整合聲學指引規則至 HMM 最佳路徑搜尋之歌聲分段方法

Singing-voice Signal Segmentation Methods Integrating Acoustic-guiding Rules into HMM Based Best-path Searching

古鴻炎[1]、許瓊之[2*]

Hung-Yan Gu、Cyong-Zhii Syu

[1]作者一，國立台灣科技大學 資訊工程系 教授

[2]作者二，國立台灣科技大學 資訊工程研究所 研究生

E-mail: guhy@csie.ntust.edu.tw, Tel: (02)27376684

摘要

本研究對於歌聲信號裡聲、韻母時間位置之自動分段的問題，提出了一種整合聲學指引規則至 HMM (hidden Markov model)維特比(Viterbi)解碼中作最佳路徑搜尋的方法，可用以顯著提升基於 HMM 之基本分段方法的準確率。實作上我們製作了三種版本的自動分段程式，分別使用不同的維特比解碼演算法，來作相互的效能比較。在訓練 HMM 之階段，使用了 HTK 軟體對 TCC-300 語料庫中選出的語句，去訓練出聲、韻母 HMM 模型；然後透過強制對齊，對自備的歌聲語料，分析各聲、韻母 HMM 之各狀態上的駐留時長參數，如此就可帶入伽瑪(gamma)機率分佈，去計算外顯式狀態時長機率。在測試階段，實驗的結果顯示，使用外顯式狀態時長機率之修正的維特比解碼可以比基本維特比解碼在 10 ms 之容忍度內提升 7.55% 的準確率；進一步依各音框偵測出的基頻值與能量值，並依聲學知識去設計聲、韻母相關的限制規則，再把規則整合至維特比解碼的步驟中，此方法比起基本的維特比解碼方法，可讓準確率從 31.73% 提升到 61.33%；接著再把聲、韻母 HMM 駐留時長的限制規則整合進去，則可讓準確率再提升至 66.86%；此外若再加入一種靜音相關的後處理步驟來更正聲、韻母邊界，則準確率更可提升到 68.45%。

關鍵詞：歌聲自動分段、隱藏式馬可夫模型、維特比解碼、聲學指引規則、外顯式狀態時長

Abstract

In this paper, we propose singing voice signal segmentation methods that integrate acoustic-guiding rules into HMM (hidden Markov model) based best-path searching to greatly improve the segmentation accuracies for HMM based segmentation of syllable initials and finals. In practice, we have programmed three versions of Viterbi decoding algorithms for automatic segmentation of initials and finals, and then compare their performances. In the training stage, the software package, HTK, is used to train syllable initial and finial HMM models with some selected sentences from TCC-300 corpus. Next, we estimate the state-duration parameters of each HMM state by means of forcedly aligning our recorded singing voice signals. Then, the parameters of state durations can be used to calculate gamma distribution based explicit state-duration probabilities. In the testing stage, the results of the experiments show that the Viterbi decoding algorithm using explicit state duration probability can obtain the segmentation accuracy rate which is 7.55% higher than the Viterbi decoding algorithm using implicit state transition probability under the tolerance range of 10 ms. Furthermore, we base on the detected fundamental frequency and energy from each frame to

design acoustic-knowledge related constraint rules, and integrate these acoustic-guiding rules into the improved Viterbi decoding algorithm. By using this Viterbi decoding algorithm, the segmentation accuracy rate can be promoted from 31.73% to 61.33% as compared with the basic Viterbi decoding algorithm. In addition, if we integrate more rules to constrain the duration of syllable initials and finals, the segmentation accurate rate is raised to 66.86%. Finally, we add a post-processing step that adjusts the boundaries of syllable initials and finals according to the detected silence frames. As a result, the segmentation accuracy rate is further raised to 68.45%.

Keywords: automatic singing-voice segmentation, hidden Markov model, Viterbi decoding, acoustic-guiding rule, explicit state duration

一、導言

在語音信號處理之領域,一個語句中各音素(phoneme)的左右邊界所在的時間位置,是很重要的資訊,若可取得正確的音素邊界之時間資料,則可訓練出比較穩定的模型,進而改進語音處理系統的效能(如語音辨識之正確率)。實務上,以人工來標記音素邊界之時間位置的語料庫並不多,因為採取人工標記的作法,不僅會耗費大量的人力與時間,並且會因不同標記者對於時間邊界認定的不同而缺乏一致性。目前,強迫對齊(forced alignment)[1]之自動標記(或稱為自動分段)方式雖然較為快速,但卻存在自動標記出的時間位置不夠準確的問題。因此,一個能夠作自動標記、且能準確標記出音素時間位置的軟體,是語音處理的一個重要組件(或工具),它可以為各種語音處理相關的研究建立語料庫。

