新世代自動語音辨識技術–第二階段

– 國語及方言之音節階層事件偵測及其相關研究

計畫編號：97-2221-E-009-080-MY3

目錄

中文摘要

中文關鍵詞

新世代自動語音辨識系統，發音特徵變化點，語音屬性，整合式語音音節端點與屬性偵測器

在新世代自動語音辨識技術中，將結合語音與語言學知識，以多種語音屬性(attribution)與語音事件(event)偵測器群，盡可能從語音信號中擷取各種聲學訊息，以提供後級『語音事件及相關知識整合』及『語音證據確認』單元，做語音辨認甚至於語意瞭解，以期突破傳統隱藏式馬可夫模型方式的困境。新世代自動語音辨識技術或稱之為以偵測器為基礎(detection-based)的架構，不再是直接對整句語音信號做辨認，而是偵測出語音信號中我們感興趣的部分，如：詞、片語或觀念(concept)而已。此時偵測器群不只是像傳統語音辨認架構中之參數抽取所扮演的角色，它能找出語音信號中的時序資訊以及語音特徵，所以新世代自動語音辨識技術中的發音特徵變化點(landmark)之偵測就變成十分的重要了。

在本計畫中以精確的偵測語音信號中的發音特徵變化點(landmark)為起點，將進行下列研究：

(1) 具高解析度 TIMIT 音素端點偵測器－
計畫中首先充分利用語言學家的知識以建立準確至語音信號取樣點的發音特徵變化點偵測器，使用取樣點式之語音聲學參數製作一個可靠的音素端點偵測器。經實驗證實，本計畫中所提出使用取樣點式之語音聲學參數之語音音素端點偵測器效能遠優於使用音框升學參數之系統；
(2) 國語語音之類音素端點自動分段器－
計畫中使用取樣點式之語音聲學參數，來對國內之 TCC-300 語料庫及 Treebank 語料庫做語音類音素端點自動分段；
(3) 方言之類音素端點自動分段器－
計畫中將進一步製作台灣常用之方言－客家話之類音素端點自動分段，以證實計畫中所提出之取樣點式之語音聲學參數及類音素端點自動分段器是可以跨方言的；
(4) 使用取樣點式聲學參數之音素端點偵測器之應用－
在使用取樣點式聲學參數之音素端點偵測器將語音信號切割為一個個音段後，計畫中使用音段式取樣點式聲學參數製作了音段式語音屬性辨認器，經實驗證實及效能亦優於音框式語音屬性辨認器。

本計畫提供其它子計畫所需之語音屬性與事件之資訊，以期建立一套新世代自動語音辨識架構；同時所建立之整合式語音音節及其相關的端點偵測器與語音屬性偵測器也將提供我們以

工程的觀點去探討語言學上的一些現象。

Abstract

Keyword
next-generation automatic speech recognition, speech landmark, speech attribute, integrated boundary and attribute detection

In the next-generation automatic speech recognition paradigm, two types of speech detectors, i.e., landmark (to find the articulation change points in time) and attribute (to find the manner and place of the articulatory) detectors are the fundamental building blocks to reliably phone, word or phrase detection. Especially, landmark detectors are the most important front-end for the following “event merge” and “evidence verification” stages.

In this project, we will focus on developing accurate and reliable landmark detectors and studying the optimal way to integrate them with our well-established attribute detectors (done in previous projects). The following items will be carefully studied and implemented:

(1) Phone boundary detector using sample-based acoustic parameter－
High-resolution sample-based landmark detectors will be developed using articulation parameters. The sample-based acoustic features were proposed to model the rapid spectral changes in speech signal. Both the precision and accuracy of the sample-based phone boundary detector were shown to be better than those of frame-based algorithms.
(2) Force alignment of Mandarin －
The proposed sample-based acoustic features were also used in the force alignment of Mandarin speech, two databases, TCC-300 and Treebank databases were force aligmnet in this project. And, the phonetic unit used in the project was the phone-like units.
(3) Force alignment of Hakka －
Hakka were the most frequently used dialects in Taiwan. In this project, the cross-dialect capacities of the proposed sample-based acoustic festures were cross-examined using Hakka dialects database.
(4) Applications of phone boundary detector using sample-based acoustic parameter －
After phone boundary detection, the speech signal was cut into segments by the boundary candidates. Some segmental parameters found from the sample-based acoustic parameter were used for the pronunciation manner recognition. The accuracy was proved better than the result using frame-based parameters, like MFCC.

In brief summary, the cross-dialect boundary and attribute detector proposed in this sub-project will provide other sub-projects the necessary components to successfully build the next-generation automatic speech recognition paradigm. Moreover, the proposed sample-based acoustic parameters will be cross-examined with linguistic knowledge.

第一章 前言

1.1 研究動機

音素是語音當中最小的單位，且每種語言中某些部分音素的特性是非常相似的，音素之間也能透過適當結合進而組成為音節、詞甚至是片語。正確音素分段位置在語音辨認的研究中可以提升辨識模型的可靠度與統計上一致性進而提升辨識率[1]，也扮演著語音合成方面合成聲音品質提升的重要因素之一。在全球有人工時間標記音素位置的語料庫不多，最著名的是 TIMIT 語料庫，其同時也是本計畫中所使用的語料庫，但是一個大型的連續語音資料庫，使用人工標記音素位置的方式，不僅非常耗時且人工時間標記音素位置也伴隨著一個缺點，就是以人工做標記的動作時，會因為主觀上認定音素端點位置不同，使得標記的位置缺乏一致性，因此一個能夠自動標記且具有精確音素分段的語料庫是非常重要的。

在語音信號處理中，自動音素之分段是一個非常重要的問題，儘管在過去有非常多自動音素分段的研究[2]，一個具有高精準度的自動音素分段演算法，仍是一個可待持續研究的課題。故在本研究中提出取樣點式（sample-based）音素端點偵測方法的架構，來提高音素端點偵測(phone boundary detection)及自動分段位置(phone alignment)的精準度與準確度之效能。

在本計畫中，將以獲得一個良好的音素端點偵測以及自動語音分段系統為目標，因此本研究結合語言學家所提出的（Articulation Parameter，AP），並提出取樣點式音素端點偵測方法的架構，利用數個頻段來區分不同發音特徵之方法，應用於將語音信號做分段可提高時間解析度由音框進一步地精準至取樣點，並在此提出一些取樣點式的聲學參數以用於描述不同語音信號變化時的聲學特性，依此來調整音素位置之標記。

1.2 過去相關研究

在過去一些自動音素分段與偵測的研究中，主要可分為以數學模型為基礎（Model-based）及以量測為基礎（Metric-based）或是上述兩種方法結合。

在 Model-based 方法中，最常被使用的就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，其效能可在正負 20 ms 之內佔有 90%的包含比率（inclusion rate），而傳統 HMM 是以整段語句所得到最大相似度函數為訓練準則，故其自動分段之位置並非為最佳之音節或音素端點。近年來有學者提出一些方法，其中以最小邊界錯誤（Minimum Boundary Error， MBE）為訓練準則之 HMM[3]，就使用自動與給定之已知端點間誤差最小化作為 HMM 模型之訓練準則，在 TIMIT 語音語料庫中，MBE-HMM 自動分段之邊界與人工標記音素端點誤差範圍 10 ms

之內的比率高達 79.75%，與傳統 ML-trained HMM 模型其百分比 71.23%相比，提昇許多；然而其自動音素分段位置只有 7.89%的邊界在人工標記位置誤差 20 ms 之外。此外，也可進一步使用其它圖形識別的方法如支撐向量機[4]（Support Vector Machine，SVM）、類神經網路[5]（Neural Network，NN），皆可用來對 HMM 之自動分段位置再作進一步地修正以獲得更好的結果。

而在 Metric-based 方法中，我們知道語音信號在一個音素中穩定的信號，其聲學參數變化的速率就是決定一個音素邊界的重要線索，回顧一些文獻如 Rabiner[6]使用頻譜轉換量測（Spectral transition measure）的音素端點偵測方法，應用在 TIMIT 語料庫[7]其效能可達到在誤差 20ms 的容忍範圍內，只有 15%的音素端點位置為偵測漏失 （Missed Detection rate，MD）、22.0% 誤報率（False Alarm rate，FA）。Kotropoulos[8]結合 Kullback-Leibler（KL）距離及貝式資訊法則（Bayesian Information Criterion，BIC）所提出的 DISTBIC 演算法來偵測語音信號之音素邊界端點，其效能在 NTIMIT 語料庫亦可達到 25.7% MD 與 23.3% FA 的結果。

在先前的語音分段或是端點偵測的研究，無論 model-based 或 metric-based 的方法中，常用的語音信號參數多與信號頻譜相關；這些參數描述了發音特徵使得語音信號的特性不同，且一般假設語音信號在短時間內為穩定的特性，故使用音框式（frame-based）的聲學參數，例如梅爾倒頻譜係數（Mel-Frequency Cepstral Coefficients，MFCCs）。然而，在做頻譜分析時會造成時間與頻譜（time-spectrum）上之不確定性（uncertain），所以頻譜參數越精確就會犧牲時間精確度；但在音框式的架構中必須要讓頻譜解析度越精細，以提昇辨認音素能力，而發音器官變化很快的音素如爆破音，其音長可能小於一個音框，使得音框式的方法之語音分段位置與實際正確端點位置之間產生誤差，因此對於音素端點偵測及自動語音分段之研究來說，提昇時間解析度，必可降低大量因音框之時間解析度所造成的誤差。

除此之外，在李錦輝教授所提出之 detection-based ASR 中，我們認為 phone boundary detection 擔任了一個提出系統"同步信號"的重要腳色，如圖 1.1 所示。有了 phone boundary 資訊後，不論語音特徵偵測器(attribution detector)或語音辨認的解碼(decoder)工作都可以同步進行，將有助於提升系統效能。而使用語音信號取樣點為單位的 phone boundary detector 更可大幅提高同步信號的精確性。

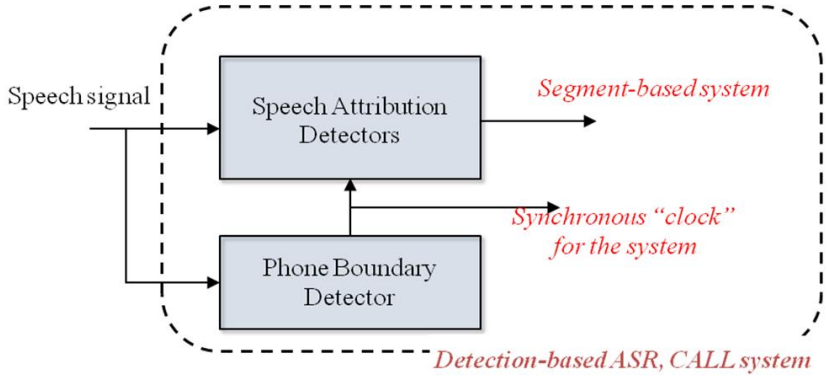


圖 1.1: 使用 phone boundary detection 的同步語音解碼系統示意圖

第二章 取樣點式之語音聲學參數

傳統語音聲學參數與本計畫所提出之取樣式聲學參數最大的差異即是時間與頻譜的取捨，在傳統上抽取聲學參數方式通常假設語音信號為短時間穩定而依固定的取樣點數作為一個音框，音框可視需要改變音框平移以及音框長度，並以此音框為單位抽取語音信號的聲學參數。音框平移的寬度影響時域上音素標記的精準度，音框長度影響著語音信號在頻譜之細膩程度。但在音素分段的觀點，上述這兩種影響卻是不必要的，語音信號的特性雖表現於頻譜分佈上，不過語音信號為時變的，音框式之時間解析度較大，音素之端點位置即使標記在正確的音框內仍會與實際正確端點位置之間產生誤差。本計畫所使用的聲學參數結合語言學家所提出的聲學參數，並應用於本計畫所提出之音素端點偵測以及自動音素分段的研究方法。

2.1 取樣點式聲學參數之語音特徵

本計畫提出一些取樣點式聲學參數如子頻段之信號波封[9]（sub-band signal envelope）、上升率[9]（rate of rise，ROR）、頻譜熵[10]（spectral entropy）、頻譜 KL 距離（spectral KL distance），這節將列舉數個計劃中所提出之聲學參數範例以觀察在不同語音信號或是語音屬性的變化時呈現出的聲學特性為何。以下，進一步介紹本研究所使用的語音特徵參數：

2.1.1 子頻段信號波封

在語言學家所提出的聲學參數中，有許多帶通濾波器能量（band-energy），它們各自能用來區別不同的發音方式或發音位置，常見的頻段[9]（filter bank）有以下：

0.0 – 0.4 kHz 0.8 – 1.5 kHz 1.2 – 2.0 kHz
2.0 – 3.5 kHz 3.5 – 5.0 kHz 5.0 – 8.0 kHz

例如在摩擦音、塞擦音中，在頻譜中之高頻段成份能量極強，低頻段成分能量較弱，鼻音韻尾或是母音的部分則是在低頻段的成份能量極強。這些頻段中能量在有明顯變化的時候，可視為是語音信號開始改變的地方。但語言學家所使用的聲學參數為信號波封（signal envelope），而非現今語音辨認器中常用的能量。故我們將這六個頻段能量取出它的波封來當作本研究中所使用的聲學參數。

在製作一個波封檢測器（envelope detector）的同時，為了保持在波封變化時之信號能正確地描述信號的波封變化，其變化即為頻段信號波封的表示方式；使用希爾伯特變換（Hilbert

transform）來求取輸入信號的波封是一個適當且普遍的方法，其中 $H(x[n])$ 為輸入信號 $x[n]$ 的希爾伯特變換，若輸入信號為頻段之能量 $x[n]$，其 $H(x[n])$ 即為語言學家所使用信號波封，如下式：

$$y_i[n] = x_i[n] + j\left(x_i[n] \otimes h_d[n]\right) = e_i[n]\, e^{j\Phi_i[n]} \quad \text{for } i = 1, \cdots, 6$$

其中

$$h_d[n] = \begin{cases} 1 \ (n-N)\pi \quad , n & \text{i s} < n \text{o} < \text{d} N \text{d} \\ 0, & \text{otherwise} \end{cases} \qquad \text{(2-1)}$$

圖 2.1 即為語音信號經波封檢測器輸出之波封結果，其表示語音信號的輪廓，但是觀察輪廓時卻沒有明顯的規則可做為分辨音素端點的依據，故轉而觀察語音信號在使用六個頻段中之分佈，並依此分佈之特性來區分不同的音素。

