樂的情況下也能獲得不錯的結果。

2. 部分語音文句校準問題

對於以部份語音和部分文句對大量連續語料進行語音文句校準，兩個主要問題必須解決：(一) 如何確保整體搜尋空間的最佳路徑落在部份語音和部分文句所形成的部份搜尋空間集合裡；(二) 如何決定落在部份搜尋空間裡的部分最佳路徑，並且該部分最佳路徑與整體搜尋空間的最佳路徑是重疊的。

圖一是一個以部份語音和部分文句進行校準的示意圖，其中灰色區塊為整體搜尋空間，白色區塊表示部分搜尋空間，黑色粗實線表示整體搜尋空間的最佳路徑。在圖一中，s_K 為整體文句狀態序列的最後一個狀態，s_1 為整體文句狀態序列的第一個狀態，s_ℓ 為選定部分文句狀態序列的最後一個狀態，T 為整體語音序列的最後一個音框，t 為選定部分語音序列的最後一個音框，而 τ 則為整體搜尋空間的最佳路徑與選定部分文句狀態序列的最後一個狀態 s_ℓ 所交會對應的語音音框位置。由圖一可以發現，如果自部分搜尋空間的終點 t 才進行回溯取得其最佳路徑，結果將會與整體搜尋結果有相當大的出入（圖一白色區塊的虛線曲線），這也說明：(一) 對於部份語音區間 $[1, t]$ 和部分文句區間 $[s_1, s_\ell]$ 所形成的部分搜尋空間 $\Gamma(t,\ell)$ 並沒有將落在語音區間 $[1, t]$ 的整體搜尋空間的部份最佳路徑完全涵蓋，也就是部分文句區間 $[s_1, s_\ell]$ 是不足的；(二) 對部分文句區間 $[s_1, s_\ell]$ 而言，所使用的部分語音區間 $[1, t]$ 是過長的，適當的語音長度應小於 τ。圖一顯示，最佳路徑在語音序列 τ 之後已超出預設的部分文句區間 $[s_1, s_\ell]$。因此，除非在搜尋的同時可隨時掌握該最佳路徑的語音及文句的適當範圍，否則，也只能不斷擴大部份搜尋空間以保證整體搜尋空間的部份最佳路徑落在該部份搜尋空間裡，最極致的情況就是部份搜尋空間

等於整體搜尋空間，而這也是我們所不樂見的情況。顯然的，如何決定存在於部分搜尋空間裡的最佳路徑，並且使得該部分最佳路徑與整體搜尋空間的最佳路徑是重疊的，是解決上述問題的關鍵。因此，下一節，我們將介紹一個可靠路徑估測演算法來解決上述問題。



圖一：部分語音文句校準示意圖

3. 可靠路徑估測

可靠路徑估測是以維特比演算法為基礎的作法，並且配合最大相似度來施行，因此，以下我們先對基本的維特比演算法和其特性做一簡單描述。

3.1. 維特比演算法

假設由語音序列 $X_T=(x_1x_2..x_T)$ 和文句狀態序列 $S_K=(s_1s_2..s_K)$ 所構成的搜尋空間 $\Gamma(T,K)$，存在一個與 X_T 相對應的最佳狀態序列 $Q_T=(q_1q_2...q_T)$，以最大相似度(ML，Maximum Likelihood)來進行比對時，我們可以在時間t得到對應到文句狀態 s_i 的最佳狀態序列為 $Q_t=(q_1q_2...q_t)$，若將此時的相似度分數定義為

$$\delta_t(s_i)=\max_{q_1,q_2,...,q_{t-1}}\Pr[q_1q_2,...q_{t-1},q_t=s_i,X_t\,|\,\lambda],\ 1\le i\le K \tag{1}$$

則可將t+1時間落在文句狀態 s_j 的相似度分數表示為

$$\delta_{t+1}(s_j)=\max_{1\le i\le K}\left[\,\delta_t(s_i)a_{ij}\,\right]b_j(x_{t+1}),\quad 1\le j\le K \tag{2}$$

