文本不特定之自動音素分段演算法

Text-Independent Automatic Phone Segmentation Algorithm

陸勁逢　Ching-Feng Lu
國立清華大學電機工程學系
Department of Electrical Engineering
National Tsing Hua University
g9761589@oz.nthu.edu.tw

王小川　Hsiao-ChuanWang
國立清華大學電機工程學系
Department of Electrical Engineering
National Tsing Hua University
hcwang@ee.nthu.edu.tw

摘要

本論文提出一個文本不特定循序式偵測音素邊界的演算法，在沒有提供任何已知資訊情況下，建構一個自動音素分段系統。其方法是沿著時間一次只尋找一個候選音素邊界點，找到之後即作確認，經確認後才算是偵測到的音素邊界點而進行音素分段。尋找候選音素邊界點時，採用改變音框長度之離散小波轉換，將小波係數換算成一組頻帶能量，從能量排序之頻帶尋找可能的音素存在範圍，接著以貝式資訊修正準則與正規化頻譜變異函式針對候選音素邊界點進行確認。實驗的語音資料是 TIMIT 語料庫，實驗結果以 F-值與 R-值表示其音素分段之正確性，在 20ms 容忍度下，640 個實驗語句得到的平均 F-值爲 72%，R-值爲 75%。

Abstract

This paper proposes a text-independent sequential phone boundary detection algorithm. Without any previous knowledge, an automatic phone segmentation system is constructed. The method is to sequentially search for a candidate phone boundary and follow by a verification process. The phone segmentation is accomplished when the phone boundaries are verified. The discrete wavelet transform is applied to variable-length frames. The wavelet parameters are converted to a set of band energies. The energy-sorted bands are then used to search for the candidate phone boundaries. The Bayesian information criterion corrected (BICC) and normalized spectral variation function (SVF) are applied for verifying the phone boundaries. To evaluate this proposed algorithm, the experiment was conducted on TIMIT corpus. The performance of phone segmentation was measured in F-value anf R-value. In the condition of 20-ms tolerance, the average F-value of 640 test utterances is 72% and R-value is 75%. .

關鍵詞：音素分段、音素邊界偵測、離散小波轉換、頻譜變異函式、貝式資訊修正準則

Keywords: phone segmentation, phone boundary detection, discrete wavelet transform, spectral variation function (SVF), Bayesian information criterion corrected (BICC)

一、緒論

在語音處理的研究中，決定語音正確的音素邊界點是一項很重要的工作。知道正確的音素邊界位置不僅可以提升語音辨識率，也可以提高語音合成的品質。但是有人工手動標記音素位置的語料庫並不多，這是因爲人工標音不僅耗費大量的人力與時間，而且常會因爲主觀上認定標記位置的不同而缺乏一致性，如何設計一個自動標記語料庫的系統[1]，是語音信號處理中的研究重點。

然而，在沒有提供任何已知資訊下，標記音素邊界點的位置是非常困難的。在過去一些自動音素分段(phone segmentation)與偵測的研究中[1][2]，最好的實驗結果是可以在正負 20 ms 容忍度之內，有 90%的含蓋率(inclusion rate)，但是必須透過提供一些已知的資訊幫助找出正確的標記點，這樣的作法在不能獲得語文資訊的情況下，就會不實用。一般來說，若想要在沒有已知資訊的條件下完成自動標記的工作，必須先將語音訊號轉換成一組特徵參數(feature)，接著去量測這些參數間的變化量，變化量大的位置即暗示了音素邊界點之所在。

音素分段是語音信號處理中非常重要的研究課題，所謂的分段，是將一串序列(sequence)拆成一些有意義的單元(unit)。若是針對語音做分段(segmentation)，這些分段邊界可以是音素邊界(phone boundaries)、詞邊界(word boundaries)、或句邊界(sentence boundaries)等。其中音素分段通常是語音辨認或語音合成技術中最重要的前處理。音素分段可分爲基於模型(model-based)及基於量測(metric-based)兩種方法，或是這兩種方法的結合。

在基於模型的方法中，主要是以概似法則訓練的隱藏式馬可夫模型(maximum likelihood-trained hidden Markov model, ML-trained HMM)做自動語音分段，最出色的語音分段效能是由 Hosom[3]所發表，其演算法結合了類神經網路(artificial neural network, ANN)與 HMM，以 TIMIT 語料庫進行實驗，在正負 20 ms 容忍度內可以高達 92.57%的含蓋率。雖然基於隱藏是馬可夫模型的方法有很高的正確率，但此法需要很多的訓練語料(training data)來建立一個準確的統計模型，對於某些語料數量不足的語料庫而言，往往在應用上會有許多的限制。