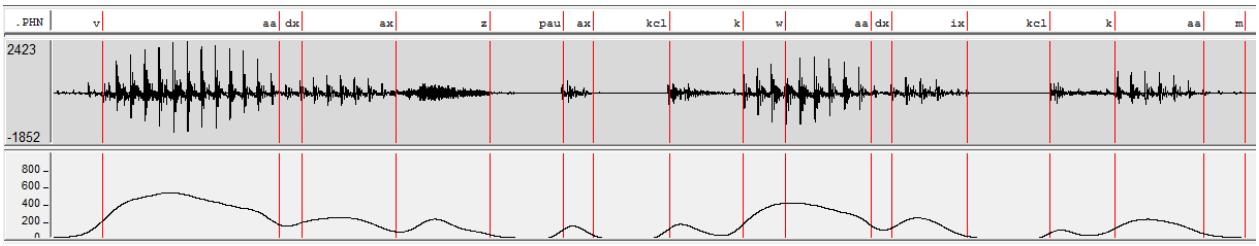


圖 2.1：取樣式語音波封聲學參數範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、音高軌跡（pitch contour）、語音信號之波封

另外，考慮語音信號之波封受到喉頭震動的影響（尤其在音高較低的男性影響越顯著），其會造成語音信號的特性與喉頭震動的周期產生某種程度的關聯性或是造成語音信號的不連貫性，使得波封出現不是預期該有的波動而產生失真。為避免如以上所述之影響，藉由調整波封檢測器的低通濾波器頻寬（passband bandwidth）、截止頻率的衰減斜率來達到其參數物理意義之目的。由簡單的頻寬-濾波器階數定性分析發現，低通濾波器頻寬在 30Hz 至 50Hz 之間並使用相同之濾波器階數，其語音信號波封的輸出結果沒有太大的差異，但其波封變動卻與不同之濾波器階數影響最大，圖 2.2 即是顯現出以上所述之觀察結果。

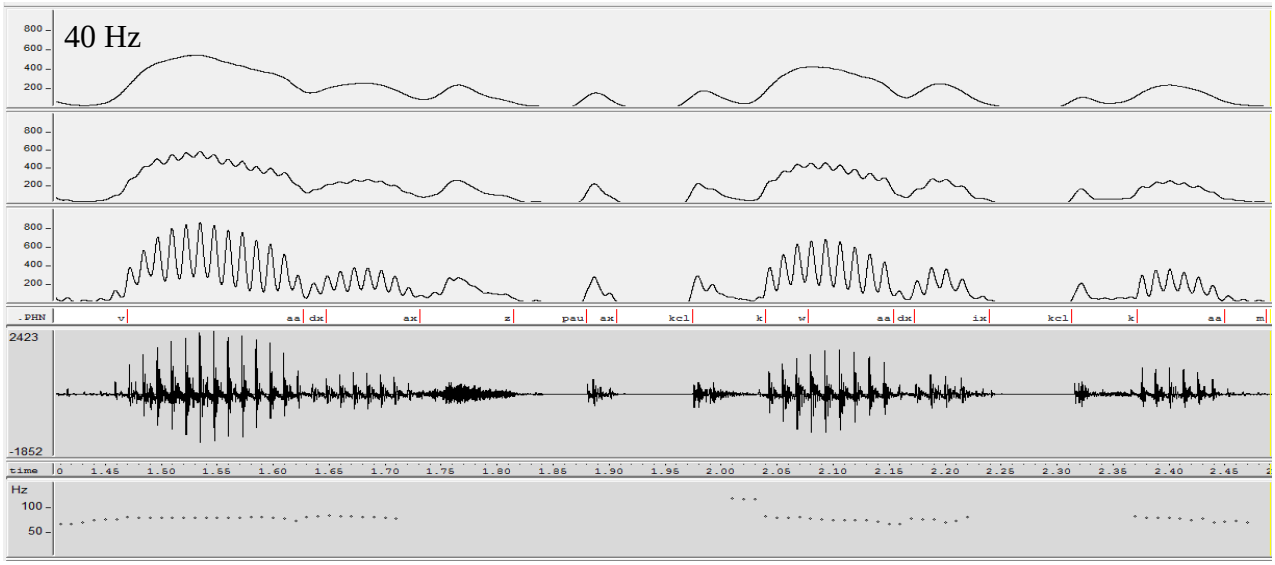


圖 2.2：不同階數之波封檢測器輸出結果，由上至下分別表示波封檢測器使用 40Hz 之 641 階、321 階、161 階低通濾波器的語音信號波封輸出結果、音素層級之人為時間標記的文字轉寫、語音信號、音高軌跡（pitch contour）

2.1.2 上升率

語言學家所稱之上升率，可用於描述語音信號之聲學參數變動的情況，因此藉由參數變動量而觀察發現可能存在的音素端點；其計算方法可對應於音框式抽取語音特徵參數的一階時間軸導數（time derivatives）的方式：在有限的視窗寬度（window width）內，第 n 個取樣點的上升率 $ROR_x[n]$ 依據對應的輸入參數所計算如下式：

$$ROR_x[n] = \frac{\sum_{i=-w}^{w} i \cdot x[n+i]}{\sum_{i=-w}^{w} i^2} \tag{2-2}$$

其中 $x[n+i]$ 為輸入參數資料，w 為計算上升率所使用的視窗寬度。本研究使用語音波形之波封的上升率、頻譜熵之上升率、各頻段信號波封的上升率等當作語音信號的聲學參數，來評量各取樣點式聲學參數的變化率。

透過觀察下圖 2.3 可以發現由人為時間標記對應於語音信號之波封急遽上升的時候，即是該區域波封上升率之局部最大值（local maximum）之端點。在此處之上升率參數可指出語音信號之波封變動最大的端點位置，這種情況尤其好發在音節結構的前端音節頭至音節核的部分，如摩擦音至母音、塞擦音至母音…等等的音素轉換端點，由以上觀察的聲學參數之特性，我們將其輸入參數至換成各頻段的信號波封，那麼我們即可由各頻段信號波封所計算的

上升率來分別找到對應每個頻段其信號波封變動量大的端點。如圖 2.4 各頻段的波封上升率可以對應於聲譜圖（spectrogram）的顏色深淺程度，也就對應至各頻段信號波封的大小變化；語音信號在六個頻段之中之分佈由強(亮)轉弱(灰暗)，其轉變程度越大上升率越高。然而，觀察每個頻段之波封上升率為局部最大值之端點，其會因為信號波封變動量的不同而使得在某一段時間內各頻段之端點位置並不一致，要如何在此一區段時間選擇一個適當的音素轉換端點，將在下節討論。

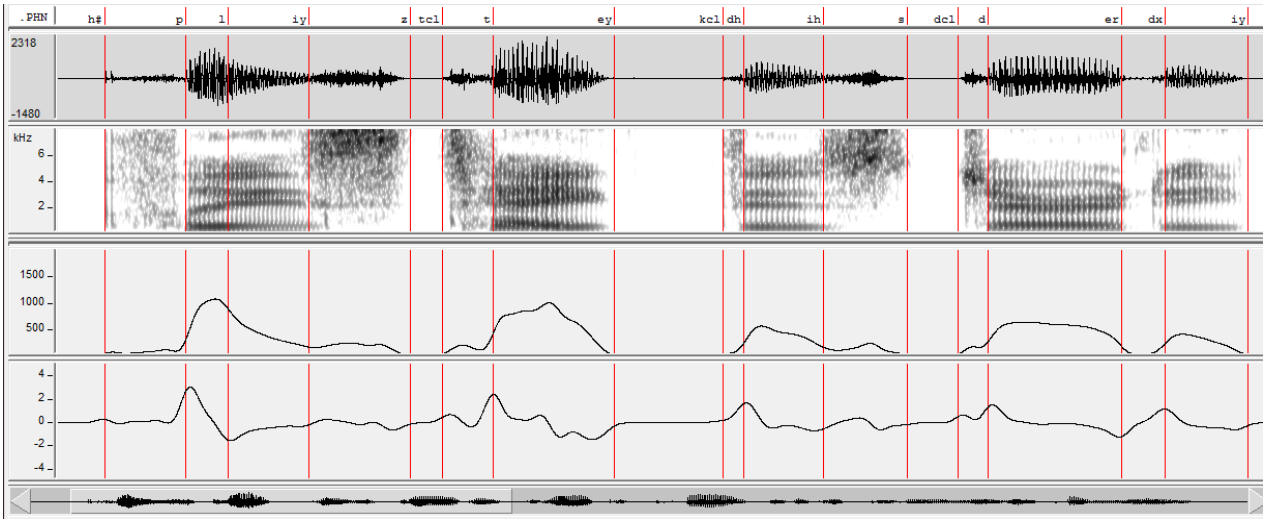


圖 2.3：取樣式聲學參數之上升率範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、語音信號之波封、波封之上升率

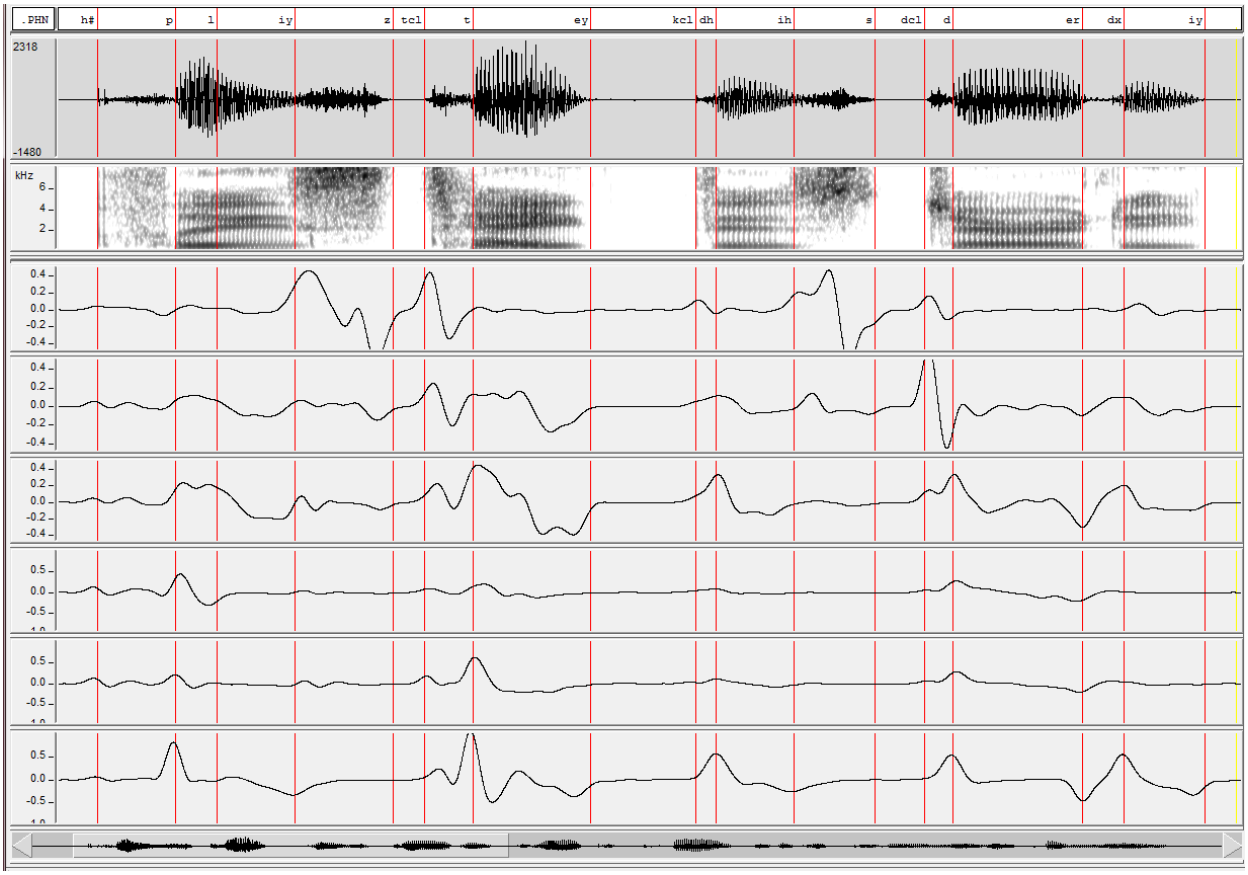


圖 2.4：取樣式子頻段信號波封聲學參數範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、第六個至第一個頻段之信號波封上升率

2.1.3 頻譜熵

頻譜熵可用來描述信號在頻譜上的集中之分佈程度，若語音信號越集中在某一個頻段則頻譜熵越小。在此，本研究使用先前所述之六個頻段，將頻譜熵 $H_s[n]$ 定義如下式表示：

$$H_s[n] = -\sum_i E_i[n]\log\left(E_i[n]\right) \tag{2-3}$$

其中

$$E_i[n] = \frac{e_i}{\sum_{j=1}^{6} e_j} \tag{2-4}$$

$E_i[n]$ 為第 i 個頻段之第 n 點正規化之後的子頻段信號波封。由語音信號對應到頻譜熵的表現上如圖 2.5，可以發現短停頓、靜音內之語音特性只有非語音的雜訊。如背景雜訊在各個頻段都會出現，所以頻譜熵值較高是可以預期的；而母音在頻譜上的能量則較集中於低頻段至中頻段的部分，其頻譜熵值相對較低。同樣地，可依頻譜熵在不同之音素在頻譜上的分佈之

間的變動，求取頻譜熵的上升率。

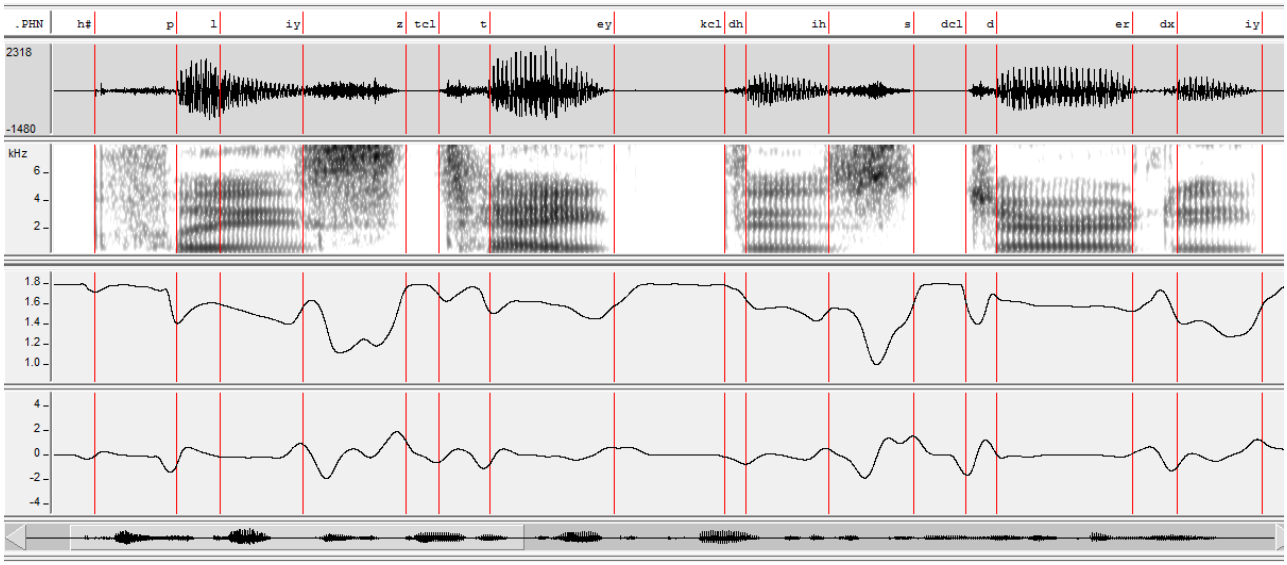


圖 2.5：取樣式頻譜熵聲學參數範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、頻譜熵、頻譜熵之上升率