其中，λ 為比對所使用的聲學模型，a_{ij} 為狀態 s_i 轉移到狀態 s_j 的轉移機率，$b_j(x_{t+1})$ 為語音 x_{t+1} 在狀態 s_j 的機率分布。不斷反覆Eq.(2)直到語音序列終點T，可得終點T的相似度分數為

$$\max_{1\le k\le K}\left[\,\delta_T(s_k)\right] \tag{3}$$

也就是可由終點T裡選出一個具有最大相似度分數的文句狀態 s_k 與語音序列終點T對應。之後，可由語音序列終點T與文句狀態 s_k 的交會點 $\varphi(T,s_k)$ 回溯至搜尋空間 $\Gamma(T,K)$ 原點 $\varphi(1,s_1)$ 得到最佳

路徑 $path(T,s_k)$。(詳細演算法可參考[1])。

3.2. 維特比演算法特性

維特比演算法以最大相似度來施行，存在有以下兩個特性：

特性一：假設部分搜尋空間 $\Gamma(t,\ell)$ 存在一終止於時間 t 對應到文句狀態 s_i 的路徑 $path(t,s_i)$，該路徑為搜尋空間 $\Gamma(T,K)$ 最佳路徑的一部份。若在 $\Gamma(t,\ell)$ 裡有另一路徑 $path(t,s_j)$ 與 $path(t,s_i)$ 在 $\Gamma(t,\ell)$ 空間裡有一對應到時間 τ 與文句狀態 s_n 交會點 $\varphi(\tau,s_n)$，則由交會點 $\varphi(\tau,s_n)$ 至搜尋空間的原點 $\varphi(1,s_1)$ 之間的最佳路徑 $path(\tau,s_n)$ 必為 $path(t,s_i)$ 的一部份，也必為 $\Gamma(T,K)$ 最佳路徑的一部份。

說明：由於路徑 $path(t,s_i)$ 為 $\Gamma(T,K)$ 最佳路徑的一部份，而 $\varphi(\tau,s_n)$ 又位在 $path(t,s_i)$ 中，因此，由 $\varphi(\tau,s_n)$ 所決定出來的最佳路徑 $path(\tau,s_n)$ 必然是屬於 $\Gamma(T,K)$ 最佳路徑的一部份。

特性二：由 $\varphi(\tau,s_n)$ 所決定出來的最佳路徑 $path(\tau,s_n)$ 必定只有一條，且必定是使得 $\varphi(\tau,s_n)$ 所在位置的相似度分數 $\delta_\tau(s_n)$ 最大的狀態序列 $Q_\tau=(q_1q_2...q_{\tau-1},q_\tau=s_n)$。

說明：維特比算法以最大相似度來施行，每一時間對應到每一狀態的相似度分數都是最大的，且必定僅有一最佳狀態序列與之對應。

3.3. 可靠路徑估測

由於以最大相似度來施行，最佳路徑終點通常都具有較大的相似度分數，因此，如果最佳路徑落在部分搜尋空間 $\Gamma(t,\ell)$ 裡，則在時間 t 由部分文句狀態序列 $S_\ell=(s_1s_2...s_\ell)$ 選出 N 個具有較大相似度分數的狀態，是非常有可能將最佳路徑的終點狀態涵蓋進來。以這些具有較大相似度分數的狀態所回溯出來的路徑就有可能包含 $\Gamma(t,\ell)$ 的最佳路徑。所以，假設這些路徑中包含有最佳路徑，並且這些路徑有共同的交會點，依照特性一，該交會點必定落在最佳路徑裡，並且，由交會點回溯到 $\Gamma(t,\ell)$ 原點的路徑，必定為最佳路徑的一部份，依照特性二，由交會點所回溯的最佳路徑僅有一條，所以可由任一經過交會點的路徑取得該最佳路徑。由於是以 N 個具有較大相似度分數的狀態來取得部分搜尋空間 $\Gamma(t,\ell)$ 裡可能的最佳路徑，我們稱這個取得最佳路徑方式為可靠路徑估測。