在基於量測的方法中，並不用依賴任何訓練的語料，主要的方法是使用頻譜失真量測(spectral distortion measure)，或是頻譜變異函式(spectral variation function, SVF)[4]作爲分段依據，以 F-值表示其分段之正確性，得到的結果是 F-值爲 67.7%。以差量倒頻譜函式(delta cepstral function, DCF)進行分段實驗，其 F-值爲 70.1%，最好的實驗結果是使用 DIST-ICOMPR 法，其 F-值高達 74.7%。

本篇論文主要的目的是要結合一些先前的方法，如 Almpanidis 等[4]所提出的 DISTBIC 以及 Brugnara[5]和 Mitchell[6]所使用的頻譜變異函式等，提出新的演算法，發展出文本不特定的音素分段系統。所提出的音素分段系統在 TIMIT 語料庫上的實驗結果顯示，在 20ms 容忍度下，多分率約爲-6%，即傾向於少分，但擊中率(hit rate)與精確率(precision rate)仍分別有 70%與 75%。文獻上其他音素分段方法，傾向於多分以提升擊中率，本文所提出的方法是希望在少分的情形下仍維持一個不錯的擊中率，以避免多

分時需要作錯誤邊界點之偵測。

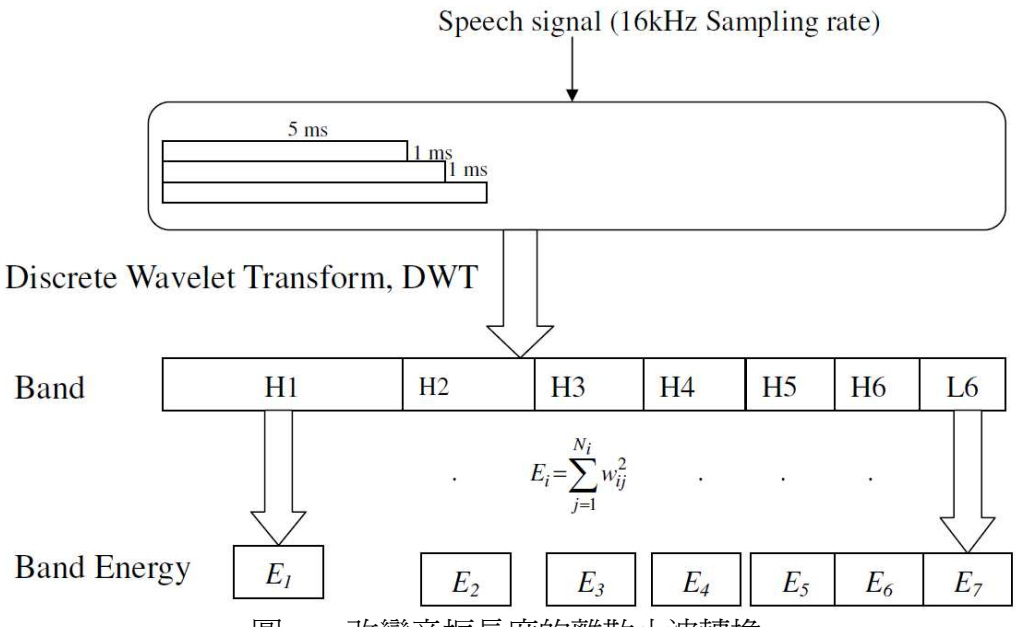
二、語音特徵參數

本文使用的語音特徵參數爲梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)與離散小波轉換係數[7]，梅爾倒頻譜係數是在固定音框長度爲 320 個取樣點的情況下計算，音框移動爲 32 個取樣點。離散小波轉換則有兩種計算方式，一是在固定長度音框中計算，其係數標示成 WLP_FL，一是在改變長度音框中計算，其係數標示成 WLP_VL。假設對一個音框長度爲 256 個取樣點的語音波形，作六層解析度的離散小波轉換，在 16 kHz 取樣頻率下，小波轉換後的七個頻帶如表一所示，

表一、離散小波轉換後的七個頻帶

頻帶	H1	H2	H3	H4	H5	H6	L6
頻率範圍 (Hz)	4000~ 8000	2000~ 4000	1000~ 2000	500~ 1000	250~ 500	125~ 250	0~ 125
小波係數個數	128	64	32	16	8	4	4

固定音框長度的離散小波轉換，其 256 個係數(WLP_FL)組成一個參數向量，將用於音素邊界點的確認。改變音框長度的離散小波轉換，則是用來尋找候選音素邊界點，其係數(WLP_VL)的計算如圖一所示。