2.1.3 頻譜 KL 距離

將頻譜視為一個機率分佈的問題，因此可以利用頻譜 KL 距離來描述兩段時間點之頻譜相似程度。在語音信號中計算兩點不同時間(n 與 m)的頻譜 KL 距離，$d_{KL}(n,m)$，可以由下式表示：

$$d_{KL}[n,m]=\sum_{i=1}^{6}\left(E_i[n]-E_i[m]\right)\log\left(\frac{E_i[n]}{E_i[m]}\right) \tag{2-5}$$

而本研究目前為考慮**相鄰語音信號取樣點**之頻譜信號分佈特性，則將(2-5)式改寫為以下：

$$d_{KL}[n]=\sum_{i=1}^{6}\left(E_i[n]-E_i[n+1]\right)\log\left(\frac{E_i[n]}{E_i[n+1]}\right) \tag{2-6}$$

不同音素轉換的時候，其發音的方法或是部位也會跟著轉移，使得不同音素之語音信號轉換至頻譜上的分布情形也會跟著不同，頻譜 KL 距離即是度量在頻譜間的相似程度，且此一度量之特性具有一致性。那麼經由簡單調整一個臨限值（threshold），即可初步地得到一序列（sequence）經由頻譜 KL 距離所挑選出來是具有音素端點可能性的位置。

藉由聲譜圖可以清楚地觀察到在相鄰音素之間的信號分佈變化，如圖 2.6 中同一音素內之頻譜信號分佈為局部穩定的狀態，並在不同音素轉換的區域音其頻譜分佈差異大，使頻譜 KL 距離明顯增大。

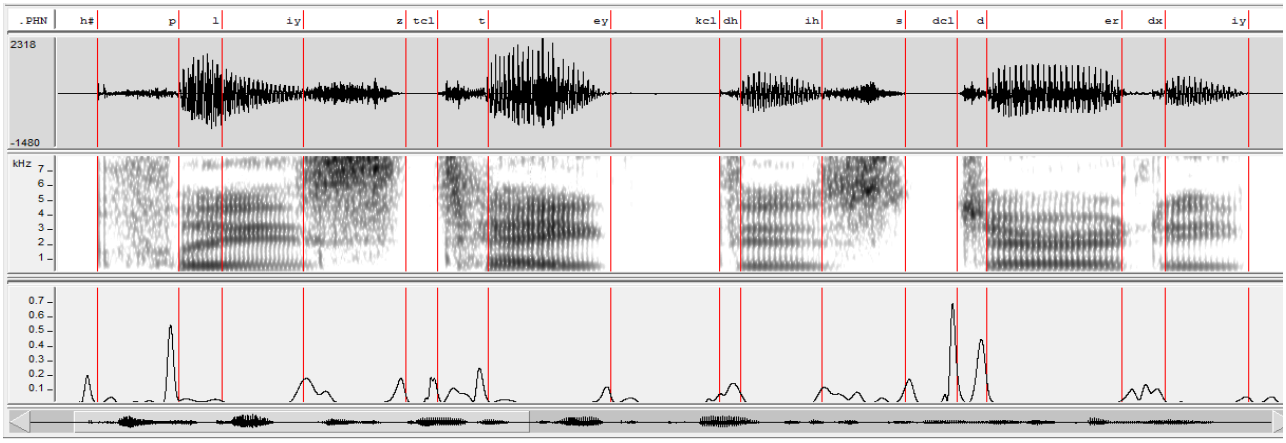


圖 2.6：取樣式頻譜 KL 距離聲學參數範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、頻譜 KL 距離

由前 2.1.1 節所述波封檢測器內低通濾波器之階數，影響著頻段內之信號波封的變動。利用各頻段分佈所計算出來的頻譜 KL 距離也有如下圖 2.7 的差異，在圖中我可以觀察到隨著濾波器階數越低，則頻譜 KL 距離的大小因信號變化而受影響的程度也會增加。假若使用臨限值來挑選一序列音素之候選端點，在高階數的部分，音素端點之候選端點少，其端點雖能表現出信號的重大變化，但有部分的音素端點卻因為臨限值之遮蔽而消失；相對地在低階數的部分，情況卻是完全相反，序列中音素候選端點幾乎能包含原有之音素端點，不過因為其頻譜 KL 距離易受信號變化影響的效應，使得音素候選端點序列中增加極多冗餘的端點。那麼以音素端點偵測的觀點考量，就必須在音素候選端點的數目與參數的穩定度上做一個取捨（trade-off），以達到最佳的結果。

綜合以上所敘述之取樣點式聲學參數，其子頻段信號波封、聲學參數的上升率、頻譜熵及頻譜 KL 距離等語音特徵參數的變化，確實能得到在語音信號變化的時候，可以觀察這些參數的語音特性達到分辨不同音素端點位置之目的。

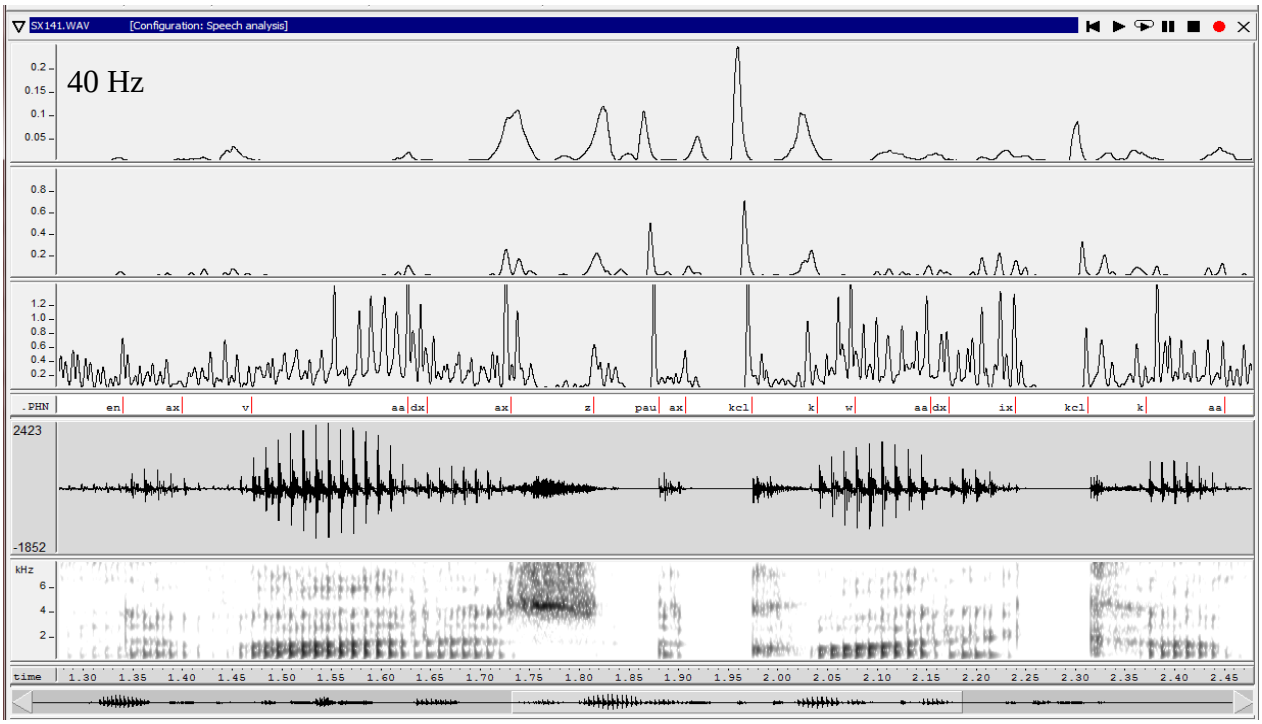


圖 2.7：不同階數之波封檢測器對頻譜 KL 距離的影響，由上至下分別表示波封檢測器使用 40Hz 之 641 階、321 階、161 階 FIR 低通濾波器輸出結果所計算的頻譜 KL 距離、音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖

第三章 語音音素端點偵測器

本計畫中將以國語及台灣方言之語料庫進行音素端點偵測或是自動語音分段的實驗。但現今的國語及台灣方言之語料庫均無人工標示音素端點資訊，也就是說將沒有標準答案；所以在計畫中才會先從一個有人工切割位置的語料庫 TIMIT 著手研究音素端點偵測器。

3.1 TIMIT 語料庫簡介

本計畫中以 TIMIT[8]（The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus，TIMIT）語料庫作為主要實驗與分析之語料。 TIMIT 語料庫是屬於由朗讀句子的語音（read speech）所組成。而語料庫中的這些朗讀語句皆是由德州儀器（Texas Instruments，TI）、麻省理工學院（Massachusetts Institute of Technology， MIT）以及史丹佛研究機構（Stanford Research Institute，SRI）來共同設計而成。其語料庫的語句是德州儀器請美國不同區域的人朗讀並錄製成語音檔案，麻省理工學院進行人工轉寫的步驟。

TIMIT 語料庫中其包含有 6300 個語句，這些語句分別是由美國主要八種不同口音（Dialect）地區的 630 個語者，每位語者朗讀 10 個語句錄製而成。語料庫語句在收錄時以 16kHz 之取樣率經 16 位元量化來錄製單聲道音檔，音檔檔頭為 1024 位元組（byte），以提供語音辨識為主要應用。

每個語者朗讀的 10 個語句中之語句類型，包含 2 句方言（SA）語句，為了顯現不同地區語者口音之差異；5 句 phonetically-compact（SX）語句，為了每個音素出現之頻率能夠相近；以及 3 句 phonetically-diverse（SI）語句，其語句是從當時現存的文字語料庫資源挑出來的，如布朗文字語料庫（Brown Corpus，Kuchera and Francis，1967）等等。

TIMIT 語料庫廣泛地用於各方面有關之語音研究，其原因在於語料庫內之資訊囊括完整的文字轉寫及對應不同層級之人為時間標記；文字轉寫以及其對應字詞層級（word level）及音素（phone level）的人為時間標記，使得 TIMIT 語料庫成為一個平台來提供各式各樣之理論及方法之間進行語音相關研究，並基於此平台驗證其理論、方法或是評量實驗結果效能的好壞。

無論是在何種層級之文字轉寫中，皆是由標音員給予該語音信號正確的標音符號並依其語音段落之起始與結束的語音取樣點作為時間標記，如圖 3.1 所示。如前一章節所述，文字轉寫中的人為時間標記是目前最為準確對語音進行分段的方式，但其標記位置皆含有主觀的判斷且因人而異，容易造成時間標記之不一致性。因此將在本計畫實驗分析時，來討論此現象引起的相關問題。

目前語料庫之音素集包含 61 個音素，如附錄一，音素層級之文字轉寫皆是對應音素集標記而成。但是以音素端點偵測的觀點觀察語音信號的變動時，不同音素語音信號之轉變其無論在頻域或是時域上之特性應是有所差異的，利用此差異我們可以偵測音素端點存在的可能性。而在爆破音（stop consonant）發音前會有所謂的短停頓的產生，在語音學上稱為嗓音

起始時間（voice onset time，VOT），指的是爆破音成阻後持阻到除阻時間，語音學上會將此段短停頓的產生視為爆破音時長的一部份。但在音素端點的偵測內，其語音信號的特性上卻是有著極大的差異。故 TIMIT 語料庫的音素時間標記將此種情形也納入音素時間標記的範疇中，而對該爆破音之標音前的短停頓給予合適的標記符號，其對應的標記符號如下表 3.1。

另外，我們知道英語為 consonant-vowel-consonant 之音節結構，簡稱為 CVC。例如以（rime structure）表示單音節的英文詞 cat，其音節頭（onset）為“c”，音節核為“a”，音節尾（coda）為“t”。而子音在 CVC 音節結構內的位置不同會其發音也不盡相同，以本計畫之音素端點偵測的觀點，我們無須了解其音素在結構內的關係，但若以音素端點切割的方面考量，就必須考慮音節結構對音素端點的影響。

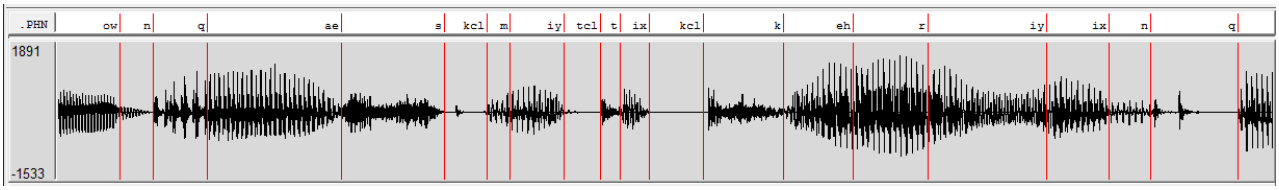


圖 3.1：音素層級之文字轉寫對應於語音信號的人為時間標記

表 3.1：爆破音對應之短停頓標記符號。

stops	*b*	*d*	*g*	*p*	*t*	*k*	*jh*	*ch*
closure intervals	*bcl*	*dcl*	*gcl*	*pcl*	*tcl*	*kcl*	*dcl*	*tcl*

TIMIT 語料庫之訓練語料與測試語料分別為 462 位語者之 4620 個語句與 168 位語者 1680 個語句所建構而成，在本計畫中使用音素層級之文字轉寫的人為時間標記之所有訓練語料來訓練音素端點偵測器的模型，並以測試語料驗證本計畫所提出方法之效能。

3.2 音素端點偵測系統

儘管在不同語言之中，人類的發音系統之構造對語音的影響，在一段語句內即顯現出其音素的語音特性皆與發音部位以及發音方法有非常大的關聯性。由第二章所述，本計畫提出取樣點式聲學參數的聲學特性來描述這些語音信號中不同語音屬性的變化，藉由量測這些變化來找出可能為音素端點的位置，這意謂著進行語音的標記中並不需要完整的音素辨認流程，也不需使用到非常準確的音素標記位置，即可簡化語料庫繁複處理的過程。

端點偵測器以音素層級之人為時間標記文字轉寫來訂定目標函數的兩種轉移狀態，分別為音素端點（T）、非音素端點（nT），對所有由候選端預選（Candidate Pre-selection）所選取之候選端點對應文字轉寫標記目標函數的種類，並用於端點偵測器的訓練。其中，對於每個候選端點其包含了自身端點的聲學特性及其與前後相鄰候選端點之間的音段聲學特性，最後經由多層感知器的學習特性，反覆疊代訓練將音素端點與非音素端點的語音特性做分類，並藉此模型達到音素端點偵測的目的。