可以了解的是，當 N 越大涵蓋最佳路徑終點狀態的可能性就越大；反之，當 N 小時，就不一定保證能涵蓋最佳路徑的終點狀態，尤其當訓練聲學模型所使用的語音語料庫與待估測語音的特性差異很大時，可能就必須加大 N 的數量，以容忍不同的語音環境。另外，由於我們以文句狀態序列來與語音序列進行較準，也就是說，無論是中文的音節、英文的單詞或是存在於語音序列裡的靜音都轉化為狀態序列來表示(譬如：以 3 個聲母狀態和 5 個韻母狀態來表示一個含 8 個狀態的中文音節、以 3 個狀態來表示英文音素及以多個音素來描述一個英文單詞、以及以 1 個可有可無的靜音狀態來表示可能存在於語音序列裡的靜音等，將文句序列轉化為文句狀態序列來表示)，因此，所決定之可靠路徑之端點就不一定是中文的音節端點、英文的單詞端點或靜音處，亦可能是存在於中文音節、英文單詞的內部狀態位置，或是靜音狀態位置。不過，只要能決定出可靠路徑，藉由可靠路徑資訊取得語音序列與文句狀態序列的對應關係，無論是中文的音節端點、英文的單詞端點或是存在於語音序列裡的靜音位置，都是可以決定的。即便是使用者自行標示的文句段落位置，亦可藉由該可靠路徑資訊來取得其所對應的語音序列位置。