圖一、改變音框長度的離散小波轉換

在圖一中的離散小波轉換，使用不同的音框長度，初始音框長度設定爲 80 個取樣點，對應的時間長度爲 5 ms，第二個音框的起始點與第一個音框相同，但音框長度增加 16 個取樣點，換算成時間長度就是增加了 1 ms。對每一個音框進行離散小波轉換的演算，轉換之後分成七個頻帶，將對應的頻帶係數取能量和，

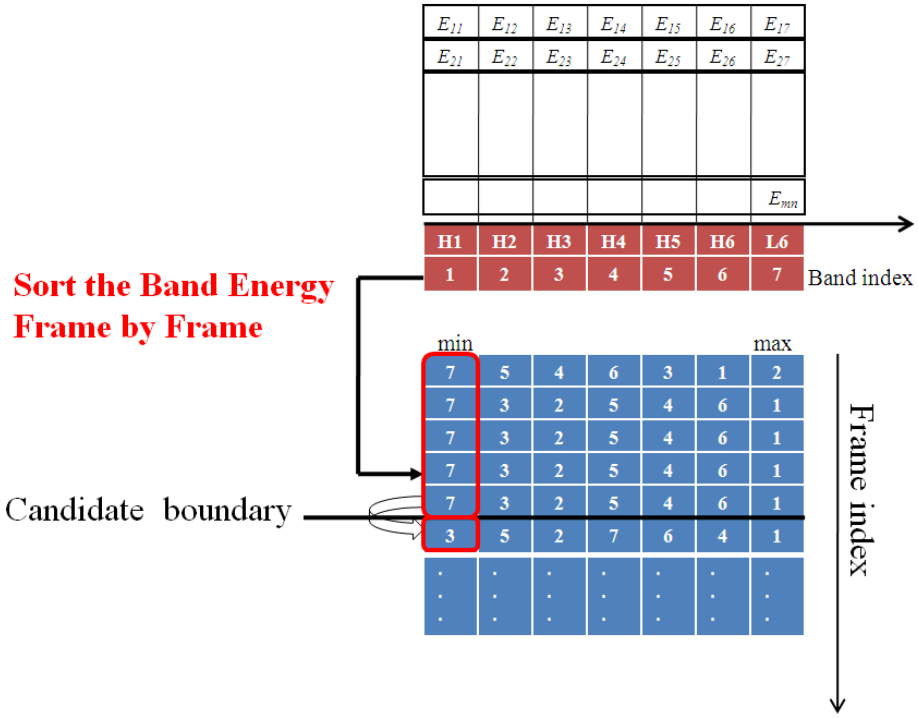
$$E_i = \sum_{j=1}^{N_i} w_{ij}^2 \qquad (1)$$

E_i 代表第 i 個頻帶的係數能量和，N_i 爲第 i 個頻帶的係數個數，w_{ij} 爲第 i 個頻帶的第 j 個小波係數。如此就將一個音框的離散小波轉換結果，表示成一個維度爲七的向量。對於不同長度的音框，都得到維度爲七的向量，因此可以組成一個矩陣。若我們連續取了 100 個改變長度的音框，就會得到一個 100x7 的矩陣，其中矩陣的列數代表音框的數目，矩陣的行數表示離散小波轉換後的頻帶數目。

三、候選音素邊界點之偵測

尋找候選音素邊界點的演算，就是用改變音框長度的離散小波轉換，轉換成七個頻帶，將對應的頻帶係數取能量和。對維度爲七的向量作觀察，針對各頻帶之小波係數能量和變化，判斷是否可能是音素邊界點。當我們觀察一個音素，此音素持續存在的時間內，各個頻帶能量大小之順序是穩定的。舉例來說，若是有一個音素存在的時間長度爲 10 ms，且此音素的能量集中在 250-500Hz 之間，則當我們從一個起點開始觀察，會有一段長度爲 10 個音框的區間，其頻帶 H5 的能量排序在最右邊，也就是能量最大。當然，也可以觀察頻帶能量最小者。本文提出的方法，是觀察能量最小者，在其穩定區間的開始與結束，會看到能量最小頻帶發生改變，改變處就是可能的音素邊界點。

圖二描述這個尋找候選音素邊界點的例子，它檢查能量最小的頻帶，發現能量最小的頻帶原來一直是 7，後來變成 3，在發生改變的音框處，就可能是音素邊界點。



圖二、尋找候選音素邊界點

四、音素邊界點之確認

尋找到一個候選的音素邊界點之後，立即進行確認，確認演算是依據貝式資訊修正準則(Bayesian information criterion corrected, BICC)與頻譜變異函式(spectral variation functions, SVF)。