雖然前人已提出多種音素自動分段(segmentation)的研究成果,且已達到相當高的正確率,但大多是針對語音信號,而對於歌聲信號作音素分段的研究成果仍嫌不足。由於歌聲信號中經常有節奏(tempo)以及旋律(melody)的變化性,而使得歌詞音素的時間位置很難被正確地偵測出來,因此本論文的研究重點是,發展一個可對歌聲信號作自動聲、韻母位置標記的方法與系統,如此,輸出的標記檔就可被用於作歌詞聲、韻母信號之自動切割與擷取,或應用於人形機器人之對嘴唱歌之表演。

HTK[2]是一套以 C 語言寫成的免費工具軟體,研發人員可用它來訓練隱藏式馬可夫模型,並可用它來製作語音辨識系統。HTK 軟體主要可以分成兩個階段,分別為訓練階段與辨識階段。在訓練階段中,HTK 會對預先準備好的訓練語料作頻譜特徵參數的抽取,即計算 MFCC 係數,接著可再執行 HTK 批次命令去訓練聲、韻母 HMM 模型。在辨識階段,則可用 HTK 提供的命令執行維特比演算法作搜尋比對,計算出詞典檔中所有詞彙可能形成的詞彙網路所對應的各條 HMM 模型序列的機率值,並從當中找出機率值最高之 HMM 序列所對應的詞彙序列來作為辨識結果。在本研究的訓練階段,我們也是使用 HTK 來建立國語聲、韻母的 HMM 模型[3];但是在分段階段則是使用自製的程式,而未使用任何的 HTK 模組。

本論文的研究方法是,拿說話語音信號去訓練出的 HMM 模型作為基礎;然後對少量的歌聲信號作聲、韻母單位之強迫對齊,以統計各個聲、韻母 HMM 狀態的駐留時長平均值與標準差,如此求得的平均駐留時長標準差與變異數值,就可帶入伽瑪(gamma)機率分佈,用以計算外顯式(explicit)狀態時長機率,以改進原先 HMM 所使用內顯式(implicit)狀態移轉機率; 此外,我們更進一步研究聲、韻母相關的聲學規則,並且提出了把聲學規則整合至 HMM 維特比解碼(Viterbi decoding)演算法的作法,如此用以改善原始 HMM 模型作自動分段之正確率太低的問題。

二、HMM 模型訓練

對於HMM模型之訓練，其流程如圖1所示，圖1的第一個方塊會對TCC-300語料庫中的語句，使用吳昌益先生發展的程式[4]去萃取頻譜特徵參數，也就是梅爾頻率倒頻譜係數(mel-frequency cepstral coefficient，MFCC)；接著在第二個方塊，採用HTK工具軟體來訓練HMM模型，在訓練程序中我們會把第一個方塊所抽取出的MFCC係數拿去取代HTK計算出的MFCC係數，原因是在圖2及圖3裡，我們對所輸入的音檔都要自行計算MFCC係數，而不使用HTK所抽取的MFCC係數。之後，經由HTK的訓練程序就可得到聲、韻母的HMM模型。

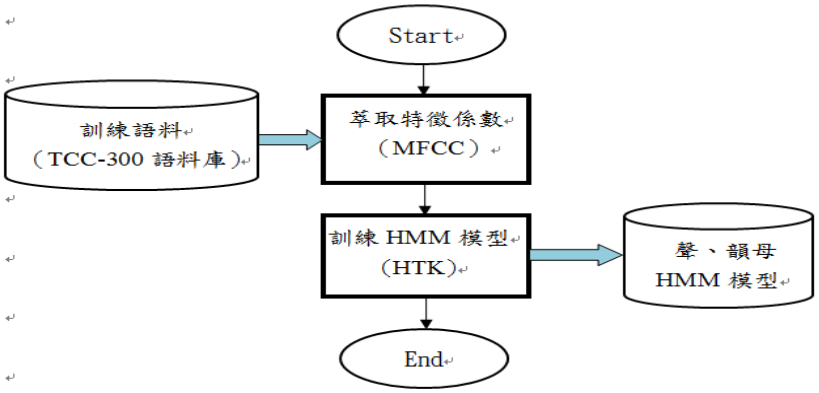

圖1　**HMM模型訓練流程**

我們使用 TCC-300 語料庫[5]來訓練聲、韻母HMM，TCC-300語料庫之錄音，是以麥克風唸讀華語語句的說話方式，原先是作為華語語音辨認研究的用途。在此，我們使用TCC-300語料庫的部分語句，詳細情形如表1所示。

表1　HMM訓練語料之資訊

語料格式	**語料內容**
語料來源	TCC300 麥克風語料庫
語料種類	語音
語料音檔格式	取樣率 16KHz，解析度 16bits
錄音人數	300 人
語料語句數	3, 000 句
語料音節數	42, 018 個