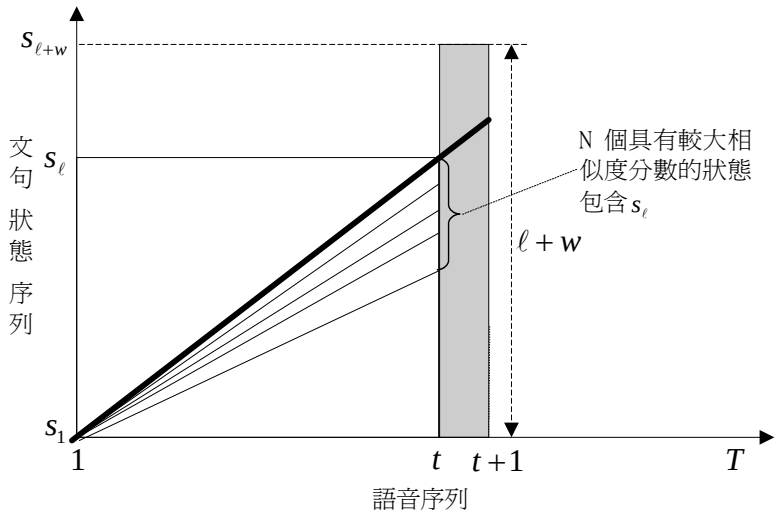
3.4. 部分搜尋空間調整

調整部分搜尋空間的情況共有以下二種：

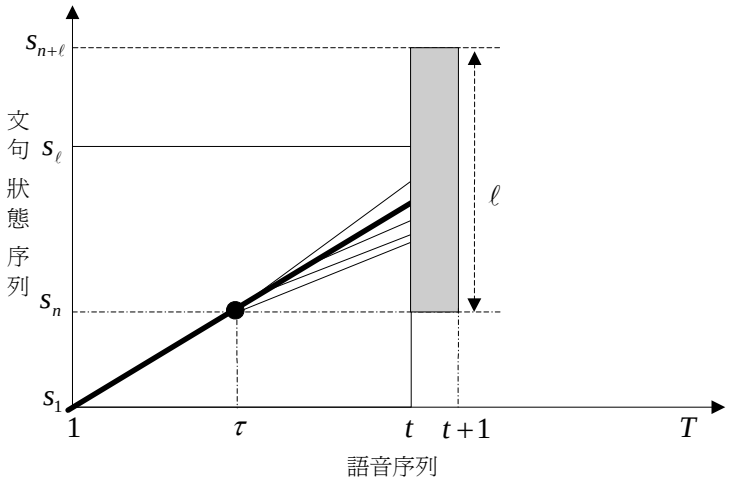
調整一：在部分搜尋空間中未找到可靠路徑，加大部分搜尋空間之文句狀態，以防止可能的最佳路徑在下一時間比對時落在部分搜尋空間外。如圖二，時間 t 所取得的 N 個具有較大相似度分數的狀態集合中包含部分文句狀態序列 $S_\ell = (s_1 s_2 ... s_\ell)$ 的最後一個狀態 s_ℓ，則在下一個時間 t+1 進行比對之前，應加大部分搜尋空間的文句狀態序列為 $\ell + w$，以防止如圖一情況，可能的最佳路徑落在部分搜尋空間之外。

調整二：一旦在部分搜尋空間裡已取得可靠路徑，由可靠路徑終點，我們可重設下一次比對的部分文句序列。圖三是一個在未加大文句狀態序列下，在時間 t 取得可靠路徑，在 t+1 時間保持下一次比對的部分文句序列為 ℓ 個狀態。圖四則是文句狀態序列加大為 $\ell + w$ 後，在時間 t 取得可靠路徑，在去除 n 個文句狀態之後，剩餘的文句狀態 $\ell + w - n$ 大於 ℓ，t+1 時間調整部分文句序列為 $\ell + w - n$ 個狀態。

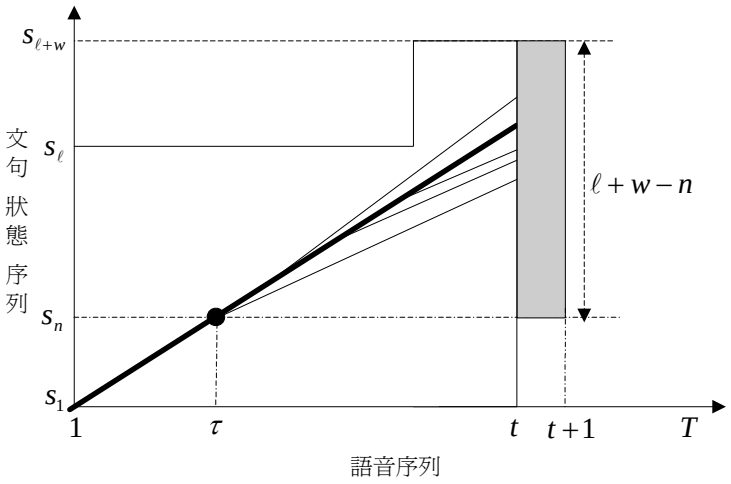
因此，藉由可靠路徑估測和上述搜尋空間的調整，我們就可以循序的以部分語音文句校準完成大量連續語料的語音文句校準工作。圖五是以部分搜尋空間來涵蓋整體搜尋空間 $\Gamma(T, K)$ 的最佳路徑示意圖。以過去傳統作法進行語音文句校準，需要對整體搜尋空間 $\Gamma(T, K)$ 做運算之後才得以決定整體搜尋空間的最佳路徑，以本文所介紹的方法則可以省去如圖五灰色區間的運算量。當整體搜尋空間不大時，本文作法與傳統作法所需的運算量或許差異不大，但是當整體搜尋空間變大，如動輒 5 分鐘以上的教學錄音帶或是光碟音軌，本文作法可節省相當多的運算量，是一種相當有效率的作法。