本計畫所建構之音素端點系統是利用英文 TIMIT 語料庫所提供之人為時間標記的文字轉寫作為音素端點偵測器模型初始化訓練之目標。採用半監督式的訓練方式，來獲得一個端點偵測器模型。利用訓練後的音素端點偵測器模型，對不同語料庫進行音素端點的偵測，實驗結果將於下章節做分析。圖 3.1 為訓練音素端點偵測系統之流程圖，分為抽取聲學參數以及音素端點模型之訓練方式兩個部分。

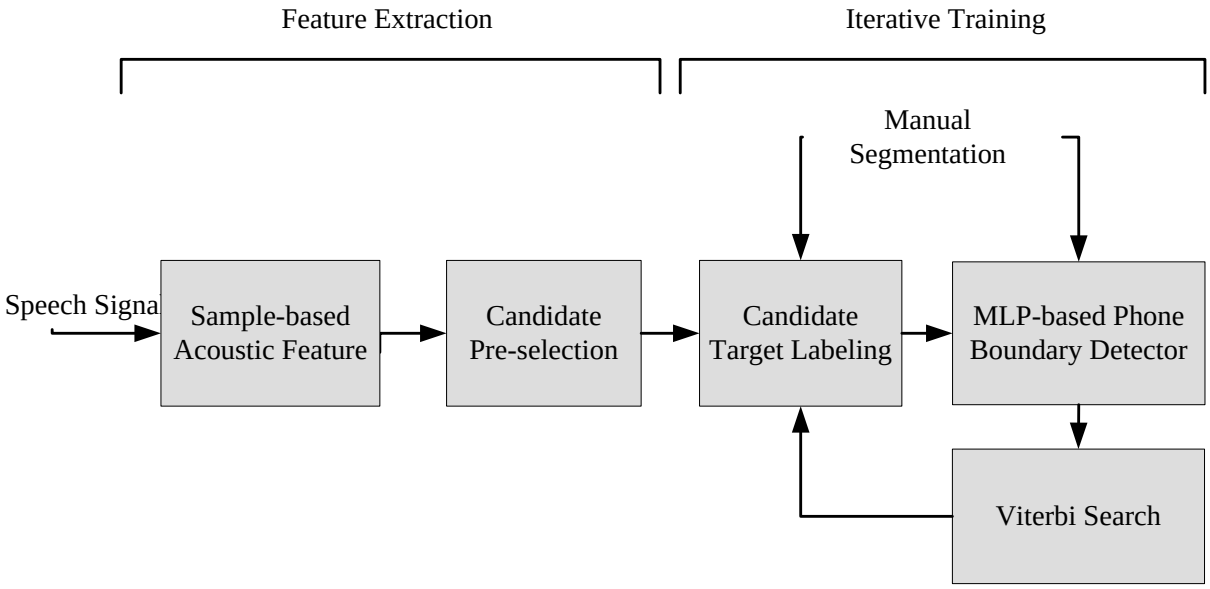


圖 3.1：使用多層感知器架構之音素端點偵測器

3.2.1 端點候選者之預挑選方式（Candidate Pre-selection）

取樣點式的音素端點偵測架構中，首先使用計算同第二章節所述之取樣點式聲學參數，來得到 6 個子頻段信號波封，值得注意的是在此一計算過程當中做了一些適當的調整。即計算時將這 6 個子頻段信號波封輸出加上一個臨限值，此臨限值是為降低每個頻段微弱信號部分的變動影響，例如雜訊。

$$E_i[n]=\begin{cases}\dfrac{e_i[n]}{\sum_{j=1}^{6}e_j[n]}, & \dfrac{e_i[n]}{\sum_{j=1}^{6}e_j[n]}>\eta \\ \eta, & \text{otherwise}\end{cases} \tag{3-1}$$

從語音信號中抽取聲學參數之後，為了減少在端點偵測器內過於龐大的資料計算量，經由預選擇即如同 2.1.3 節所敘述，藉由簡單設定一個臨限值（Th_d）的方法來挑選可能較大之音素端點位置；由於頻譜 KL 距離在挑選出語音信號相鄰時間中的變化上是一種很好的量測方式，故若頻譜 KL 距離滿足下式：

$$d_{KL}[n-1] < d_{KL}[n],\ d_{KL}[n] > d_{KL}[n+1] \text{ and } d_{KL}[n] \geq Th_d \tag{3-2}$$

則代表為挑選出來的候選端點值，最後得到這一序列音素的候選端點，$\{c_j; j=1,...,N\}$。

經過預選擇步驟後，會將音素端點候選者之數目大量降低，也就是可以降低音素端點偵測器之運算量。

在此實驗過程中依照觀察頻譜 KL 距離與人為時間標記之間的關係發現一些現象，舉例來說對於人為時間標記中之摩擦音至母音、流音之間的音素轉換端點，在聲譜圖中可觀察到端點兩邊頻譜信號分佈的差異極大如圖 3.2 中的（/k/-/l/）、（/t/-/ix/）之轉換端點，圖中可以看到人為時間標記的位置並不一定是相鄰區域中頻譜 KL 距離局部極大值的端點，而是黑色箭頭所指向的端點；另外，圖中偏右旁的（/k/-/l/）音素轉換端點之相鄰區域中並無特別大的頻譜 KL 距離，那麼要如何選擇最適當的音素候選端點能減少訓練音素端點偵測器所需要達到收斂的次數？ 此問題即為先前所描述其人為時間標記之語料庫其標音員之主觀性所產生時間標記位置之不一致性的問題。

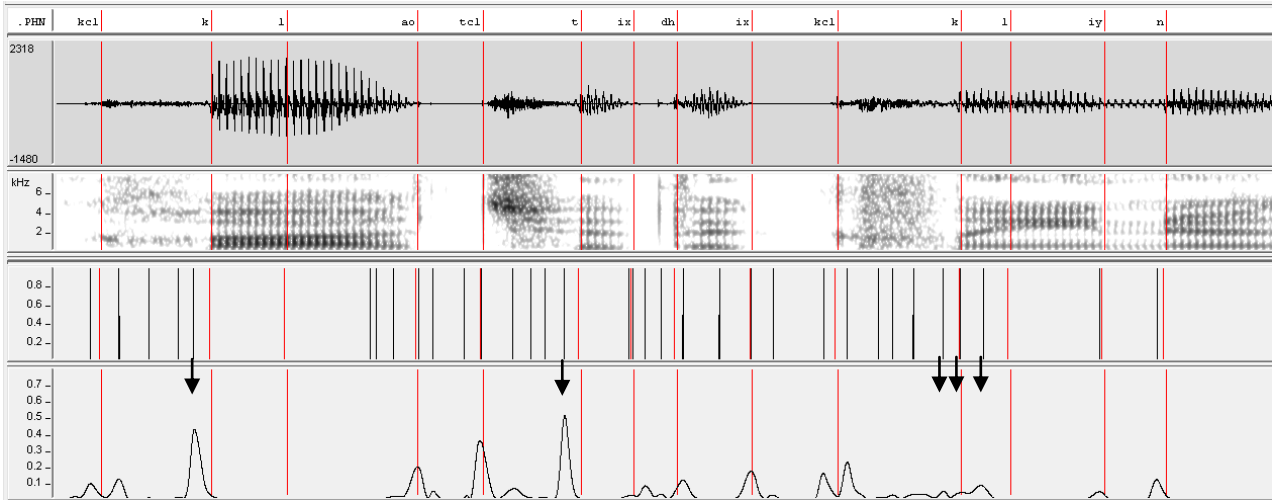


圖 3.2：調整音素候選端點之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素候選端點、頻譜 KL 距離

因此，本計畫提出一個演算法用以挑選出候選端點序列中最佳的音素候選端點作為半監督式學習的目標（Target）。

其演算法的敘述如下：

(1) 在人為之時間標記音素端點之相鄰區域選擇適當的範圍，本計畫使用相鄰音素端點之中點作為上限（Upper bound，UB）與下限（Lower bound，LB）且前後以不超過 30 毫秒的範圍作為挑選候選端點的區域 R。

(2) 在區域 R 內頻譜 KL 距離挑選出來之候選端點即為第 i 個音素端點之候選端點子序列 $\{c_{i,j}; j=1,...,k\}$，並將此子序列依候選端點與該音素端點之距離由近至遠排序。

(3) 將排序好的候選端點子序列依照臨限值[1]（Th_c）判斷，得到此序列中最佳的音素候選端點 $c_{i,j}$，並標記此候選端點為第 i 個音素端點所要學習的目標。

(4) 重複(1)、(2)、(3)的步驟直至所有音素端點皆經過計算後，求得所有最佳之音素候選端點並完成學習目標的標記。

藉由候選端點會將語音信號分割成很多音段（Segment），反而言之，這些音段相較於由頻譜 KL 挑選之音素候選端點的語音特性是可視為穩定的，故即可使用這些音段之語音信號求取一些音段式（Segment-based）的聲學參數來描述候選端點兩旁之語音特性，以協助進行音素之端點偵測。

首先，本計畫使用音段式的子頻段信號波封（Segmental sub-band signal envelope）來表示 2 個相鄰的音段$[c_{k-1},c_k]$、$[c_k,c_{k+1}]$內其語音信號在頻譜的分佈情形，在此以下圖 3.3 來作說明。圖中候選端點 k 之高度表示頻譜 KL 距離數值之大小，其前、後音段(Segment k-1、Segment k)則分別表示在候選端點間其語音特性的狀態，假若候選端點相鄰兩旁音段之頻譜信號分佈差異極大，代表其語音信號轉變而造成其分佈差異，那麼即可增加此一輔助資訊來提升音素端點偵測之效能。因此，本研究定義候選端點相鄰音段 $ES_i(k)$ 為在第 k 個音段 $[c_{k-1},c_k]$ 中其子頻段信號波封經正規化後的平均值，如下式：

$$ES_i[c_{k-1},c_k]=\left(\sum_{n=c_{k-1}+\Delta}^{c_k-\Delta}E_i[n]\right)/(c_k-c_{k-1}-2\delta) \tag{3-3}$$

其中 δ 表示與候選端點 k 相距的取樣點個數。

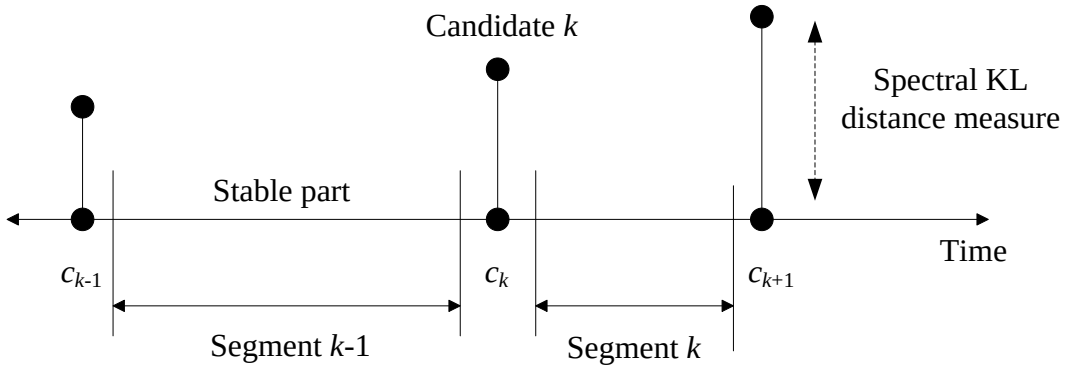


圖 3.3：利用候選端點將語音信號分割成片段的示意圖

[1]經由觀察頻譜 KL 距離對應語音信號變化之數值我們設定一臨限值，假若其候選端點之頻譜 KL 距離大於臨限值我們便認為其端點是極有可能為音素端點的位置。

另外，我們對候選端點，c_j，左右各取一小段語音信號來量測其相似度。這兩小段信號的區間，B_j^- 及 B_j^+，它們分別表示為

$$B_j^- = [c_j - r_j^-, c_j - 1] \ , \ B_j^+ = [c_j, c_j + r_j^+],$$

其中 r_j^- and r_j^+ 為兩的區間內的語音樣本數分別為

$$r_j^- = \begin{cases} r_{\min}, & c_j - 1 - c_{j-1} < r_{\min} \\ c_j - 1 - c_{j-1}, & r_{\min} \le c_j - 1 - c_{j-1} \le r_{\max} \\ r_{\max}, & r_{\max} < c_j - 1 - c_{j-1} \end{cases}$$

及

$$r_j^+ = \begin{cases} r_{\min}, & c_{j+1} - c_j < r_{\min} \\ c_{j+1} - c_j, & r_{\min} \le c_{j+1} - c_j \le r_{\max} \\ r_{\max}, & r_{\max} < c_{j+1} - c_j \end{cases}$$

其中 $r_{\max}$ 及 $r_{\min}$ 及音段最大及最小長度。如果我們將此兩音段之子頻段信號波封參數視為高斯分布。則我們可以使用下列 KL 距離來描述這兩音段的相似度，

$$D_{KL}[c_j] = \frac{1}{2} tr[(\Sigma_- - \Sigma_+)(\Sigma_-^{-1} - \Sigma_+^{-1})] + \frac{1}{2} tr[(\mu_- - \mu_+)^T (\Sigma_-^{-1} + \Sigma_+^{-1})(\mu_- - \mu_+)] \quad (3\text{-}4)$$

上式中，μ_- 及 μ_+ 分別表示兩音段子頻段信號波封參數之平均向量；Σ_- 及 Σ_+ 為子頻段信號波封參數之變異矩陣。

接著，考慮相鄰候選端點之時間關聯性與其端點間語音特性之相關性，對於每個候選端點建立一個 27 維的聲學參數向量，對於第 k 個候選端點，c_k，其聲學參數向量包括以下聲學參數：

(1) 目前候選端點及前、後候選端點之參數：

$$d_{KL}[c_k],\ D_{KL}[c_k],\ H[c_k],\ \Delta H[c_k],\ \left(E_i[c_k]; i = 0, \cdots, 6\right)$$

其中 $\Delta H_s[c_j]$ 為頻譜熵之一階差量。

(2) 目前音段及前、後音段之參數：

$$\left(ES_i\left[c_{k-1}, c_k\right],\ ES_i\left[c_k, c_{k+1}\right]; i = 1, \cdots, 6\right),\ c_k - c_{k-1},\ c_{k+1} - c_k$$