(一) 貝式資訊修正準則

貝式資訊準則(Bayesian information criterion, BIC)常用來偵測聲音訊號的改變，如說話人的切換、通道改變、或環境改變 [8][9]，通常需要相當長的一段聲音訊號，才能偵測其發生改變之所在。Almpanidis 等在討論音素邊界偵測時[4]，建議使用貝式資訊修正準則(BICC)，它可以用在短的聲音訊號，本文就利用貝式資訊修正準則(BICC)，配合頻譜變異函式(SVF)進行音素邊界點之確認。

如果給一序列的語音特徵向量，$\{\mathbf{S}_i, i=1,2,...,N\}$，它可能是來自連接兩個音段的一串語音特徵向量，也可能是只有一個音段的一串語音特徵向量。一個音段的語音特徵向量分佈可以對應一個機率模型，因此語音特徵向量，$\{\mathbf{S}_i, i=1,2,...,N\}$，可能對應兩個機率模型，也可能對應一個機率模型。如何評估一串語音特徵向量所對應到的機率模型，就是所謂的模型選擇問題。貝式資訊修正準則(BICC)是一個選擇模型的準則，它計算這組向量對應到一個模型 M_j 的對數最大概似值(log-maximum likelihood)，如(2)式示。

$$BICC(M_j) = -2\ln\{l[\mathbf{S}_1,\mathbf{S}_2,...,\mathbf{S}_N \mid M_j]\} + \frac{kN\ln N}{N-k-1}, \quad k = Q + \frac{Q(Q+1)}{2} \tag{2}$$

$l[\mathbf{S}_1,\mathbf{S}_2,...,\mathbf{S}_N]$ 爲特徵向量序列對應到模型 M_j 的最大概似值(Maximum Likelihood)，Q 爲特徵向量的維度， N 爲整段信號之特徵向量數量。

如果這一個序列的語音特徵向量 $\{\mathbf{S}_i, i=1,2,...,N\}$ ，是由兩個模型所產生，其前段 $\{\mathbf{S}_i, i=1,2,...,m\}$ 對應到模型 M_a ，其後段 $\{\mathbf{S}_i, i=m+1,m+2,...,N\}$ 對應到模型 M_b 。如果語音特徵向量 $\{\mathbf{S}_i, i=1,2,...,N\}$ 是由一個模型所產生，整段對應到模型 M_{ab} 。要判斷此語音特徵向量 $\{\mathbf{S}_i, i=1,2,...,N\}$ 是由兩個模型所產生，或是由一個模型所產生，我們計算其間的差異，如(3)式示。

$$\Delta BICC = BICC(M_a) + BICC(M_b) - BICC(M_{ab}) \tag{3}$$

若是 $\Delta BICC > 0$ ，則暗示是由一個模型所產生，反之，若是 $\Delta BICC < 0$ ，則暗示是由兩個模型所產生。在實際應用中，會設定幾個門檻值，如高門檻、低門檻等，並訂一個判斷程序，來決定是否是由兩個模型所產生，如此就可以判斷是否真的是一個音素邊界點存在。

(二) 頻譜變異函式

頻譜變異函式(Spectral Variation Functions, SVF)是由 Brugnara 提出[5]，當時的 SVF 主要是藉由計算兩個正規化倒頻譜向量(Normalized Cepstral Vector)之角度來判斷兩個特徵向量之差異性，但是在本文中，使用的特徵向量爲離散小波轉換係數。SVF 之數學定義如(4)式所示：

$$SVF(i)=\frac{\langle \mathbf{S}_{i-1},\mathbf{S}_{i+1}\rangle}{\|\mathbf{S}_{i-1}\|\,\|\mathbf{S}_{i+1}\|} \tag{4}$$

其中$\|\bullet\|$表示向量範數(vector norm)，$\langle \bullet,\bullet\rangle$表示兩個向量的內積。若是針對一連串的音框計算其 SVF，這些音框的 SVF 值範圍就會落在-1到 1 之間，兩個音框相似度越高，則兩個音框在向量空間中的夾角越小 SVF 的值就越接近 1。反之，若是兩個音框相似度越低，兩個音框在向量空間中的夾角越大 SVF 的值就越接近-1。如果作一個位移與正規化演算，改爲如(5)式所示：

$$C(i)=\frac{1}{2}\left(1-\frac{SVF(i)}{\max_i |SVF(i)|}\right) \tag{5}$$

如此一來$C(i)$的範圍就會落在 0 到 1 之間，當兩個音框相似度越高$C(i)$就越趨近於 0，當兩個音框相似度越低$C(i)$就越趨近於 1。在本文中，SVF 會與$\Delta BICC$配合使用，以判斷是否真的是一個音素邊界點。