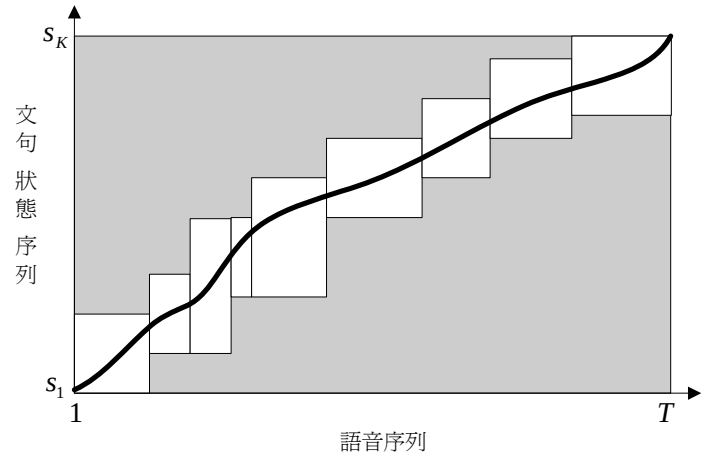
圖二：於語音序列時間 t 選出 N 個具有較大相似度分數的狀態集合，該集合含有部分文句狀態序列的最後一個狀態，於 t+1 時間加大文句狀態序列為 $\ell + w$



圖三：語音序列時間 t 取得可靠路徑，當去除可靠路徑所含 n 個文句狀態序列之後，若剩餘的文句狀態少於 ℓ 個狀態，於 t+1 保持文句狀態序列為 ℓ 個狀態



圖四：語音序列時間 t 取得可靠路徑，當去除可靠路徑所含 n 個文句狀態序列之後，若剩餘的文句狀態 $\ell+w-n$ 大於 ℓ 個狀態，於 t+1 保持 $\ell+w-n$ 個文句狀態



圖五：以部分搜尋空間集合來涵蓋整體搜尋空間 $\Gamma(T,K)$ 的最佳路徑示意圖

3.5. 部分語音文句校準演算法

以下將部分語音文句校準演算法整理如圖六。

1. 取出部分文句狀態序列$S=(s_1 s_2 ... s_\ell)$，並初始文句狀態序列的機率分數$\{\delta_1(s_1),\delta_1(s_2),...,\delta_1(s_\ell)\}$爲極小值；
2. 循序取出每一語音音框t進行下列步驟：
   2.1 依照Eq.(2)，但限制$1\le i\le \ell$和$1\le j\le \ell$，決定每一文句狀態序列的機率分數$\{\delta_t(s_1),\delta_t(s_2),...,\delta_t(s_\ell)\}$，並紀錄其前一語音音框t-1的最佳狀態位置；
   2.2 依照機率分數選出N個具有較大機率分數的文句狀態；
   2.3 以t和N個具有較大機率分數的文句狀態進行回溯，求出路徑的交會點$\varphi(\tau,s_n)$；
   2.4 可靠路徑資訊紀錄和搜尋空間調整：
      2.4.1 若$\varphi(\tau,s_n)$非部分搜尋空間的原點$\varphi(1,s_1)$，則由$\varphi(\tau,s_n)$回溯至$\varphi(1,s_1)$的路徑爲可靠路徑，輸出可靠路徑數據，依照部分搜尋空間調整二，重設部分文句狀態序列，並初始文句狀態序列裡新增加狀態的機率分數爲極小值；
      2.4.2 若$\varphi(\tau,s_n)$爲部分搜尋空間的原點$\varphi(1,s_1)$，則無可靠路徑，若N個具有較大機率分數的文句狀態含部分文句序列最後一個狀態，依照部分搜尋空間調整一，加大部分文句狀態序列，並初始文句狀態序列裡新增加狀態的機率分數爲極小值；
   2.5 若尚有語音信號，重複進行2.1至2.4步驟；
3. 剩餘的部分搜尋空間以部分搜尋空間的語音信號終點和文句狀態序列終點進行回溯，求出其最佳路徑並輸出最佳路徑數據。