(一)使用 HTK 訓練 HMM 模型之流程

HMM模型是一種具有狀態觀念的統計模型，應用於語音信號處理時，HMM是用來描述一個語音單元發音過程的頻譜特性變化，所以必須限制其狀態轉移方式為由左至右，或是停留在原狀態。我們使用HTK來訓練國語聲、韻母的HMM模型，其訓練流程如圖2所示。

(二)語料預處理 至 MFCC 係數計算

在使用HTK訓練HMM模型之前，需要先準備兩種檔案，第1種是每一個語句音檔所對應的說話內容之mlf檔(HTK規定之名稱)，mlf檔案內記錄音檔的說話內容之拼音符號；第2種則是phoneme.pam，此檔案記載各個聲學HMM模型之代號，用來對上述mlf檔裡的羅馬拼音作拆解，在此我們將國語的聲、韻母單元看成是待建立的聲學HMM模型。

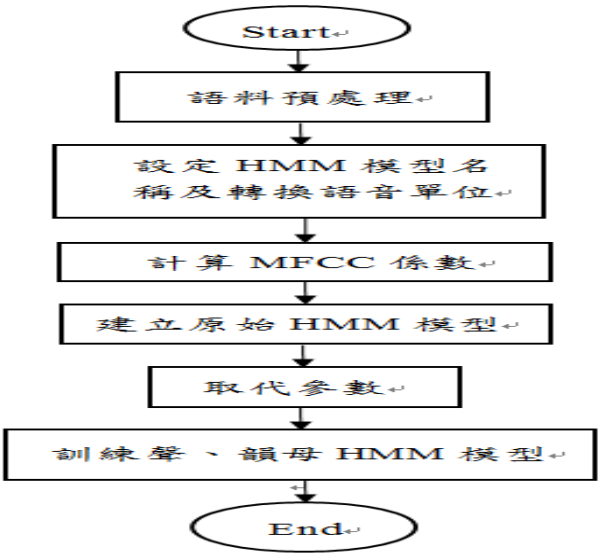

圖2　使用HTK訓練HMM模型之流程

準備好mlf檔及phoneme.pam之後，就可以執行HTK的批次命令來訓練聲、韻母HMM模型，在圖2的第二個方塊，執行HTK指令HLEd就可把音節為單位的mlf檔轉換成聲、韻母為單位的mlf檔，並且產生所有會使用到的聲、韻母HMM模型名稱。在圖2的第三個方塊，進行語音頻譜特徵參數的抽取，在此我們使用HTK的HCopy指令，來求取MFCC係數。

(三)建立原始 HMM 模型 至 訓練聲、韻母 HMM

在圖 2 的第四個方塊，要先建立 HMM 模型結構之列表，以記錄模型結構的參數，例如一個 HMM 的狀態個數及一個狀態有多少個高斯(Gaussian)混合個數數等。接著，在圖 2 的第五個方塊，對模型的高斯混合個數作擴增，並且會將 HTK 所擷取出的 MFCC 係數取代為圖 1 第一個方塊所抽取的 MFCC 係數。

在最後一個步驟，即圖 2 的第六個方塊，執行 HTK 的 HERest 指令，來跑 EM (expectation maximization)演算法[3]，我們作了 30 次的反複估計處理，以求取 HMM 模型參數，如此就可完成聲、韻母 HMM 模型的訓練。

三、HMM 狀態之外顯式駐留時長

關於 HMM 狀態駐留時長之分析，其流程如圖 3 所示，首先從歌聲語料庫輸入各樂句之音檔，接著使用相同於圖三的 MFCC 求取程式，去算出各音框的 MFCC 係數;然後依據各樂句之標記檔去選取對應的聲、韻母 HMM 模型，據以作維特比解碼之比對，以求得各個聲、韻母 HMM 狀態的駐留時長平均值與變異數，作為 HMM 狀態駐留時長之參數。

(一)歌聲語料資訊

對於圖 3 裡狀態駐留時長參數的估計，我們從歌聲語料中選取 40 首歌、並且切割成各樂句的音檔，再作維特比解碼，以計算外顯式狀態時長機率分佈的參數，此部分歌聲語料的詳細情形如表 2 所示。