其中 $c_k - c_{k-1}$, $c_{k+1} - c_k$ 表示目前端點與前後相鄰端點之時間資訊。

最後，由語音信號所抽取之每個聲學參數向量皆存在聲學參數檔案內，以提供後級音素

端點偵測器之訓練使用。圖 3.4 展示了抽取聲學參數演算法的整體架構。

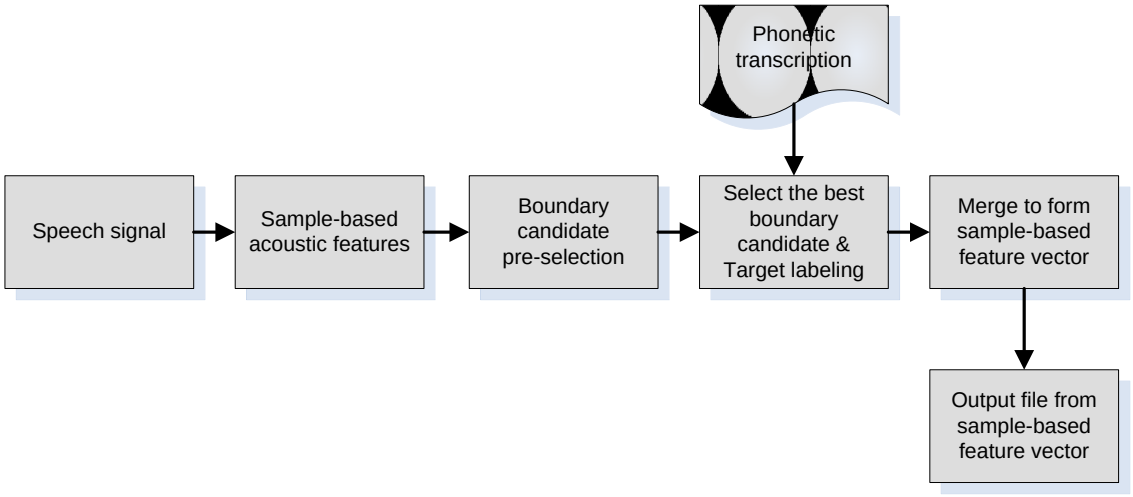


圖 3.4：聲學參數抽取演算法的系統架構圖

3.3 使用多層感知器及 RNN(Recurrent Neural Network) 之類神經網路架構之音素端點偵測器

完成語音之聲學參數萃取後，本節中將介紹音素端點偵測器模型之演算法，本研究使用 NIKO toolkit[12]多層感知器之類神經網路架構，將基於使誤差值最小化的準則（Error minimization）採用倒傳遞演算法（Back-propagation algorithm）將先前所建置之取樣式聲學參數進行參數資料的分群訓練與模型目標函數的更新。

在有 TIMIT 語料庫人為時間標記之文字轉寫作為模型初始化訓練後，為實現半監督式的訓練方式，以下將介紹訓練音素端點偵測器模型反覆疊代的步驟，其流程圖如圖 3.5：

- ***Step1*：將多層感知器輸出之概似度（likelihood）正規化為機率**

依照目標函數的個數將多層感知器之輸出層對應每個輸入聲學參數向量所產生之概似度作正規化，則得到該參數向量在各個目標函數機率。

- ***Step2*：更新文字轉寫之自動時間標記**

接著，使用維特比搜尋演算法（Viterbi search algorithm）重新將文字轉寫作強迫對齊，以得到一個更新後的自動語音分段位置。

- ***Step3*：重新標記目標函數**

在有一個經重新自動分段後的文字轉寫，由文字轉寫內的時間標記將端點位置再重新標記目標函數，並作為下一次多層感知器之學習目標。

- ***Step4*：更新多層感知器之目標函數**

置換多層感知器的目標函數，繼續訓練音素端點偵測器之模型。

➢ ***Step5*：重覆 *Step1* 到 *Step4* 至收斂**

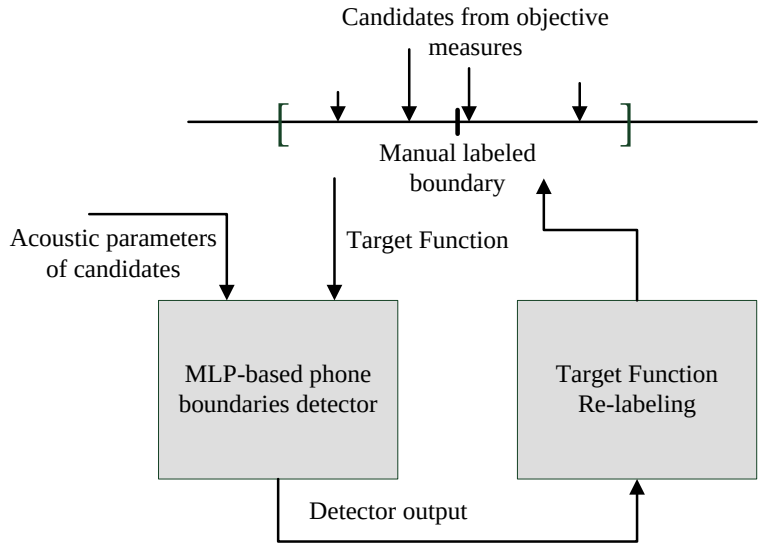


圖 3.5：音素端點偵測器模型反覆疊代之流程圖

3.4 音素端點偵測實驗結果分析

使用 TIMIT 語料庫來驗證本計畫提出音素端點偵測器的偵測效能，並依照 TIMIT 語料庫所建議訓練語料 4620 個語句及測試語料 1680 個語句的分類，用於音素偵測實驗。首先，表 3.2 統計了訓練語料與測試語料所處理的語音取樣點、音素邊界候選端點（Candidate）以及語料中所要偵測之音素邊界總數（Phone boundary）的數據，由此可推得訓練語料約 1314 個語音取樣點也就是平均約 82.125 毫秒有一個音素端點的存在，而測試語料則是平均每音素端點相隔約 82.83 毫秒，皆與平均音素長度為 50 至 100 毫秒或是約為 5~10 個音框長度的統計量相符；挑選音素邊界候選端點時適當設定臨限值，分別在訓練語料及測試語料挑選出 534189 與 194201 個可能為音素邊界的候選端點，以提供音素端點偵測器的訓練及測試。

在實驗中，我們使用了兩種類神經網路架構，多層感知器(MLP)及 Recurrent Neural Network(RNN)，其隱藏層神經元數目分別為 75 及 80 個。最後比對經人為標記的音素層級之文字轉寫而得到偵測音素邊界端點其誤報率與偵測漏失率相等時之錯誤率（Equal error rate，EER）效能為 11.6%與 8.6%。而偵測漏失率與誤報率的定義如下式表示：
偵測漏失率為未偵測到之音素邊界端點個數 D 在總音素邊界端點個數 N 中所佔的比例。

$$\text{Miss Detection rate} = \frac{D}{N} \times 100\% \tag{3-5}$$

誤報率表示誤偵測為音素邊界端點個數 I 在總音素邊界端點個數 N 與 I 之總和中所佔的比例。

$$\text{False Alarm rate} = \frac{I}{I+N} \times 100\% \tag{3-6}$$

表 3.2：TIMIT 語料庫的統計資料結果

TIMIT corpus	Sample	Candidate	Phone boundary
Training part	226727341	534189	172461
Test part	82786737	194201	62466

在測試語料中所挑選出的候選音素端點，可藉由加上不同的臨限值來控制音素端點偵測器所偵測的音素端點個數，因此實驗中對應不同的的臨限值描繪出誤報率與偵測漏失率的對應曲線圖為圖 3.6 所表示，圖中♦點為 Rabiner 在數據中近乎 EER 的數值點，而本計畫測試語料使用 MLP 及 RNN 的實驗結果分別以黑色線實線及虛線表示，而傳統 HMM 所辨認出之音節結果則為●點。然而，誤報率與偵測漏失率為成反比的，在本計畫音素端點偵測的觀點中，誤報率的增加代表著有更多音素候選端點被誤認為音素邊界端點的可能性被提高，但音素候選端點是以評量相鄰語音取樣點頻譜差異的頻譜 KL 距離所挑選出來，有些音素的連音現象造成不明顯的頻譜變化，這些部分為較難偵測的音素端點，藉著調降臨限值使誤報率增高，造成對應較難偵測的音素邊界端點也可一併偵測出來，進而減低音素端點偵測的漏失。音素端點偵測的目標為減低人為標記語料庫的繁複過程，過大的偵測漏失率即為音素偵測實驗最不想見的結果。在此，找出誤報率與偵測漏失率之間的取捨平衡點亦即當誤報率與偵測漏失率相同，作為實驗結果的比較方式。

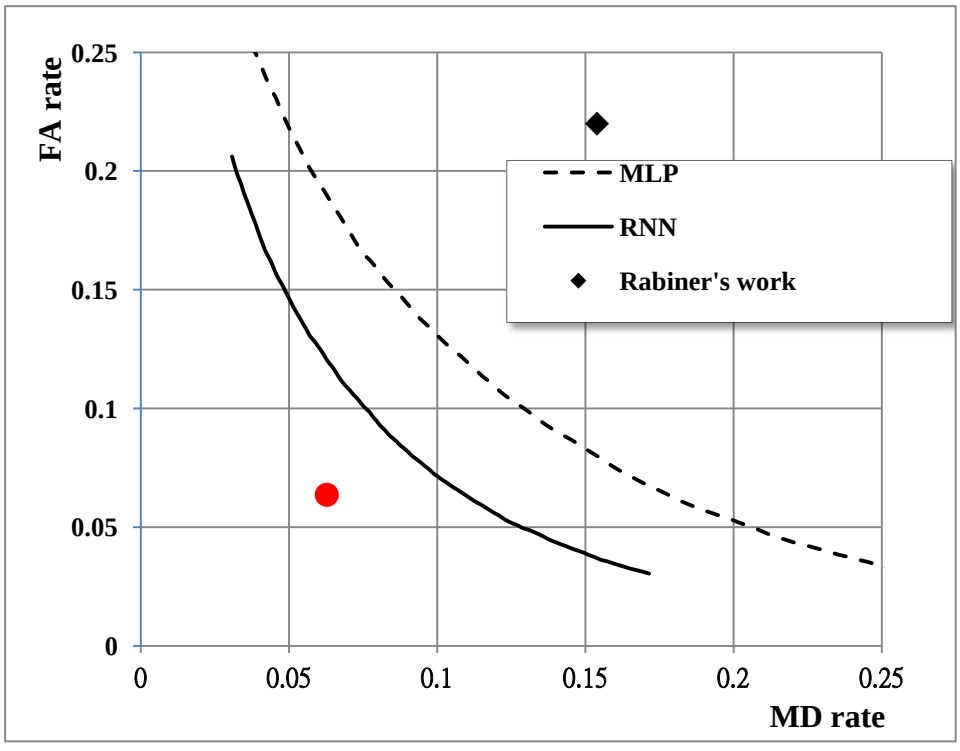


圖 3.6：音素端點偵測器於 TIMIT 語料庫誤報率與偵測漏失率之對應曲線圖

接著，為了能與傳統音框式方法比較實驗的結果，我們統計每 5 毫秒所包含到偵測音素

邊界的比例，並計算被偵測到音素端點落在相同或是相鄰音框之內的包含比例，以評量本計畫之音素端點偵測器之效能好壞。其中表 3.3 顯示在 EER 的情形下，偵測到的音素邊界端點在不同絕對偏差值內（5、10、15 毫秒）的包含比率，而在相同音框內為 41.72%，相鄰音框範圍內為 87.32%，兩種評量之實驗結果皆優於 Rabiner（27%/ 10ms, 70%/ 20ms），可易見時間解析度較細的取樣點式的音素端點偵測方法有較高的效能。圖 3.7 顯示了音素端點偵測器之實驗結果與人為標記之間的差異在不同絕對偏差值的差異的區間內，佔有總音素端點個數的比例。絕對偏差值越小代表著與人為標記位置越相近，亦表示偵測出之音素候選端點越準確。

表 3.3：使用音框式計算音素邊界偵測結果的方式的統計結果，音框平移為 10ms

Methods	In the same frame	within ±1 frame
HMM	27.5%	67.3%
Rabiner's [17]	22.8%	59.2%
MLP	36.0%	73.9%
RNN	37.3%	77.0%

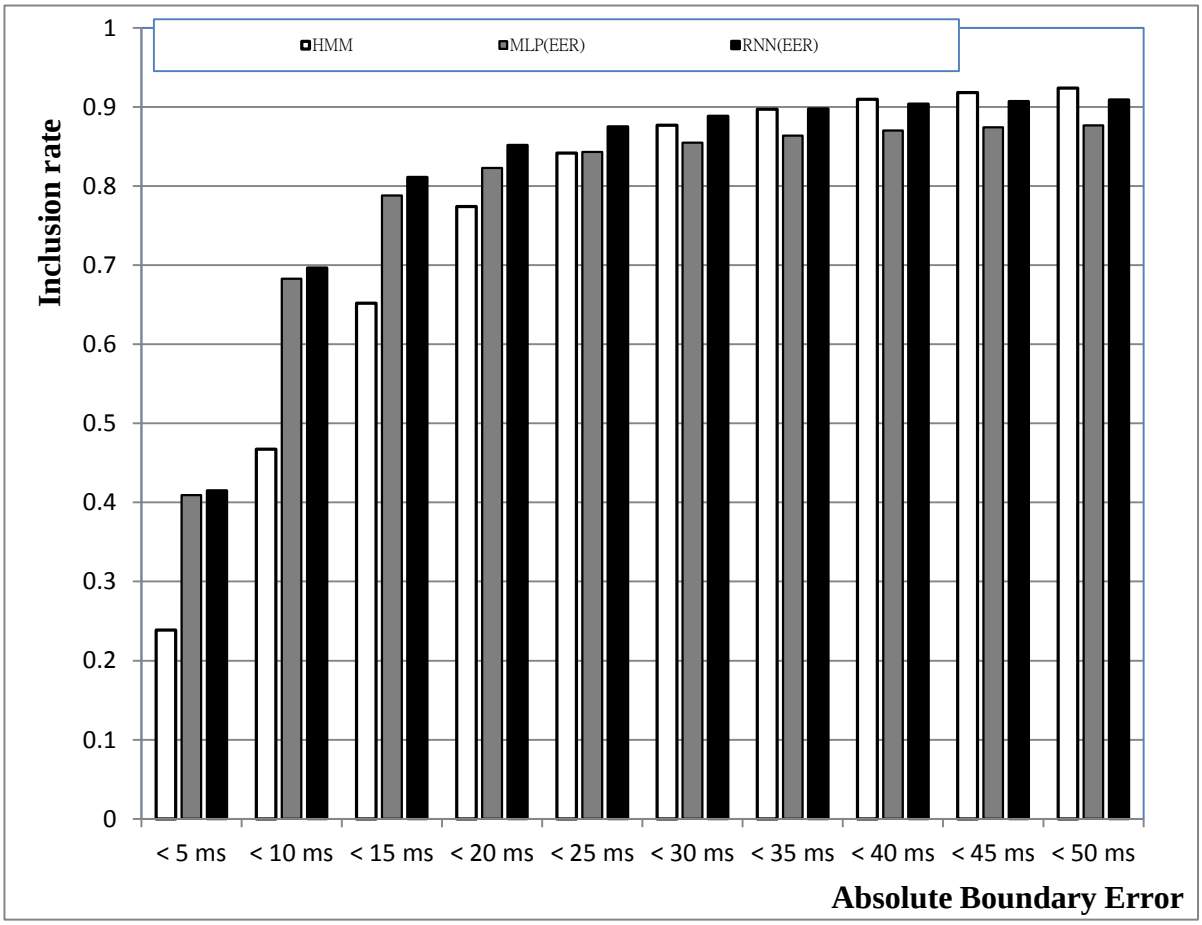


圖 3.7：音素端點偵測器實驗結果與人為標記之絕對偏差值直方圖

由先前所述，有些音素的連音現象其頻譜為平滑的變化，使得這些音素端點非常地難偵測，以下本計畫也列舉出觀察語音波形時較難辨別音素轉換對其音素端點偵測的數據。在表 3.4 及 3.5 中，我們統計了 EER 的的情況下，不同發音方式的音素連接時所偵測到音素端點之絕對誤差值及均方誤差值。

表 3.4：TIMIT 語料庫中發音方法與前後音素不同發音方法之其偵測端點 MAE 統計資料
(The two values in table are MAEs of RNN and HMM in ms, * means sample counts less than 100.)