圖六：部分語音文句校準演算法

由於機率分數隨著語音信號不斷累積可能會產生溢流（Run-Off）問題，因此，可在重設部分文句狀態序列時，將累積的機率分數進行重設，以避免溢流情況發生。

4. 實驗與結果討論

4.1. 測試語料

我們使用下列幾套語料來驗證上述作法，並以語句邊界的正確率做為語音文句校準的評估。

語料一（DB1）：來自工研院的104自動總機系統[4]所蒐集的人名語音，這些語音都是以8 KHz、16-bit、Mono格式所錄製的電話語音，並且可能夾雜一些背景雜訊，如打字的聲音、話筒撞擊聲或有說話的背景等。我們將其中的751句串接成一約23分15秒的長串語音，共包含2247個音節，句子與句子之間我們以一個分隔符號來作區隔。我們分別以傳統的維特比演算法定出這751句人名語音每一句語音的語音起點和終點邊界，即去除靜音部分，並且求得這些語音邊界對應長串語音的絕對位置作為正確的邊界位置，共包含1502個邊界。

語料二（DB2）：來自教學用的語音光碟[5]，這些語音都是以44.1 KHz、16-bit、Stereo格式儲存。我們取出其中4段音軌，並且將語音格式轉換為8 KHz、16-bit、Mono格式，之後，將這些語音串接成一約23分48秒的長串語音，共包含5175個音節。我們以人工方式將該語音資料分成421句，之後，如語料一的處理方式，插入分隔符號於句子與句子之間，並分別定出每一句語音的語音起點和終點邊界和其所對應長串語音的絕對位置作為正確的邊界位置，共包含842個邊界。

含背景音樂的語料（DB1+MU, DB2+MU)：將上述數語料（DB1, DB2）以語料裡的平均音量為基準，加入對應強度約10 dB SNR的古典樂（四季，維瓦第)，以觀察受背景音樂干擾的語音文句校準情形。同樣使用DB1和DB2所求得的邊界位置為正確答案來進行評估。

4.2. 聲學模型

實驗所使用的聲學模型係自MAT電話語音語料庫[6]訓練而得，語音語料庫的格式為8 KHz、16-bit、Mono。聲學模型共含100個右相關聲母模型、38個左右無關的韻母模型和1個靜音模型。每個聲母模型以3個狀態來描述，韻母模型以5個狀態來描述，靜音模型則僅以1個狀態來描述。聲、韻母模型狀態使用10個混合數，靜音使用64個混合數。

除了上述的聲學模型之外，在進行校準時，我們也使用具有 64 個高斯混合數的語音/非語音模型來進行語音/非語音判斷，並將判斷為非語音的樣本收集起來訓練一個僅有一個混合數的背景模型。將該背景模型與前述靜音模型合併成 65 個混合數，作為部分語音文句校準時吸收靜音和背景雜訊之用。

4.3. 實驗條件設定和語音邊界偵測

不加背景音樂的語料（DB1, DB2)，我們以N=40來估測可靠路徑。加了背景音樂的語料（DB1+MU, DB2+MU)，則以N=80來估測可靠路徑。

部分文句狀態序列長度 ℓ 設為 20 個音節（也就是 180 個狀態，含音節與音節之間的一個靜音狀態)。加大文句狀態序列的序列長度 w 也是使用 20 個音節來防止最佳路徑可能落在部分搜尋空間之外的情況。