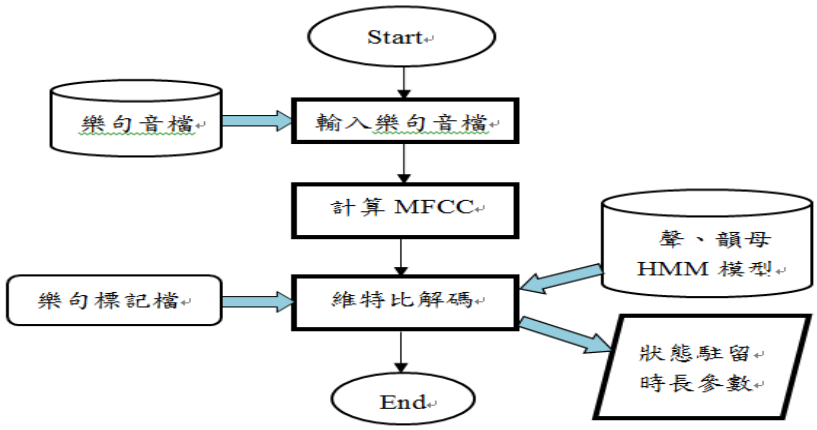

圖3　狀態駐留時長參數之估計流程

表 2　估計狀態駐留時長參數之歌聲語料資訊

語料格式	語料內容
語料來源	本實驗室所錄製
語料種類/首數	歌聲/40 首
語料音檔格式	取樣率 16 KHz，解析度 16 bits
錄音人數/性別	1 位/女性
語料音節數	3,966 個

(二)外顯式狀態駐留時長之機率計算

本論文拿少量的歌聲信號給第 2 節訓練出的聲、韻母 HMM 模型，去作維特比解碼，這是為了求取每個聲、韻母 HMM 狀態的平均駐留時長與變異數值，作為 HMM 狀態駐留時長之參數。當把一個聲、韻母的一個 HMM 狀態的平均駐留時長值與變異數當作參數，帶入伽瑪機率分佈[6]，就能求得該狀態的外顯式時長機率，這樣就可改進原本 HMM 使用內顯式狀態移轉機率的缺點。

伽瑪機率分佈如公式(1)所示[7]：

$$f(x)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta} \quad , \quad x\geqq 0 \tag{1}$$

$$= \quad 0 \qquad\qquad\qquad , \quad \text{elsewhere}$$

其中，α 與 β 是決定此分佈曲線形狀的兩個參數，且兩個參數值只能為正數，此外伽瑪函式 $\Gamma(\alpha)$ 定義如公式(2)所示[7]：

$$\Gamma(\alpha)=\int_{0}^{\infty}x^{\alpha-1}e^{-x}dx \tag{2}$$

上式裡 x 表示在一個狀態上停留 x 個音框。關於 α 與 β 參數值的計算，可用公式(3)。

$$\alpha=\frac{E(X)^2}{Var(X)}, \qquad \beta=\frac{Var(X)}{E(X)} \ , \tag{3}$$

四、聲、韻母自動分段

在此一整首歌聲作聲、韻母自動分段處理的流程如圖 4 所示，第一步先輸入整首歌的歌詞作分析，即由中文字查出音節拼音;接著輸入整首歌聲的音檔，對整首歌聲作樂句切割，程式實作上先統計出各個片段之連續靜音的音框個數，再依各片段靜音之長度

作排序，然後依據樂句句數由長至短，選取靜音片段作為樂句分割點，如此就可輸出整首歌樂句分段的標記檔，但因為歌詞分析結果與歌聲樂句分段會有不一致的情況，所以我們再作人工更正，以獲得正確的樂句分段標記檔。

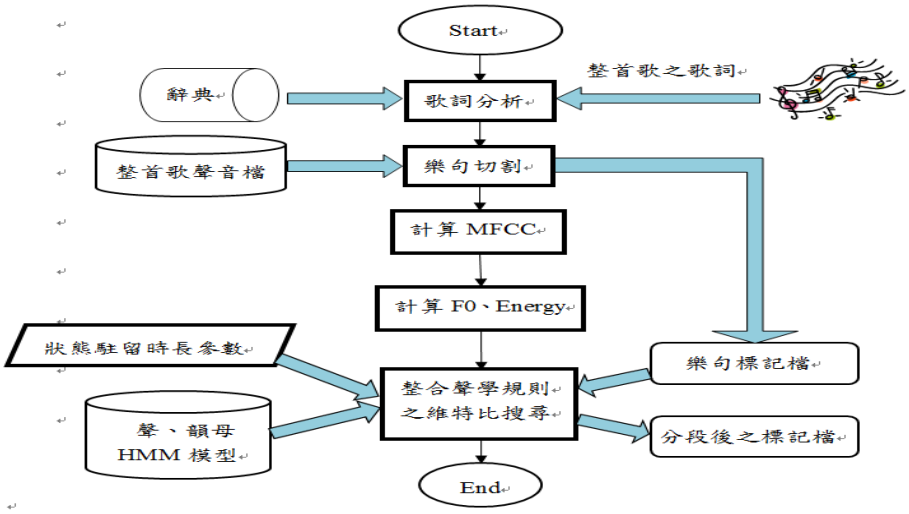

圖4　歌聲聲、韻母自動分段的處理流程

接著，對各樂句分段的時間範圍使用相同於圖四的 MFCC 程式去算出各音框的 MFCC 係數，然後計算各音框的能量值及使用吳俊欣先生發展的基頻估計程式模組[8]，算出各音框的基頻值;之後，我們可利用圖 3 所得到的狀態駐留時長參數，帶入伽瑪分佈的公式，去計算出外顯式狀態駐留時長機率，並且把聲、韻母的聲學規則整合至維特比解碼方法之中以搜尋出最佳的狀態路徑，最後就可以回溯(backtrack)找出分段點，並輸出分段後之標記檔。