	Affricate	Fricative	Stop	Glide	Vowel	Nasal	Silence
Affricate	-	6.4/6.5*	10.1/6.9*	7.3/10.0	6.8/13.7	4.9/15.3*	6.1/12.8
Fricative	2.3/17.0	7.2/7.0	13.6/13.1*	9.5/14.9	7.9/13.3	7.1/12.5	6.5/11.7
Stop	-	6.1/7.3	12.4/12.0*	11.2/15.0	7.5/13.1	7.6/9.6	7.1/14.4
Glide	-	7.0/9.5	10.4/12.8	11.0/21.2	7.9/13.6	6.4/11.2	6.3/12.7
Vowel	-	6.3/9.8	7.9/11.8	9.9/15.9	8.8/17.6	6.8/11.5	6.9/13.6
Nasal	7.6/11.3*	6.2/8.2	11.1/13.2	11.6/15.3	7.2/13.3	5.6/11.2*	6.9/12.1
Silence	6.3/12.5	6.0/7.5	7.3/8.2	11.7/14.1	7.4/12.1	5.2/9.9	7.0/18.9

表 3.5：TIMIT 語料庫中發音方法與前後音素不同發音方法之其偵測端點 RMSE 統計資料
(The two values in table are RMSEs of RNN and HMM in ms, * means sample counts less than 100.)

	Affricate	Fricative	Stop	Glide	Vowel	Nasal	Silence
Affricate	-	7.4/8.0*	13.1/11.3*	8.8/13.1*	10.8/18.2	6.4/17.7*	7.8/15.6
Fricative	2.3/17.0	8.6/9.0	17.4/16.6*	13.8/20.3	12.4/17.9	11.2/18.4	7.9/14.2
Stop	-	8.1/9.5	17.2/16.2*	16.5/20.1	12.1/17.3	10.8/12.5	8.3/18.2
Glide	-	14.0/14.3	15.0/19.4	16.8/29.2	12.5/17.6	9.0/13.6	9.9/17.3
Vowel	-	10.2/13.6	12.5/17.2	14.5/21.3	14.6/24.6	10.4/15.3	10.0/17.9
Nasal	10.5/15.5*	9.8/10.9	15.0/18.5	16.8/20.4	11.9/17.6	8.6/12.7*	8.7/16.4
Silence	10.2/16.2	7.9/10.1	9.4/12.3	15.8/20.0	11.8/16.3	7.1/12.7	9.4/23.0

接著我們對所偵測之音素端點做定性分析：

- 偵測漏失率分析

本計畫所提出之方法為利用取樣點式參數的萃取，依照音素變化時語音信號在頻譜之間的變化程度來進行音素邊界端點偵測，若相鄰音素之頻譜變化的程度越大，則越可能被偵測為音素的邊界。可以看到相鄰音素是相同與不同的發音方式對照下，實驗結果觀察發現不同的發音方式相較於相同發音方式其大部分之偵測漏失率都有大幅降低的現象。因此以下將針對偵測漏失率較高的摩擦音、鼻音、母音以及靜音等數種發音方法來提出討論。

(1) 前後相鄰音素為摩擦音

摩擦音發音時會由於發音器官彼此靠攏而形成狹窄的氣流通道，使得氣流通過通道時造成摩擦產生出聲音，如發出 s 的音必須讓氣流通過閉合牙齒之間的縫隙來產生。摩擦音在頻譜上的分佈多集中在高頻部分。圖 3.8 舉出前後音素為（/k/、/s/）皆屬於摩擦音的分類，由音素端點偵測器輸出概似度的觀察中，在（/k/、/s/）音素的區間中所有的音素候選端點之概似度皆非常地低，亦即偵測器不認為這些候選端點是音素的端點。

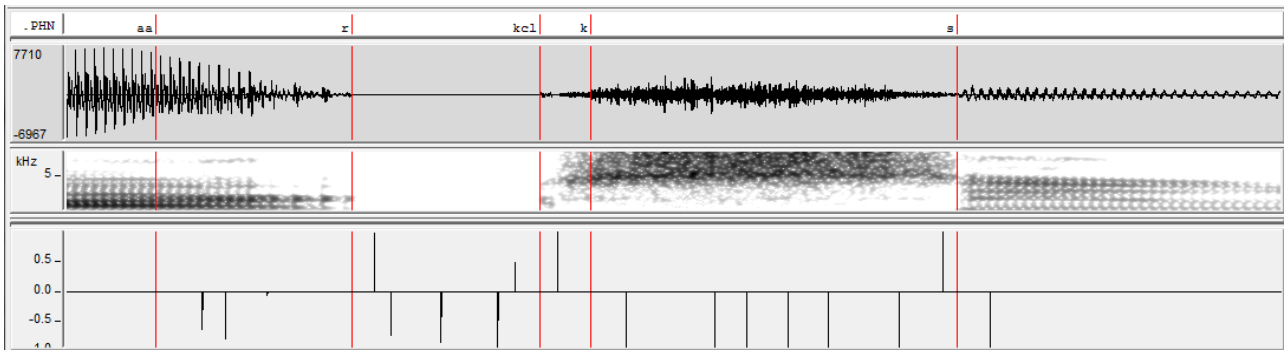


圖 3.8：音素端點偵測前後音素為摩擦音之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

(2) 前後相鄰音素為鼻音

鼻音發音時口腔中的軟顎下垂，造成氣流無法通往口腔而轉往通過鼻腔發音，如發出/m/的音時，須雙唇緊閉來讓氣流通過鼻腔產生，也因為如此使得鼻音在頻譜上的分佈多集中在聲譜圖之低頻部分。圖 3.9 舉出前後音素為（/m/、/n/）皆屬於鼻音的分類，在（/m/、/n/）音素的區間中，相鄰音素頻譜間平滑的變化造成音素候選端點的個數較少；僅觀察語音波形也亦難標記正確的音素端點位置，這也就是前後音素為鼻音時偵測漏失率較高的原因之一。即便音素端點偵測器輸出概似度藉由調整臨限值後，增加偵測出候選端點之個數，其音素候選端點仍與人為標記位置有一段誤差存在。

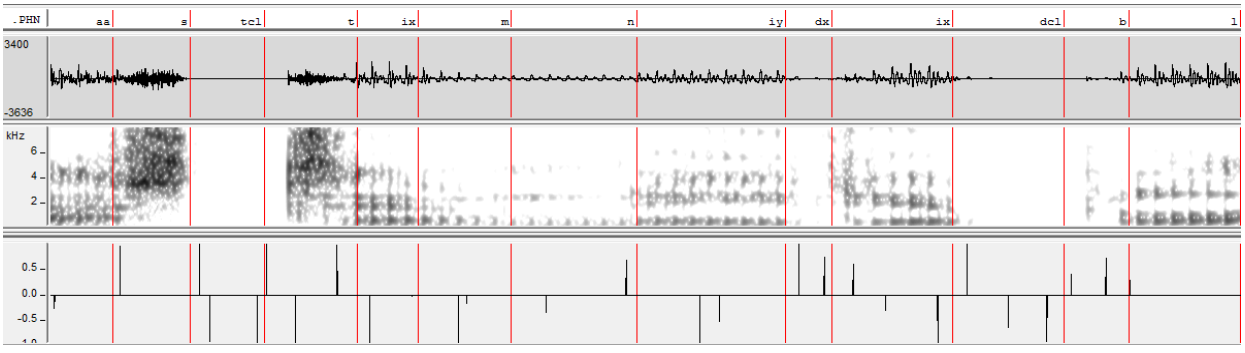


圖 3.9：音素端點偵測前後音素為鼻音之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

(3) 前後相鄰音素為母音

母音是氣流由肺通過聲帶時，使聲帶產生週期性的震動且讓氣流不受阻礙地通過口腔通道，再以舌頭或是雙唇的調整而發出聲音。不同口腔通道、舌頭位置等所發出的母音在頻譜上亦有不同的分佈，但在時域上的語音波形中皆可明顯觀察出週期性的訊號。圖 3.10 舉出前後音素為（/er/、/axr/）皆屬於母音的分類，相鄰音素頻譜間平滑的變化產生的音素候選端點個數不多，就算偵測器輸出概似度藉由調整臨限值後，增加偵測出候選端點之個數，其音素候選端點仍與人為標記位置有一段誤差存在；同樣觀察語音波形也亦難標記正確的音素端點位置。

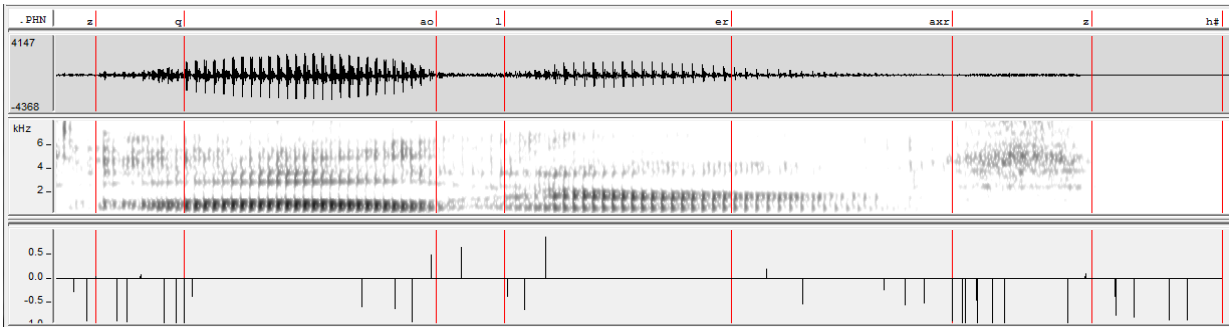


圖 3.10：音素端點偵測前後音素為母音之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

(4) 前後相鄰音素為靜音

靜音通常表示無任何語音信號的區段，但在 TIMIT 語料庫語句內的某一字詞音素與音素間的短停頓則以/*epi*/表示。圖 3.11 舉出前後音素為（/tcl/-/h#/）皆屬於靜音的分類，同前後音素為鼻音的情形相似，僅觀察語音波形也亦難標記正確的音素端點位置，為造成前後音素為靜音時偵測漏失率較高的原因。由音素端點偵測器輸出概似度的觀察中，在（/tcl/-/h#/）音素的區間中音素候選端點之概似度同樣非常地低，顯示出偵測器偵測不出這些候選端點是音素的端點，藉由調整臨限值也亦難偵測出音素端點。

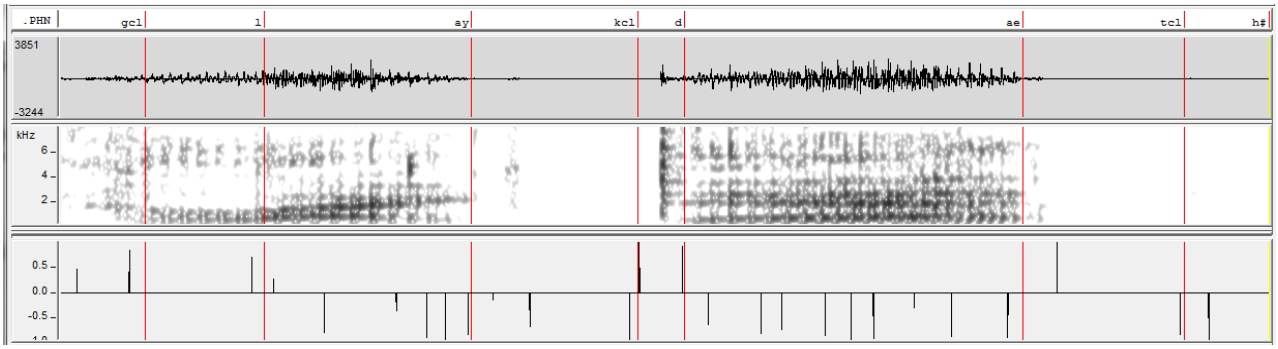


圖 3.11：音素端點偵測前後音素為靜音之範例，由上至下分別表示音素層級之人為時間標記的文字轉寫、語音信號、聲譜圖、音素端點偵測器輸出之概似度

- **誤報率分析**

由先前所述前後音素為不同發音方法的偵測漏失率較低，但誤報率與偵測漏失率是成反

比的，亦即在不同的發音方式的轉換期間語音信號頻譜的劇烈變化容易產生誤報的情形，然而本計畫以取樣點式聲學參數挑選音素候選端點的方式與傳統音框解析度對照之下，在此情況卻是更加容易產生較多的音素候選端點，可能造成誤報率增高的情形。故以下分析在前後音素為不同發音方法時誤報率的差異並作討論。塞擦音、摩擦音以及母音等發音方式之邊界。

第四章　使用取樣點式聲學參數之語音類音素端點自動分段

4.1 語料庫簡介

本計畫中將對國語及不同語言之語料庫進行自動語音分段的實驗， 首先將先介紹計畫中所使用之國語及方言語料庫。

4.1.1 國語 TCC-300 語料庫簡介

本計畫中使用 TCC-300 麥克風語音資料庫是由國立交通大學、國立成功大學、國立台灣大學所共同錄製，中華民國計算語言學學會所發行，此語料庫屬於麥克風朗讀語音，主要目的是為提供語音辨認研究，檔案統計資料如表 4.1 所示。台灣大學語料庫主要包含詞以及短句，文字經過設計，考慮音節與其相連出現之機率，共 100 人，每人錄製一句而成；成功大學及交通大學為長文語料，其語句內容由中研院提供之 500 萬詞詞類標示語料庫中選取，每篇文章包含數百個字，再切割成 3 至 4 段，每段至多 231 字，分別各 100 人，每人錄製一句朗讀來錄製，且每人所朗讀之文章皆不相同。每個學校之語句取樣頻率皆為 16000 赫茲（Hertz），取樣位元數為 16 位元。音檔檔頭為 4096 位元組（byte），副檔名為*.vat。

表 4.1：TCC-300 語料庫檔案統計資料

學校	語音檔案(*.vat)	文字檔案(*.tab)	群集(Group)
交通大學	1238	1238	5
成功大學	1170	1170	5
台灣大學	6509	6509	1

屬於聲調語言之國語音節結構如下圖所示可將音節分為聲母和韻母，韻母可再細分為介音與韻腳，而韻腳包含主要元音和韻尾，而本計畫使用之 TCC-300 國語語料庫是以類音素單元做為自動語音標記的基本語音單元，類音素即是將國語結構分為聲母、韻母（但韻母不包含鼻音韻尾）以及鼻音韻尾等三個部份以依照語音之特性簡化結構。