五、音素分段演算法

本文提出的音素分段演算法分成三個步驟，簡述如下。

(一) 候選音素邊界點的偵測

從一個「起始點」開始，尋找下一個可能的音素邊界點，使用的方法就是第三節所描述的演算法。以「起始點」開始，作改變長度音框的離散小波轉換，計算改變音框長度的小波係數(WLP_VL)，轉換成七個頻帶能量，依據頻帶能量排序，觀察能量最小者，看其頻帶改變，找到候選音素邊界點的所在。

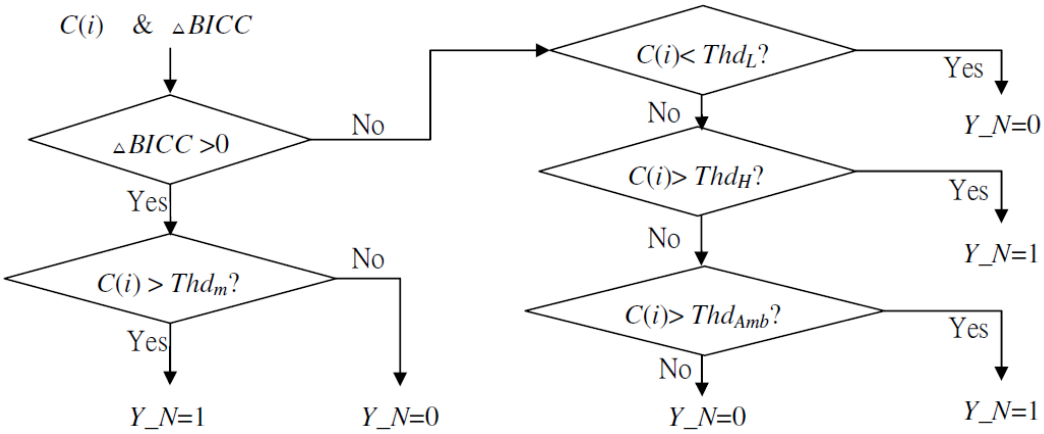
(二) 音素邊界點的確認

找到候選音素邊界點之後，即進行音素邊界點的確認。以候選音素邊界點爲中心，向前與向後各取 500 個取樣點，對這段語音訊號計算各音框的 MFCC 與固定音框長度的小波係數。計算 MFCC 時音框長度爲 320 個取樣點，音框移動爲 32 個取樣點，得到一序列音框的 MFCC 向量，用來計算$\Delta BICC$，作音素邊界點的判斷。另外是做固定音框長度的離散小波轉換，音框長度爲 256 個取樣點，音框移動爲 128 個取樣點，對這一序列音框的小波係數(WLP_FL)計算其頻譜變異函式(SVF)，配合$\Delta BICC$進行音素邊界點的確認。

確認程序說明如下；先檢查$\Delta BICC$是否大於 0，若$\Delta BICC>0$，就表示此候選音素邊界點已經有很大的機會是多分的邊界點，於是利用$C(i)$的值再進一步的確認。此時我們給定一個門檻值Thd_m，若$C(i)<Thd_m$，此候選音素邊界點就是多分的點，令$Y_N=0$。反之，$C(i)>Thd_m$，就確定此候選音素邊界點爲一正確的音素邊界點，令$Y_N=1$。

若$\Delta BICC<0$，就表示此候選音素邊界點有很大的機會成爲一個正確的邊界點，於是再檢查$C(i)$的值，此時給定的門檻值是一個極小的數值Thd_L，只要$C(i)<Thd_L$，就表示此候選音素邊界點是靜音部分多分的邊界點，若是$C(i)>Thd_L$，我們就再給定一個比Thd_m大的門檻值Thd_H，若是$C(i)>Thd_H$，就表示$C(i)$與$\Delta BICC$這兩個條件都暗

示候選音素邊界點是正確的音素邊界點。最後，若是$\Delta BICC < 0$，但是$C(i) < Thd_H$，就表示$\Delta BICC$暗示此邊界是正確的，但$C(i)$卻還不足以完全的暗示此邊界是正確的。爲了解決這樣模稜兩可的情形，我們增設一個門檻值Thd_{Amb}，再比較$C(i)$與Thd_{Amb}，以決定是否真的是音素邊界點。以下列出我們選擇的門檻值：$Thd_L = 0.001$、$Thd_m = 0.5$、$Thd_{Amb} = 0.35$、與$Thd_H = 0.65$，這些門檻值的設定是根據實驗調整而得到。確認程序的流程圖如圖三所示:；