(一) 維特比解碼

維特比解碼演算法可以分成四個部分，分別是初始、遞迴、終止以及回溯[9]。假設狀態序列為 $q=\{q_1,q_2,\ldots,q_t,\ldots,q_T\}$，$q_t$ 為示時刻 t 時所停留的狀態;觀測序列為 $o=\{o_1,o_2,\ldots,o_t,\ldots,o_T\}$，代表輸入的語音特徵向量序列; λ 表示某一個聲、韻母 HMM 模型的參數。則欲找出最佳路徑之機率，可先定義機率變數 $\delta_{t,i}(m)$，如公式(4)：

$$\delta_{t,i}(m)=\max_{q_1,q_2,q_3,\ldots,q_{t-1}} P[\,q_1,\,q_2,q_3,\ldots,q_{t-1},q_t=i,o_1,o_2,o_3,\ldots,o_t\mid\lambda\,] \qquad (4)$$

即令 $\delta_{t,i}(m)$ 表示時刻 t 時停留於狀態 i 上且使用第 m 個高斯混合的最佳路徑之累乘機率。若已知時刻 t 時各個狀態 i 上的 $\delta_{t,i}(m)$ 值，則在 t+1 時刻時停留在狀態 j 且使用第 m 個高斯混合的最佳路徑之累乘機率，可用公式(5)來作遞迴計算:

$$\delta_{t+1,j}(m)=\max\{\max_{0\le k\le M-1}[\delta_{t,j-1}(k)a_{j-1,j}\,],\,\delta_{t,j}(m)a_{jj}\,\}\cdot b_{j,m}(o_{t+1}) \qquad (5)$$

其中 $b_{j,m}(o_{t+1})$ 表示在狀態 j 上以高斯混合 m 觀測 o_{t+1} 的機率。

再著，定義一個指標變數 $\Psi_{t+1,j}(m)$，用來記錄時刻 t+1 時停留在狀態 j 上且使用第 m 個高斯混合時，應選取前一個時刻要從哪一個狀態上的哪一個高斯混合轉移過來。

(二) 外顯式時長機率之維特比解碼

在實作維特比解碼之路徑選取上，我們先把路徑行走分為斜角行走及水平行走兩種，如圖 5 所示。若是節點 Z 選擇斜角行走的情況，也就是圖 8 中虛線線段，表示在時刻 t

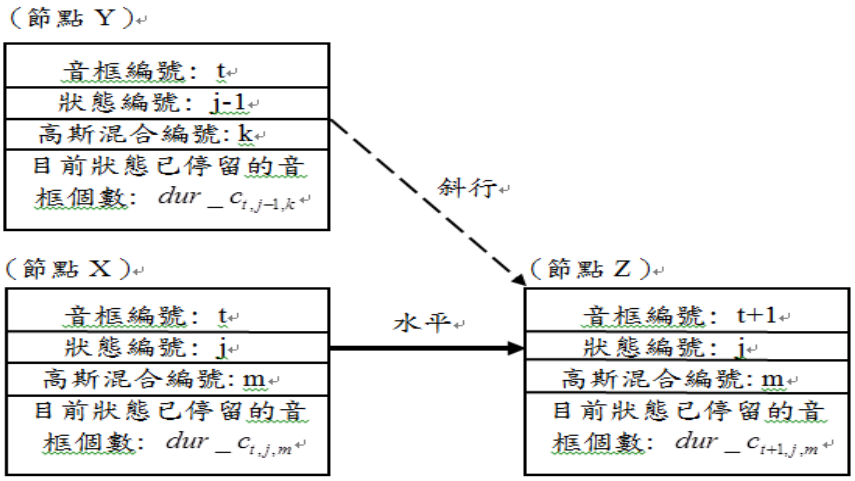

圖5　路徑行走方向及目前狀態已停留之音框個數

時停留在 j-1 狀態上，使用第 k 個高斯混合的累乘機率$\delta_{t,j-1}(k)$是 16 個高斯混合中的最大者且發生大於節點 X 之累乘機率$\delta_{t,j}(m)$之情況，因此在節點 Z 上就要把目前狀態已停留的音框個數重設為 1，並且要記錄是從 j-1 狀態的哪一個高斯混合轉移過來。若是節點 Z 選擇水平行走的情況，也就是圖 5 中實線線段，表示節點 X 的累乘機率$\delta_{t,j}(m)$大於節點 Y 上所有的 16 個高斯混合之累乘機率$\delta_{t,j-1}(k)$，$0 \leqq k \leqq 15$，之情況，因此就要依據節點 X 上已停留的音框個數加 1，然後存入成節點 Z 的已停留音框個數。前述概念之具體描述就如公式(6):

$$\mathbf{dur\_c}_{t+1,j,m} = \begin{cases} 1 & ，if\ 斜角行走\,(\delta_{t,j-1}(k) > \delta_{t,j}(m)) \\ dur\_c_{t,j,m} + 1 & ，if\ 水平行走 \end{cases} \quad (6)$$