在 TCC-300 語音資料庫之語料選取方面，我們使用交通大學與成功大學所錄製的長文語料，並隨機選取六分之五的部份當作訓練語料，其它部分為測試語料。本計畫提出自動標記音素位置之方法是以兩個階段（two-stage）來達成自動標音的目標，故需要有一個初始位

置來訓練一個自動端點標示偵測器，以進行第二階段更進一步地修正。由於 TCC-300 語音資料庫沒有人工標記的音素切割位置，利用 HTK（Hidden Markov Toolkit）使用 SAT（speaker adaptation transform，feature MLLR）及 SA（speaker adaptation，MLLR）技術訓練 HMM 類音素模型，獲得較佳的 HMM 模型後進行強迫對齊（force alignment）之自動標示結果，作為 TCC-300 語料庫之類音素初始切割位置，以提供本計畫使用。

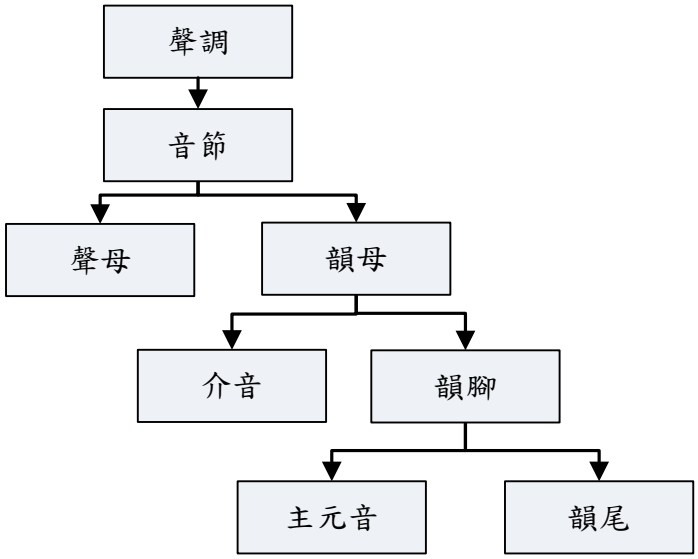


圖 4.1：國語音節結構圖

4.1.2 國語 Treebank 語料庫簡介

Treebank 語料庫包含 425 個語句且含有 56237 個音節，由一個專業的女性播音員所錄製。此語料庫屬於麥克風朗讀語音，主要目的是為提供語音韻律標記與建立韻律模型之研究。語句取樣頻率皆為 16000 赫茲（Hertz），取樣位元數為 16 位元，副檔名為*.pcm。

在文字轉寫方面，因 Treebank 語料庫內含人為時間標記之音節與聲/韻母層級文字轉寫，本計畫以此兩種層級之文字轉寫作為實驗結果之標準答案以評量實驗結果之效能。另外，藉由 HTK toolkit 訓練音節以及聲/韻母（initial/final）語音單元之 HMM 模型，對語料庫進行強迫對齊，獲得初始自動分段位置用於實驗使用與測試。選擇梅爾倒頻譜係數作為語音聲學參數，參數設定為 38 維，其中包含 12 階的梅爾倒頻譜係數與能量之對數值（log energy）及其一階差量與二階差量並扣除原本的能量對數值總計 38 維，音框長度（frame length）設為 32 毫秒，音框平移（frame shift）設為 5 毫秒。

Treebank 語料庫在訓練及測試語料的選擇上，扣除語句中含有英文的 4 個語句，剩下 421 句以 9：1 的比例隨機選取，得訓練語料為 379 句和測試語料為 42 句。

4.1.3 客語語料庫簡介

本計畫為使用四縣客家話語料庫，文章出處為龔萬灶老師所撰寫的「阿啾箭个故鄉」，音檔取樣頻率為皆以 20k 赫茲及取樣位元數為 16 位元之單聲道錄製而成，副檔名為*.pcm 格式。語料庫之語者為龔老師共錄製語音檔案 639 個，包含 42 篇文章共有 63158 個音節。語音檔是由發音人在普通房間依照文稿唸出，屬於朗讀式語音並依照錄製之日期、文章編號來命名。

在文字轉寫方面，因客語音節結構與國語相同，在此本計畫以聲/韻母作為語料庫的文字轉寫之基本單元，而客語語料庫無人為時間標示之音素端點位置可提供正確的端點進行訓練。藉由 HTK 訓練聲/韻母之 HMM 模型，對語料庫進行強迫對齊以獲得四縣客語文字轉寫之初始自動分段位置。使用梅爾倒頻譜係數做為聲學參數，參數設定為 38 維，其中包含 12 階的梅爾倒頻譜係數與能量之對數值及其一階差量與二階差量並扣除原本的能量對數值總計 38 維，音框長度設為 32 毫秒，音框平移設為 5 毫秒。

客語語料庫在訓練及測試語料的選擇上，同樣以 9：1 的比例隨機選取，訓練語料為 587 句和測試語料為 73 句。

表 4.2：客語語音發音方法的分類表。

發音方法(Manner)	發音方法對應之音素					
爆破音 Stop	*b*	*p*	*d*	*t*	*g*	*k*
鼻音 Nasal	*m*	*n*	*ng*			
摩擦音 Fricative	*f*	*s*	*h*	*v*		
塞擦音 Affricate	*z*	*c*				
流音 Liquid	*l*					
韻母音 Vowel	others					

4.2 類音素標示位置起始值

由於上述的語料庫均無人工的正確標記語音位置，而我們在計畫中想使用監督式的學習方式來製作類音素端點標示工作，所以如何使用自動的方法來獲得一個無人工的正確標記語音位置語料庫之可靠的類音素標示位置起始值是一個重要的課題。

過去的研究常以音框式之 HMM 架構為基礎來獲得之語音標記位置作為正確標示，此方法雖然可以達成自動語音分段的目的，但最終還是需要人工進一步修正，離正確語音的標記位置之間仍有許多改進的空間。以 2.1 節所提出之取樣式聲學參數之特性，對國語語料庫 TCC-300 進行自動分段的初步實驗，其步驟如下：

首先，利用 SAT(Speaker Adaptation Transform，feature MLLR)及 SA(Speaker Adaptation，

MLLR）後的出語者調適 HMM 模型來做 TCC-300 的類音素單元之初始自動語音分段位置，接著利用此初始位置依照發音方法的不同做分類，如表 4.3。並由初始位置當作參考位置再利用取樣式聲學參數的特性來調整音素端點之標記位置。以下比較 HMM 之初始位置及以取樣式聲學參數特性修正後之語音分段位置。

表 4.3：國語語音發音方法的分類表。

發音方法(Manner)	發音方法對應之音素					
爆破音 Stop	*b*	*p*	*d*	*t*	*g*	*k*
鼻音 Nasal	*m*	*n*	(*n_n*)	(*ng*)		
摩擦音 Fricative	*f*	*s*	*x*	*h*	*sh*	
塞擦音 Affricate	*q*	*j*	*c*	*z*	*zh*	*ch*
流音 Liquid	*l*	*r*				
韻母音 Vowel	others					

先前在觀察 HMM 自動語音分段位置的準確度時，發現短停頓常會有無法標記出來或是標記位置錯誤之情形，而使得某些音素之平均音長有過長的現象，如塞擦音與爆破音等。在此本計畫使用信號波封與各頻段之信號波封來判斷語音段是否為短停頓的狀態。由圖 4.2 可以觀察到短停頓中各個頻段之信號波封與其它有語音信號的地方相比其數值幾乎非常地低且根據語音屬性不同而有不同的頻譜分佈情形。在此，簡單以信號波封與各頻段之信號波封來標記短停頓的端點。短停頓標記修正之演算法如下：

(1) 前端點：在原端點位置之前後 30 毫秒的範圍內，判斷語音波形之波封是否**小於**波封之臨限值而得到一個交集點，再經由交集點附近距離 10 毫秒內來判斷各個頻段之信號波封是否**小於**頻段波封之臨限值的條件作聯集來決定是否有短停頓的狀態。

(2) 後端點：在原端點位置之前後 30 毫秒的範圍內，判斷語音波形之波封是否**大於**波封之臨限值而得到一個交集點，再經由交集點附近距離 10 毫秒內來判斷各個頻段之信號波封是否**大於**頻段波封之臨限值的條件作聯集來決定是否有短停頓的狀態。

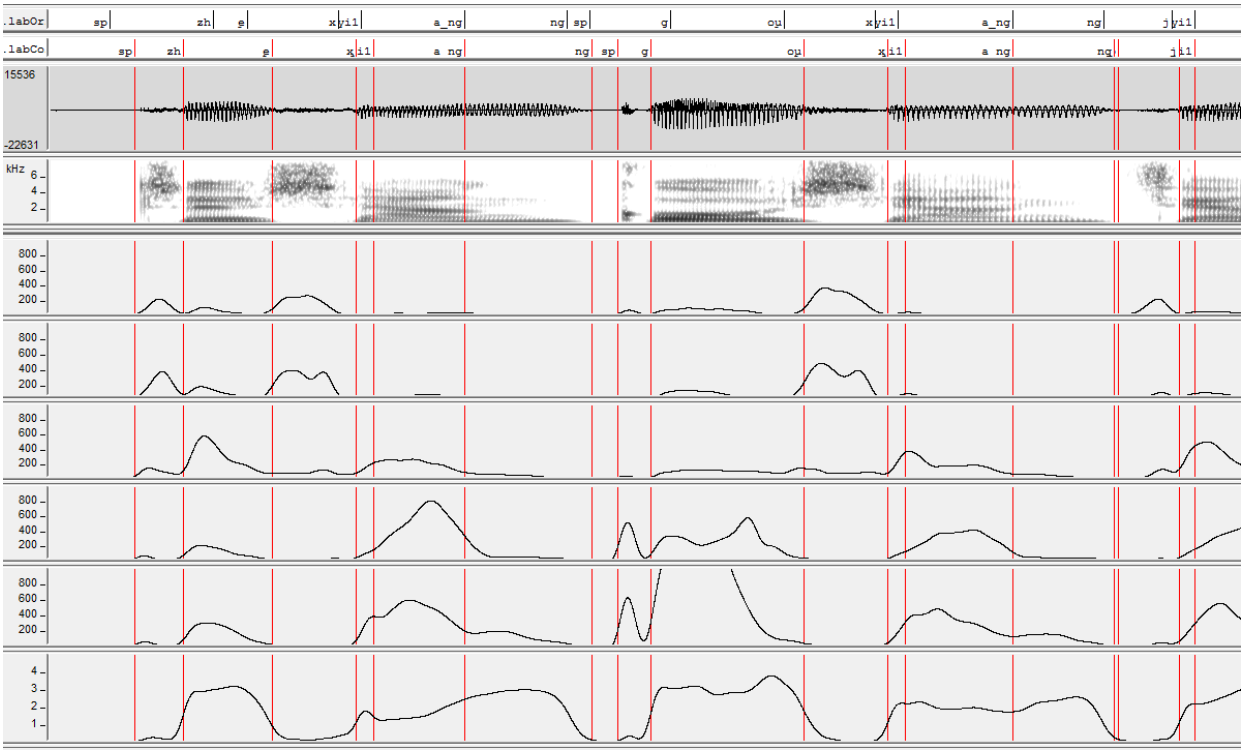


圖 4.2：國語語句端點位置自動調整(短停頓)演算法則之範例，最上方兩列標音位置分別表示是 HMM 自動語音分段及修正後之語音標記位置；接著由上至下的圖形分別表示語音波形、聲譜圖、第六個至第一個頻段的信號波封

接下來觀察摩擦音、塞擦音等發音方法之音素，其在於頻譜中與相鄰母音與短停頓有極大的頻譜差異。在此，使用頻譜 KL 距離、頻譜熵及頻譜熵的上升率來調整音素的端點。圖 4.3 所示，由摩擦音與塞擦音頻譜中可觀察到頻譜 KL 距離在母音轉換至摩擦音、塞擦音之間有較高的峰值，且摩擦音、塞擦音相鄰母音的端點，其頻譜熵值上升與下降速度很快，分別在頻譜熵的上升率中造成極大、極小的峰值。頻譜熵的上升率之峰值位置與人所期望的正確端點位置差距不遠，由先前研究可以了解頻譜熵、 頻譜 KL 距離等已知在音框式量測信號變化量方法中是非常有用的聲學參數，同樣在取樣式聲學參數量測信號變化量的效果一樣明顯，且語音之分段位置更精準。

摩擦音、塞擦音程式修正演算法如下式：

(1) 後端點：找到此一區段頻譜熵上升率的相對極小值，在小範圍的搜尋 KL distance 相對極大值。

(2) 前端點：利用後端點的位置當做參考位置，判斷前面是否有短停頓，有則利用短停頓的方式偵測前端點，若無短停頓則搜尋一段範圍找到此一區段頻譜熵上升率的相對極大值。

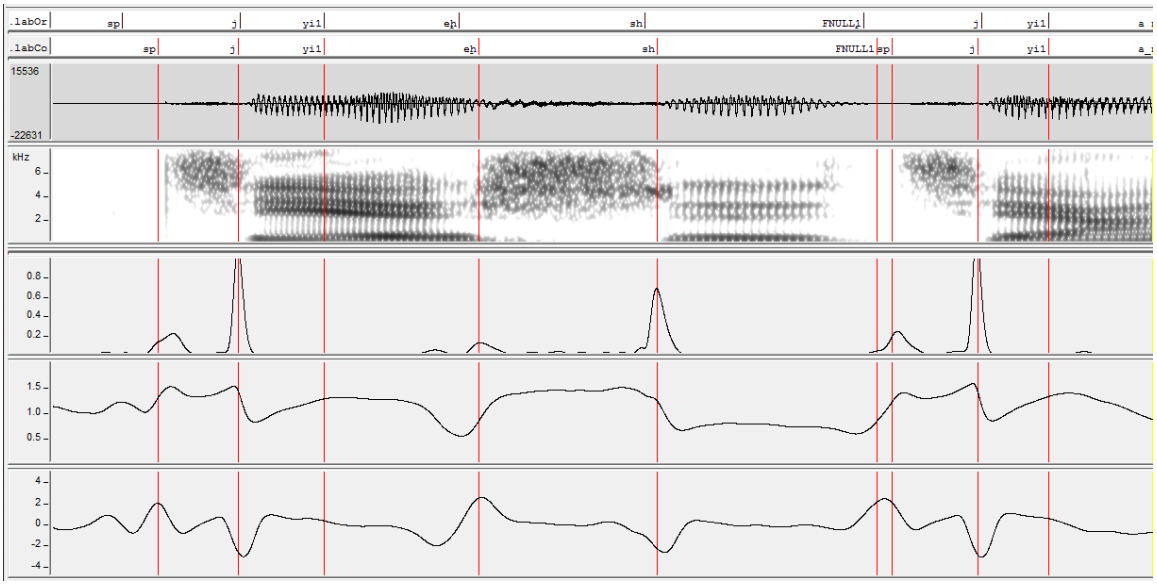


圖 4.3：國語語句端點位置自動調整(摩擦音、塞擦音)演算法則之範例，最上方兩列標音位置分別表示是 HMM 自動語音分段及修正後之語音標記位置；接著由上至下的圖形分別表示語音波形、聲譜圖、頻譜 KL 距離、頻譜熵、頻譜熵上升率