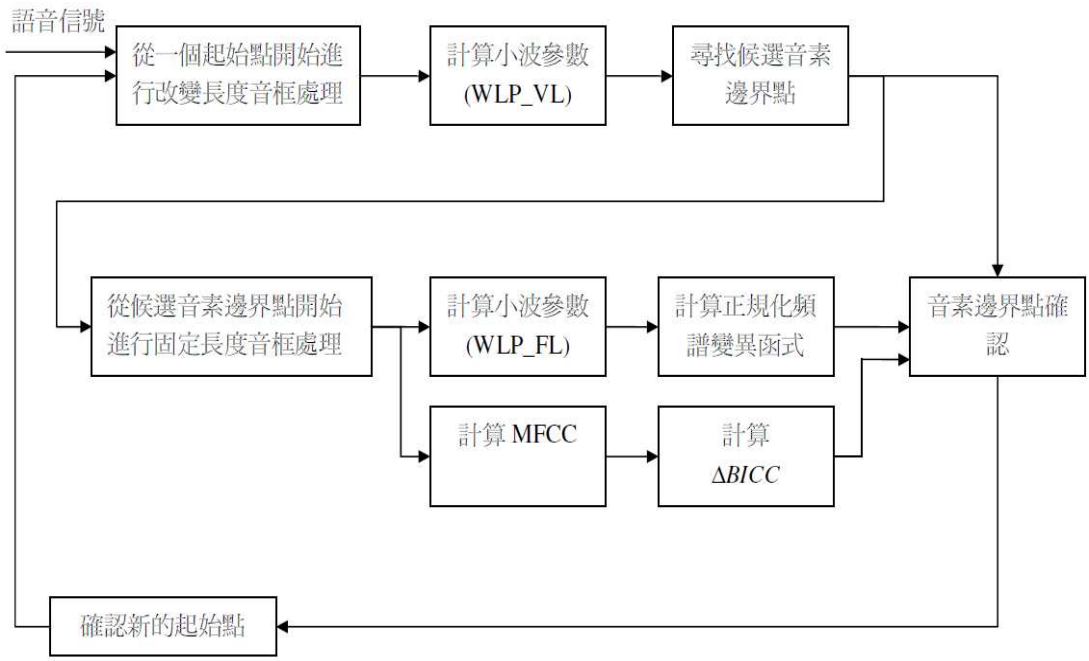


圖三、確認音素邊界點之流程圖

(三) 確定新的「起始點」

如果能確認音素邊界點之所在，就以此音素邊界點爲新的「起始點」，回到步驟(一)，繼續尋找下一個可能的音素邊界點。如果不能確認有音素邊界點存在，就表示這個候選音素邊界點可能是誤判，因此將「起始點」向後移 1 ms，回到步驟(一)重新尋找候選音素邊界點。

以上程序持續進行到全段語音訊號結束，即完成音素邊界點的標記。爲了確保所標記的音素邊界點是合理的，會對相鄰的音素邊界點作間距的判斷，若是小於一個設定值，就很可能是多標出來的，應該刪去。圖四歸納了上述的音素分段演算法，以流程圖表示。



圖四、音素分段演算之流程圖

六、實驗

(一)實驗語料

本研究採用 TIMIT 訓練集(Train Set)之語料進行音素分段之測試評估，實驗語料數共有 640 句，由 40 位男性語者以及 40 位女性語者提供。每位語者扣除“SA”之語句後，有 8 句話用來進行分段實驗，實驗語料數量統計表如表二所示：

表二、實驗語料數量統計表

性別	語者數	句數/語者	語料總數/性別
男	40	8	320
女	40	8	320
總計	80	8	640

(二)音素邊界點偵測結果之評量方式

在評估自動音素分段系統的好壞時，本文使用 F-值、R-值、以及音素邊界分類擊中率三種評分方式。這三種評分方式詳述於下：

(1) F-值

F-值用來評估偵測音素邊界點的正確性，其定義爲；

$$F = \frac{2 \times PRC \times RCL}{PRC + RCL} \times 100 \qquad (6)$$

假警報率： $FAR = \frac{N_{fa}}{N_{ref} + N_{fa}} \times 100$　　精確率： $PRC = \frac{N_{hit}}{N_b} \times 100$

召回率： $RCL = \frac{N_{hit}}{N_{ref}} \times 100$

(2) R-值

R-值是另一種評量分段演算法好壞的估測值，它是由 Rasanen 等人於 2009 年提出[10]，其定義如下；

$$R = 1 - \frac{r_1 + abs(r_2)}{200} \tag{7}$$

擊中率： $HR = \frac{N_{hit}}{N_{ref}} \times 100$　　多分率： $OS = (\frac{N_b}{N_{ref}} - 1) \times 100$