其中$dur_c_{t+1,j,m}$表示在時刻 t+1 時，停留在狀態 j 上且使用第 m 個高斯混合的已停留於狀態 j 的音框個數。

依據公式(6)記錄各個節點已停留於某一狀態之音框個數$dur_c_{t,j,m}$，我們就可以修正維特比解碼演算法之遞迴步驟，把外顯式狀態時長機率含蓋進來，如公式(7)：

$$\delta_{t+1,j}(m) = \begin{cases} \delta_{t,j-1}(k) \cdot d_{j-1,j} \cdot b_{j,m}(o_{t+1}),\ if\ 斜角行走,\ 0 \le k < M \\ \delta_{t,j}(m) \cdot 1 \cdot b_{j,m}(o_{t+1}),\ if\ 水平行走 \end{cases} \quad (7)$$

其中 $d_{j-1,j} = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$ 代表外顯式狀態時長機率，變數 x 表示前一狀態 j-1 已停留的音框個數，即$dur_c_{t,j-1,k}$。

(三) 整合聲學規則之擴充式維特比解碼

在此，執行維特比解碼之前要先作靜音音框偵測的步驟，接著才把基頻及能量的相關規則整合到維特比解碼的演算法裡，並且把限制聲、韻母 HMM 狀態駐留音框總數的條件，整合至維特比解碼裡，如此整合了多個聲學規則之維特比解碼，就稱為擴充式的維特比解碼演算法。

靜音偵測

當偵測到一個音框的基頻值為 0 且能量值低於所設定的門檻值，就將該音框歸類為靜音音框，進一步當偵測到連續的靜音音框數量超過所設定的門檻值(如設為 15)，我們就會跳過這些靜音音框，不把這些靜音音框拿去作維特比解碼的處理，所以實作上維特

比解碼使用的是虛擬的時間軸，而維特比解碼完後，需要再對映回真實的時間軸。

音框記數變數

這裡我們把圖 5 中路徑**斜行**之方式細分為兩種情況，第一種情況是同一個聲、韻母 HMM 狀態之間的轉移，稱為音段內之狀態轉移;另一種情況是，聲母 HMM 的狀態轉移至韻母 HMM 的狀態，或是韻母 HMM 的狀態轉移至聲母 HMM 的狀態，這情況稱為音段切換之狀態移轉。

為了導入聲學規則，我們必需先在路徑行走的各節點上記錄相關會用到的資訊，例如聲、韻母 HMM 各個狀態所累計停留的音框個數及記錄基頻值非 0 (即具週期性)之累計的音框個數，記錄的方式如公式(8)及(9):

$$\boldsymbol{dur\_a_{t+1,j,m}} = \begin{cases} dur\_a_{t,j-1,k} + 1 & , if\ 斜行且同一音段 \\ 1 & , if\ 斜行且切換音段 \\ dur\_a_{t,j,m} + 1 & , if\ 水平行走 \end{cases} \quad (8)$$

其中$dur_a_{t+1,\,j,\,m}$表示在時刻 t+1 時，停留在第 j 個狀態上且使用第 m 個高斯混合之累計停留的音框個數，若路徑行走選擇的是斜行且前後狀態屬於同一音段、或是路徑行走選擇水平方向，我們就可把前一時刻的狀態上所累計停留的音框個數再加 1;另外若路徑行走選擇斜行且前後時刻的狀態跨越音段，則必需把累計的音框數重設為 1。

$$\boldsymbol{f0\_a_{t+1,j,m}} = \begin{cases} f0\_a_{t,j-1,k} + 1 & , if\ 斜行且同一音段且基頻值非0 \\ f0\_a_{t,j-1,k} & , if\ 斜行且同一音段且基頻值為0 \\ 0 & , if\ 斜行且切換音段 \\ f0\_a_{t,j,m} + 1 & , if\ 水平行走且基頻值非0 \\ f0\_a_{t,j,m} & , if\ 水平行走且基頻值為0 \end{cases} \quad (9)$$

在公式(20)裡，$f0_a_{t+1,\,j,\,m}$表示在時刻 t+1 時，停留在狀態 j 上且使用第 m 個高斯混合之具有週期性音框的累計音框個數。若路徑行走選擇的是斜行且前後狀態屬於同一音段，且本時刻音框具有週期性，或是路徑行走選擇水平方向且本時刻音框具有週期性，我們就可把前一時刻的狀態上所累計的週期性音框個數再加 1;若路徑行走選擇斜行且前後時刻的狀態跨越音段，則必需把累計的音框數重設為 0。