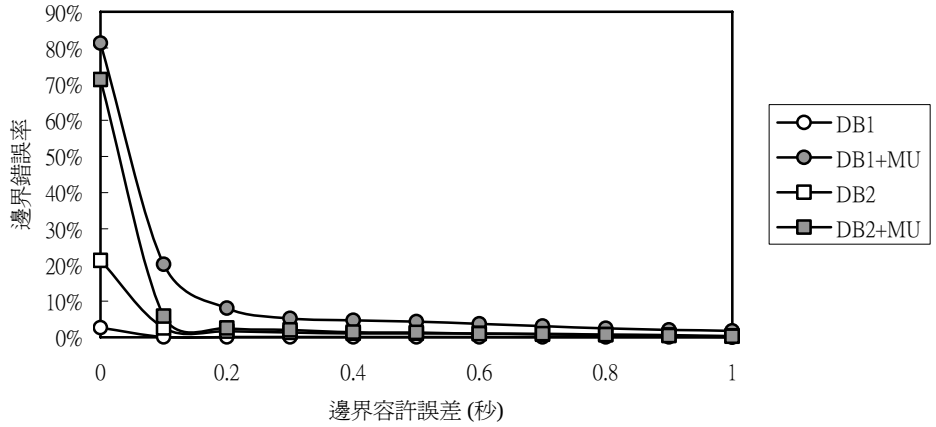
在取得可靠路徑之後，檢驗可靠路徑所對應的文句序列中是否含有分隔符號，若含有分隔符號，我們以下列方式定出前一句語音的終點邊界和後一句語音的起點邊界：(一）若分隔符號所對應的語音位置含有靜音，靜音之前為前一句語音的終點邊界，靜音之後為後一句語音的起點邊界；(二）若不含靜音，則前一句語音的終點邊界和後一句語音的起點邊界為同一邊界。另外，第一句語音的起點邊界和最後一句語音的終點邊界都以不包括靜音部分為其邊界。

4.4. 結果與討論

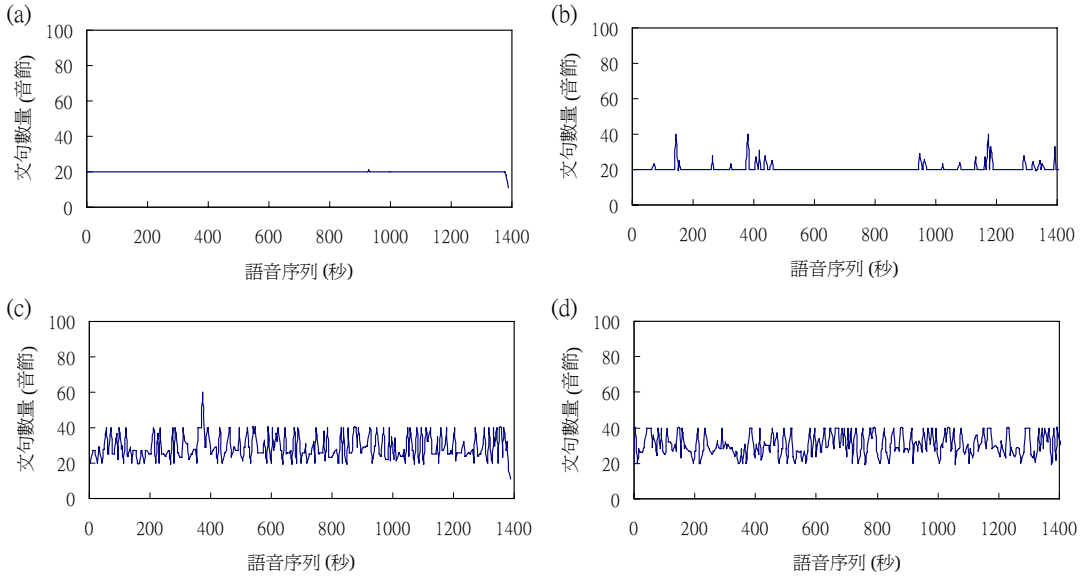
以前述條件進行實驗，4組語料都可以得到正確的邊界數量；DB1和DB1+MU可得到1502個邊界，DB2和DB2+MU可得到842個邊界。圖七顯示4組語料進行語音文句校準之後得到的邊界與正確邊界在不同容許誤差的邊界錯誤率分布情形。不加背景音樂語料DB1和DB2之偵測邊界與正確邊界有相當高的吻合度，尤其是DB1與聲學模型的訓練語料庫有較一致性的語音特性(都是電話語音)，在0.1秒的容許誤差下，可以得到幾近於零錯誤的結果(0.07%)。加入背景音樂語料DB1+MU和DB2+MU之偵測邊界與正確邊界雖然有較大的差異，但是在1秒的容許誤差下，仍然可以得到相當低的邊界錯誤率(DB1+MU：1.73%，DB2+MU：0.24%)。其中，可以注意到DB1+MU的邊界錯誤率比DB2+MU高，這是由於DB1語料庫是由不同說話人的語音串接而成，每一句語音的音量不固定，當加入固定強度的背景音樂時，每一句語音的SNR就不一定是我們所設定的10 dB。而DB2係來自教學用的語音光碟，雖然也有不同說話人的語音，但是出版者為控制其語音品質，DB2裡每一句的音量顯然是較為平均的，這也使得DB2+MU的SNR整體上較接近於我們所設定的10 dB。因此，DB1+MU的語音條件顯然較DB2+MU差一些，也使得其邊界錯誤率較高。