$r_1 = \sqrt{(100 - HR)^2 + (OS)^2}$　　$r_2 = (-OS + HR - 100)/\sqrt{2}$

以上的各個數值說明如下；

N_{ref} ：人工標音的邊界點總數　　N_b ：系統找到的邊界點總數

N_{hit} ：人工標音的邊界點被找到的總數

N_{fa} ：系統誤認是邊界點的總數

所謂「找到」是指偵測到的位置與人工標音的位置在指定的容忍度範圍內。

(3) 音素邊界分類擊中率

我們將系統找到的每一個音素邊界點，經由分類之後統計其分類擊中率。音素之種類分爲六大類，即塞音(stop)、擦音(fricative)、鼻音(nasal)、母音(vowel)、半母音與滑音(semi-vowel and glide)、以及靜音(silence)，其中我們將擦音與塞擦音視爲一個類別，其餘種類定義的音素類別皆與 TIMIT 定義之音素類別相同。一個音素的邊界點常常是前一個音素的終點與後一個音素的起點，依前述分成六類，則音素的邊界點可以有 35 類，其中靜音-靜音連接不算。類別標示以 k 表示，分類擊中率之計算如下式；

$$HR_k = \frac{N_{hit,k}}{N_{ref,k}} \times 100$$

$N_{ref,k}$ ：人工標音的邊界點中屬於 k 類之總數

$N_{hit,k}$ ：人工標音的邊界點中屬於 k 類者在容忍度範圍內被找到的總數

(三) 實驗結果與討論

以下是實驗結果之數據；

(1) F-值

表三展示三種容忍度(20 ms、15 ms 及 10 ms)下實驗得到的 F-值。

表三、以 F-值爲評量之實驗結果

	20 ms				15 ms				10 ms			
	F	PRC	RCL	FAR	F	PRC	RCL	FAR	F	PRC	RCL	FAR
女	72.6	76.9	69.4	19.5	64.6	68.3	61.8	23.8	51.2	54.2	49.0	29.6
男	72.2	74.6	70.4	21.5	62.9	65.1	61.4	26.1	48.3	50.0	47.1	32.4
全部	72.4	75.7	70.0	20.5	63.7	66.7	61.6	25.0	49.7	52.1	48.0	31.0

表三顯示隨容忍度由 10 ms、15 ms、增加到 20 ms，F-值(%)、召回率(RCL)、與精確率(PRC)也都隨之提升，而假警報率(FAR)則隨之下降。由於傳統文本不特定音素分段之研究採用 20ms 之容忍度進行實驗評估，因此我們將所有語料 F-估測值在 20 ms 容忍度下的分佈情形展示於表四。:

表四、所有語料在 20ms 容忍度下 F-值之分佈情形

F-值(%)之範圍	句數	句數/總句數	平均(%)
90 以上	4	0.63	91.85
85-90	17	2.66	86.72
80-85	63	9.84	81.88
75-80	154	24.06	77.23
70-75	184	28.75	72.44
65-70	129	20.16	67.76
60-65	58	9.06	63.02
55-60	19	2.97	57.39
50-55	6	0.94	53.43
50 以下	2	0.31	47.21
總計	640	100	72.40

(2) R-值

接著對三種容忍度(20 ms、15 ms 及 10 ms)計算其 R-值，數據如表五所示。

表五、以 R-值爲評量之實驗結果

	20 ms			15 ms			10 ms		
	R	HR	OS	R	HR	OS	R	HR	OS
女	75.1	69.4	-8.6	68.9	61.8	-8.6	58.3	49.0	-8.6
男	74.9	70.4	-4.4	67.6	61.4	-4.4	55.6	47.1	-4.4
全部	75.1	70.0	-6.5	68.2	61.6	-6.5	57.0	48.0	-6.5

如表五所示，多分率(OS)在不同的容忍度之下沒有改變的，而且是負值，這表示本系統處於少分之現象，多分率(OS)距離正值還有一段距離。隨容忍度由 10 ms、15 ms、增加到 20 ms，R-值與擊中率(HR)也隨之提升，而多分率(OS)仍維持－6.5% 。

此外，比較男性語料與女性語料在 20 ms 容忍度下的整體數據，男、女性語料的假警報率(FAR) 分別爲 21.58%與 19.52%、多分率(OS)分別爲－4.4% 與－8.6%、擊中率(HR)分別爲 70.4%與 69.4%，而精確率(PRC)分別爲 74.6%與 76.9%。男性語料之擊中率較女

性與者高，但精確率卻較女性與者低，而 F-值皆維持在 72%左右，這顯示採用不同性別之語料進行實驗對系統的偵測效果影響極小。

(3) 音素邊界分類擊中率

在容忍範圍 20ms 下，640 句實驗語料在容忍度 20ms 下，其分類擊中率，由於共計有 35 種可能的音素邊界，以下僅展示出現頻率最多 12 種情形，其在人工標音的邊界點總數中所佔百分比爲 2.65% 以上，如表六所示：