聲學規則

使用公式(6)、(8)及(9)所記載的資訊，我們據以設定聲學知識相關的檢查規則，再把這些規則整合至維特比解碼演算法的遞迴步驟中，實作上就是依據規則檢查的結果(yes 或 no)，去修改公式(7)的 $b_{j,m}(o_{t+1})$ 之值。

第一個違反聲學知識的檢查規則是，當前後時刻的狀態跨越音段時，檢查是否前一時刻的狀態屬於無聲(unvoiced)聲母之 HMM，並且累計的週期性音框數大於等於 FL (如設為 3)個音框，若條件成立就把$b_{j,m}(o_{t+1})$乘以一個很小的懲罰性機率值 P_{ini}，並且實作上我們對於累計的週期性音框數越多的情況，會乘以更小的懲罰性機率值，因為依聲學知識可知無聲聲母 HMM 上停留的語音音框應為無週期性的，不過考慮到基頻偵測仍可能發生偵測錯誤，所以才把規則設為當週期性音框總數超過 FL 個音框之條件。

第二個違反聲學知識的檢查規則是，當前後時刻發生音段切換，檢查是否前一時刻的狀態屬於韻母之HMM，並且前一時刻的狀態上累計的週期性音框數少於此韻母HMM累計停留的音框總數的 PT (如設為 85%)比例之情況，我們就會把$b_{j,m}(o_{t+1})$乘以一個很小之機率值P_{fnl}，因為依聲學知識可知韻母 HMM 上停留的音框應為週期性音框，不過考慮到基頻偵測可能發生錯誤，所以我們的規則就只要求累計的週期性音框數要佔此韻母累計音框總數的 PT 比例以上。

第三個違反聲學知識的檢查規則是針對停留於靜音狀態上的音框，依聲學知識可知靜音音框的能量應要很小，如果一個停留於靜音狀態的音框，其能量值大於能量門檻值，就屬於不合理的情況，因此我們就把$b_{j,m}(o_{t+1})$乘以一個很小的懲罰性機率值P_{sp}，以避免該路徑被行走。

五、測試實驗

在此，我們設音框長度為 400 個樣本點(25 ms)，音框位移為 80 個樣本點(5 ms)，並且從歌聲語料庫內的 40 首歌曲中選取 20 首來量測所製作的聲、韻母自動分段系統的準確率，在這 20 首歌曲中有 9 首屬於節奏偏快的歌曲;另外的 11 首屬於節奏偏慢的歌曲。

我們以準確率來評量聲、韻母自動分段方法之效能，準確率之計算如公式(10):

$$P=(\frac{N_{hit}}{N_{total}})\times 100\% \quad , \tag{10}$$

其中 N_{hit} 表示程式執行結果與標準答案比較後，屬於正確的聲、韻母左右邊界數量; N_{total} 是聲、韻母左右邊界的總數量。此外，在計算聲、韻母自動分段方法的效能時，我們使用了一個容忍範圍的設定，只要在正確的聲、韻母左右邊界點的容忍度範圍內，程式有偵測到對應的聲、韻母分段之左右邊界點，就代表是一個正確的邊界偵測結果，在此我們實驗了 5 種容忍範圍，分別是 10 ms、20 ms、30 ms、40 ms 及 50 ms。

(一)基本維特比解碼 與 外顯式狀態時長機率之維特比解碼

在執行基本維特比解碼方法(BVD)的程式後，把偵測出的聲、韻母邊界位置輸出成標記檔，再用以和人工標記的標準答案作比較，以計算出準確率。結果得到的自動分段準確率數值如表 3 所示，由表 3 可知，BVD 法應用在歌聲信號的聲、韻母自動分段結果並不理想，在 10 ms 的容忍範圍內，只有 31.73%的準確率。

表 3　基本維特比解碼之自動分段準確率

容忍範圍	≦10 ms	≦20 ms	≦30 ms	≦40 ms	≦50 ms
準確率	31.73%	53.39%	65.88%	72.27%	75.87%

當修正基本維特比解碼的遞迴步驟，即改用外顯式狀態時長機率以取代原本使用的內顯式狀態移轉機率，如公式(7)，然後執行外顯式狀態時長機率之維特比解碼方法(EVD)，結果量測得到的自動分段準確率如表 4 所示，比較表 4 與表 3 的數值，可發現在不同的容忍範圍內的準確率皆有提高，而在 10 ms 的容忍範圍內，分段的準確率可提升至 39.28%，但這與我們期望的目標仍有很大的差距。