圖八是 4 組語料的比對歷程，在 N=40 的條件下，DB1 和 DB2 每秒約處理 20 個音節，佔整體

搜尋空間的 1%以下。在 N=80 的條件下，DB1+MU 和 DB2+MU 每秒約處理 30 個音節，佔整體搜尋空間的 1.5%以下。兩者皆具有相當高的執行效率。上述實驗都是以一般的電腦系統來執行(AMD 1.0G Hz CPU, Windows 2000)，4 組語料庫都可以在 1 個實體時間內(RT，Real Time)完成切割(N=40 約 0.6 RT 完成，N=80 約 0.98 RT 完成)。



圖七：語音文句校準之後得到的邊界與正確邊界在不同容許誤差的邊界錯誤率分布



圖八：四組測試語料比對歷程

(a) DB1, N=40; (b) DB2, N=40; (c) DB1+MU, N=80; (d) DB2+MU, N=80

5. 結論與未來方向

本文提出了一種可靠路徑估測演算法，以找出可能落於部分搜尋空間裡的部分最佳路徑，以部份語音文句校準循序的定出落於部分搜尋空間裡的部分最佳路徑，以使維特比演算法仍能使用於大量連續語料的語音文句校準。實驗顯示該演算法的穩定性和使用部分校準的效率，其不但可適用

在一般無背景噪音的大量連續語料的語音文句校準，在高SNR背景音樂的情況下也能獲得不錯的結果。

雖然本文提出的演算法可有效處理大量連續語料校準,但是偵測邊界仍會與實際邊界有一些差異，在要求高精度的語料庫處理上仍不免需要人工檢驗，如何進一步提高切割精度；或者，標示出偵測邊界的信心度，以降低人工檢驗和人工校正的負擔，將是我們未來的工作重點之一。

6. 計畫相關資訊

本文係工研院資通所執行經濟部九十五年度前瞻研究專案 5301XS2310 的計畫成果之一。

7. 參考文獻

1. Rabiner L. and Juang B.-H., “Fundamentals of Speech Recognition,” New Jersey, Prentice-Hall International, 1993, pp. 339-340.
2. Robert-Ribes J. and Mukhtar R.G., “Automatic Generation of Hyperlinks Between Audio and Transcript,” Eurospeech, 1997.
3. Moreno P.J., Joerg C., Van Thong J.-M., and Glickman O., “A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments,” ICSLP, 1998.
4. 謝偉強、簡世杰、許志興、張森嘉，”工研院 104 自動總機系統的改進過程”，電腦與通訊，2001 年，第 96 期，pp. 29-34.
5. 康軒文教事業，”TOP945 兒童雙週刊中年級版第 8 期”，http://top945.knsh.com.tw，2003 年.
6. Wang H.-C., “MAT — A Project to Collect Mandarin Speech Data Through Networks in Taiwan,” Computational Linguistics and Chinese Language Processing, Vol. 2, No. 1, pp. 73-89, 1997.