表六、最常出現之 12 種音素邊界在容忍度 20ms 下之分類擊中率

音素邊界種類	$N_{hit,k}$	$N_{ref,k}$	所佔百分比	擊中率
Stop-Stop	1640	2862	11.78	57.30
Vowel-Stop	2170	2622	10.79	82.76
Stop-Vowel	2227	2572	10.59	86.58
Semivowel&Glide-Vowel	1121	1962	8.07	57.13
Fricative-Vowel	1162	1903	7.83	84.70
Vowle-Fricative	1553	1885	7.76	82.38
Vowel-Nasal	1239	1673	6.88	74.05
Vowel-Semivowels&Glide	559	1056	4.34	52.93
Nasal-Vowel	680	893	3.67	76.14
Stop-Semivowel&Glide	633	767	3.15	82.52
Stop->Fricative	456	703	2.89	64.86
Fricative-stop	396	645	2.65	61.39

除了一些傳統衡量分段系統的評量外，本文還計算了分類音素邊界的擊中率，結果顯示系統對於塞音(Stop)接母音(Vowel)、母音接塞音、擦音(Fricative)接母音、母音接擦音、母音接鼻音(Nasal)、鼻音接母音、塞音接半母音與流音(Semivowel&Glide)的邊界擁有較高的偵測率，而在鼻音接鼻音、半母音接靜音、鼻音接靜音、鼻音接塞音的邊界偵測效果仍有改善的空間。

整體而言，在 3 種容忍度 (20 ms、15 ms 及 10 ms)進行兩種分段評估(F-值及 R-值)，結果顯示在 640 句的語料在 20 ms 容忍度下有 422 句的 F-值在 70%以上，且這 422 句的平均 F-值達到 76.3%，全部的平均 F-值也能達到 72%。另一種評量分段演算法好壞的 R-值，640 句語料在±20ms 容忍度下的平均 R-值有 75%的表現，文獻[14]實驗於 TIMIT 語料庫，在多分率很低的情形下約有 73%的擊中率，其實驗結果之 R-值約爲 75%與本文之實驗結果相當。

七、結論

本文提出一次只偵測一個音素邊界點的方法，與其他文本不特定音素分段的方法比較，本系統之設定傾向於“少分”以避免多分時必然產生的錯誤邊界。系統經由 TIMIT 英語語料測試，驗結果顯示在 20 ms 容忍度下，系統的多分率約爲-6%，但擊中率(Hit Rate)與精確率(Precision Rate)仍分別有 70%與 75%的表現，整體而言 F-值達 72%以上，表示

本系統之設計具可行性。惟鼻音在低頻部分與半母音性質相近，以及塞音之持續時間(duration)過短，因此較難做出有效的偵測，此乃系統之設計有待改善之處。

參考文獻

[1] 林宥余,“高解析度之國語類音素單元端點自動標示”,電信工程學系碩士班，國立交通大學，中華民國九十八年六月

[2] I. Mporas, T. Ganchev, and N. Fakotakis, “A Hybrid Architecture For Automatic Segmentation Of Speech Waveforms,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process 2008*, pp. 4457-4460.

[3] J. Hosom, “Automatic phoneme alignment based on acoustic-phonetic modeling,” in *Proc. Int. Conf. Spoken Lang. Process.*, 2002, vol. 1, pp.357–360.

[4] G. Almpanidis, M. Kotti, and C. Kotropoulos, “Robust Detection of Phone Boundaries Using Model Selection Criteria With Few Observations,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol.17, no.2, pp.287-298, 2009

[5] F. Brugnara, R. De Mori, D. Giuliani, and M. Omologo, “Improved connected digit recognition using spectral variation functions,” in *Proc. Int. Conf. Spoken Lang. Process.*, 1992, vol. 1, pp. 627–630

[6] C. Mitchell, M. Harper, and L. Jamieson, “Using explicit segmentation to improve HMM phone recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, vol. 1, pp. 229–232.

[7] S. Mallat, “A Wavelet Tour of Signal Processing,” Academic Press 1998.

[8] S. S. Cheng and H. M. Wang, "A Sequential Metric-based Audio Segmentation Method via The Bayesian Information Criterion," *EuroSpeech 2003*, pp. 945-948.

[9] S. Chen and P. Gopalakrishnan, "Speaker, environment, and channel change detection and clustering via the Bayesian information criterion," DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[10] O. J. Rasanen, U. K. Laine, and T. Altosaar, “An Improves Speech Segmentation Quality Measure: the R-value,” *Interspeech* 2009

[11] G. Aversano, A. Esposito, and M. Marinaro, “A new Text-Independent Method for Phoneme Segmentation,” *Proc.IEEE International Workshop on Circuits and Systems*, vol.2, pp. 516-519, 2001