表 4　使用外顯式狀態時長機率之自動分段準確率

容忍範圍	≦10 ms	≦20 ms	≦30 ms	≦40 ms	≦50 ms
準確率	39.28%	60.26%	72.27%	77.97%	81.0%

(二)整合聲學規則之擴充式維特比解碼

當使用第 1 個檢查規則時，我們內定(default)的音框數門檻是 3，即檢查無聲聲母 HMM 累計的週期性音框數是否大於 3 個音框；而當使用第 2 個檢查規則時，我們內定的比例門檻值是 85%，即檢查韻母 HMM 累計的週期性音框數是否少於該韻母 HMM 累計停留的音框總數的 85%之情況。前述的門檻值是否恰當? 我們在 10ms 的容忍範圍內進行不同門檻值設定之準確率量測實驗。首先對韻母 HMM 週期性音框數比例值 80%、

85%及 90% 進行實驗，接著，再對無聲聲母 HMM 的週期性音框數門檻：≧2、≧3、≧4 及≧5，進行分段實驗，量測得到的準確率結果顯示，當韻母 HMM 週期性音框數之比例門檻設為 85%，並且把無聲聲母 HMM 的週期性音框數門檻設為 3 時，才能夠獲得最高的分段準確率數值 66.86%，所以內定的門檻值設定，的確是最好的。

關於連續的靜音音框數之門檻值設定，我們實驗了≧10、≧15、≧20 及≧25 等幾個門檻值，結果量測出的聲、韻母分段準確率如表 5 所示，由表 5 可知，將連續靜音音框數門檻設為≧20 是最好的選擇。

表 5　連續靜音音框數門檻值設定之分段準確率

連續靜音音框數門檻	≧10	≧15	≧20	≧25
準確率	63.48%	64.97%	66.86%	65.53%

六、結論

我們製作的歌聲聲、韻母自動分段程式，分成三種版本，結果顯示使用內顯式狀態移轉機率之基本的維特比解碼方法，在 10 ms 之容忍度內只能達到 31.73%之聲、韻母分段正確率;當改成使用外顯式狀態時長機率之修正的維特比解碼方法，則可讓正確率稍微提升至 39.28%；至於整合聲學規則的擴充式維特比解碼方法，則能夠將準確率更進一步提升至 66.86%。然後，當再加入一種簡單的後處理步驟，即依據靜音音框的合理位置去移動聲、韻母的時間邊界，則可再小幅提高分段準確率至 68.45%。

我們提出之整合聲學規則之維特比解碼方法，在 10 ms 容忍範圍內，比林政源的方法(HMM+DTW)[10]好很多。不過，當林政源的方法再加入一種基於 SVM 之分數預測模型[10]來作後處理，則他的方法能夠讓準確率提升到約 72%。然而，我們的分段方法相對地較為單純(只基於 HMM 與聲學規則)，實作上花費的氣力也較少(不牽涉到各種技術)，所以，我們的方法相對地較為單純(只基於 HMM 與聲學規則)，實作上花費的氣力也較少(不牽涉到各種技術)，而能獲得 68.45%的分段準確率，應可說是不錯了。

未來我們可考慮在訓練聲、韻母 HMM 模型方面，採用更能夠表現頻譜特性之特徵參數組合，來加強 HMM 模型對於不同的聲、韻母在頻譜特徵上的分辨能力；另外，對於有聲聲母銜接韻母及韻母銜接韻母情況的邊界判斷，應可再經由適合的後處理方法，而讓準確率獲得更進一步的提升。至於對含有樂器演奏之歌唱音檔作聲、韻母自動分段，那是非常具有挑戰性的問題，需要更進一步的研究。

參考文獻

[1] S. Young, “The HTK Hidden Markov Model Toolkit : Design and Philosophy”, Tech Report TR.153, Department of Engineering, Cambridge University (UK), 1993.

[2] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valchev, P. Woodland, The HTK Book (for HTK version 3.2.1), Cambridge University Engineering Department, 2002.

[3] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[4] 吳昌益,使用頻譜演進模型之國語語音合成研究,國立台灣科技大學資訊工程研究所碩士論文,2007。

[5] ACLCLP, Mandarin microphone speech corpus - TCC300, http://www.aclclp.org.tw/use_mat.php#tcc300edu .

[6] R.L. Scheaffer, Introduction to Probability and Its Applications, PWS Publishing, 1995.

[7] 林秉正,使用適應性區間模型於語者說話速度之調整,國立成功大學資訊工程研究所碩士論文,2002。

[8] 吳俊欣,MFCC特徵空間座標系統對映之語者調適方法,國立台灣科技大學資訊工程研究所碩士論文，2003。

[9] 王小川，語音訊號處理(修訂二版)，全華圖書公司，2009。

[10] 林政源,應用於中文語音與歌聲合成之自動切音研究,國立清華大學資訊工程研究所博士論文,2007。