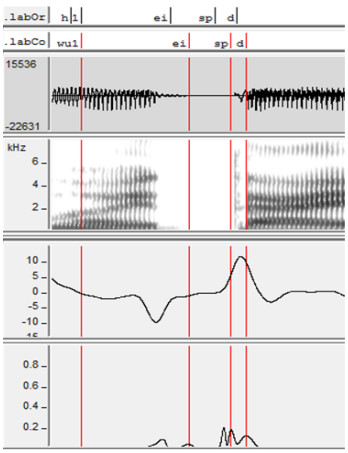
爆破音切割位置的修正時，由波形與頻譜觀察中發現通常在爆破音開始的時候會有短停頓出現，接著波封會有急遽上升的現象，故本計畫使用波封之上升率來描述其現象。如圖 4.4 中(a)、(b)小圖所示，在爆破音結束的地方，也是音素轉換的端點。

爆破音程式修正演算法如下式：

(1) 後端點：找到此一區段波封上升率的相對極大值，並在該極大值之位置找到頻譜 KL 距離的相對極大值。

(2) 前端點：利用後端點的位置當做參考位置，判斷前面是否有短停頓，有則利用短停頓的方式偵測前端點，若無短停頓則搜尋此一區段之頻譜 KL 距離的相對極大值。

(a)



(b)

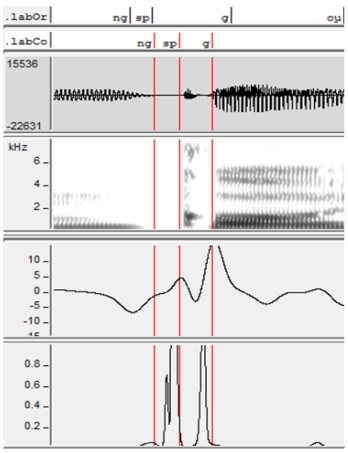


圖 4.4：國語語句端點位置自動調整(爆破音)演算法則之範例：(a) /*d*/ 和(b) /*g*/ 音素最上方兩列標音位置分別表示是 HMM 自動語音分段及修正後之語音標記位置；接著由上至下的圖形分別表示語音波形、聲譜圖、語音波封上升率、頻譜 KL 距離

另外，鼻音部分可由其語音信號之頻譜分佈多集中在 0.0 – 0.4 kHz 與 0.8 – 1.5 kHz 的低頻頻段的現象，且與相鄰的音素皆有頻譜上的差異，在此我們也使用頻譜 KL 距離來判斷。鼻音部分程式修正演算法如下式：

(1) 後端點：由原端點位置搜尋頻譜 KL 距離大於臨限值的位置。

(2) 前端點：利用後端點的位置當做參考位置，判斷前面是否有短停頓，有則利用短停頓的方式偵測前端點，若無短停頓則搜尋此一區段之頻譜 KL 距離的相對極大值。

最後，母音端點的偵測是利用相鄰母音、子音及短停頓之端點位置，當作母音的端點位置。由實驗觀察 3.1 節所述之聲學參數特性用於自動分段位置的準確度，並與原本 HMM 初始語音分段位置作為比較對象，以下列舉 2 個實驗結果之範例，圖 4.5 與圖 4.6。首先由圖 4.5 與 4.6 中，將實驗修正後的語音標記位置對應至語音波形及聲譜圖觀察，實驗結果在音素之端點位置皆能調整到適當的地方。以方形圈圈選處之聲譜圖中，以紅色線條為分界點，其前後兩段之語音信號分佈可明顯看出實驗結果能夠將端點位置近乎正確地標示出來，而其他標記位置之準確度也同樣有好的自動標記效能。另外，有些標記位置是與 HMM 的分段位置為相同標記位置，原因在於進行實驗的過程當中，若不符合自動調整演算法之條件，其標記位置則維持不變。

自動調整端點演算法之實驗結果顯示了使用取樣點式聲學參數之特性確實有助於尋找更佳的端點位置，但演算法所使用之規則是基於聲學參數對應語音信號的觀察與語言學知識相互組合而成。然而語音信號的變化並非有一定的規則可循，故本計畫將利用類神經網路之特性將各聲學參數之特性作統計分析的彙整，來找出最佳音素端點位置。

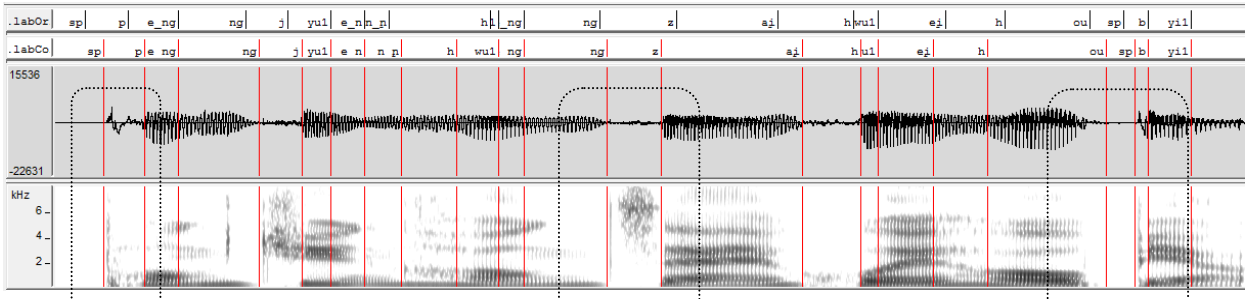


圖 4.5：自動調整國語語句端點位置實驗結果之範例一，最上方兩列標音位置分別表示 HMM 自動語音分段及修正後之語音標記位置、語音波形、聲譜圖

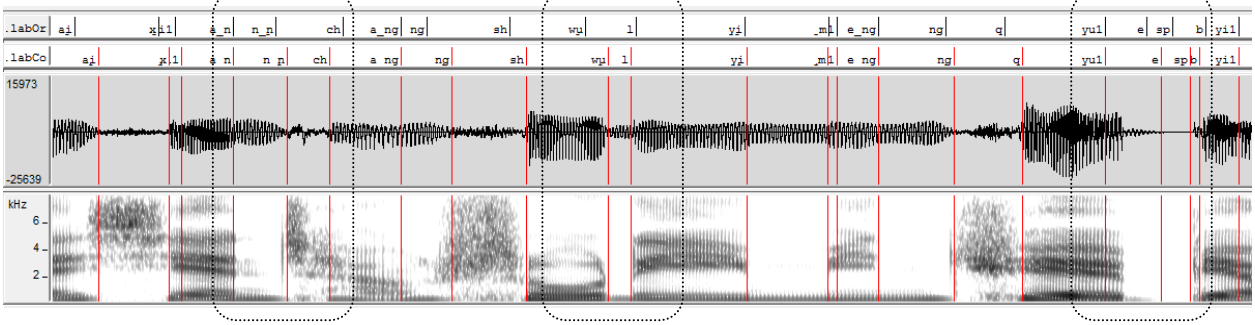


圖 4.6：自動調整國語語句端點位置實驗結果之範例二，最上方兩列標音位置分別表示 HMM 自動語音分段及修正後之語音標記位置、語音波形、聲譜圖

本計畫所建構之自動語音分段系統是分為兩階段式（Two-stage）的端點位置修正（boundary refinement）。第一階段以 MFCC 聲學參數利用 HMM 模型進行強迫對齊而得到初始的語音分段位置；第二階段由本計畫提出之取樣點式聲學參數經多層感知器對不同語音單元分類訓練端點偵測器，並依此架構對第一階段所得到之初始語音分段位置做更細部的調整，最後系統輸出對應於語音單元之文字轉寫的自動語音分段位置。圖 4.7 展示了自動語音分段系統之流程圖，其主要與音素端點偵測器架構的差別是在於目標函數的定義。自動語音分段系統之模型描述了語言之音節結構對應至語音分段之關聯性。

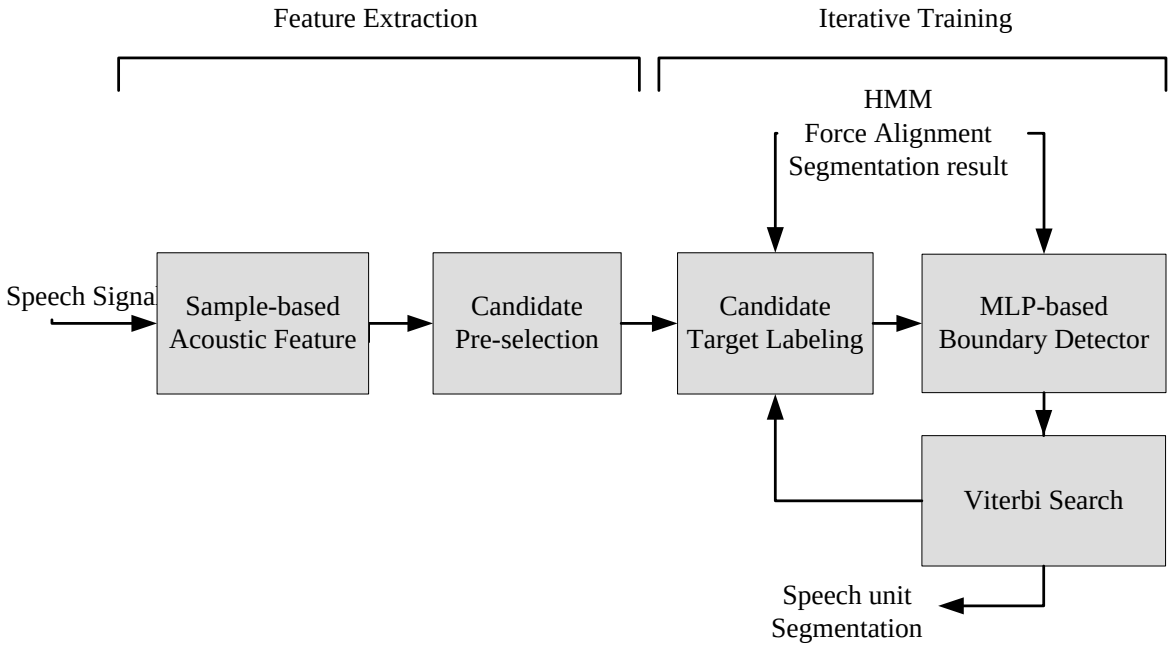


圖 4.7：使用多層感知器架構之自動語音分段系統流程圖

然而，需要做自動語音分段處理的文字轉寫必須根據基本語音單元並依照音節結構來訂定目標函數的種類，以提供端點偵測器的學習。藉由任務的不同來選擇適當的語音單元來進行處理，例如語音合成系統所需要的單元約在聲母/韻母甚至是音節的語音單元；語音辨識系統則可能需要小至音素等語音單元。同屬聲調語言之國、客語的音節結構，在本計畫中選擇處理的語音單元為客語語料庫為聲/韻母的語音單元，國語語料庫為類音素以及音節單元。以下將說明選擇不同基本語音之單元其目標函數之訂定方式：

- 音節層級

以音節結構之音節層級來訂定語音信號所對應的兩個類別（class），分別為靜音（S）與音節（V），依照不同類別彼此之間的轉移狀態，定義五種目標函數分別是 IS、SV、IV、VS、VV 等轉移狀態，如圖 4.8 表示。每個由抽取聲學參數過程中所得到的候選端點皆須要進行目標函數的標記，圖中之 IS 轉移狀態代表該候選端點仍為靜音狀態，SV 轉移狀態表示該候選端點是由靜音狀態轉換至音節狀態，依此類推...。其中需要特別注意的是圖中 VV 的轉移

狀態為表示略過靜音至下一個音節的音節端點。聲調語言中每個音節與音節之間靜音的存在可有可無，為描述此種情形本計畫加入 VV 轉移狀態來模擬音節之間無靜音的現象。

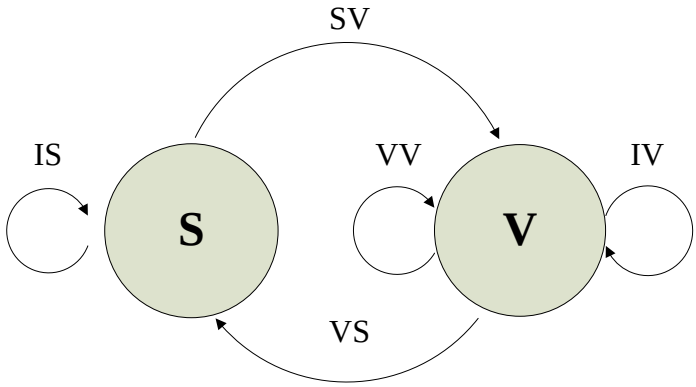


圖 4.8：音節層級目標函數之轉移狀態圖

- 聲/韻母層級

以音節結構之聲/韻母層級來訂定語音信號所對應的三個類別，分別為靜音（S）、聲母（C）和韻母（V），依照不同類別彼此之間的轉移狀態，定義七種目標函數分別是 IS、SC、IC、CV、IV、VS、VC 等轉移狀態，如圖 4.9 表示。圖中之 IS 轉移狀態代表該候選端點仍為靜音狀態，SC 轉移狀態表示該候選端點是由靜音狀態轉換至聲母狀態，同樣地依此類推…。另外，圖中 VC 的轉移狀態為模擬音節之間無靜音的現象，其代表由韻母與下一個聲母轉移狀態的端點。

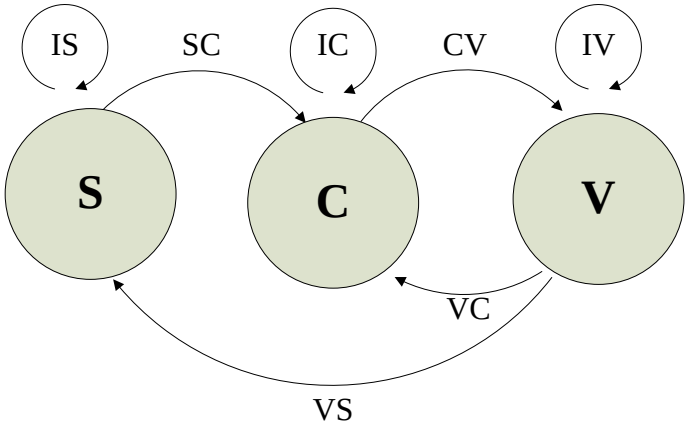


圖 4.9：聲/韻母層級目標函數之轉移狀態圖

- 類音素層級

以音節結構之類音素層級來訂定語音信號所對應的四個類別，分別為靜音(S)、聲母(C)、韻母（V）與鼻音韻尾（N），依照不同類別彼此之間的轉移狀態，定義九種目標函數分別是 IS、SC、IC、CV、IV、VN、IN、VS、VC 等轉移狀態，如圖 4.10 表示。另外，圖中為簡化目標函數之個數，本計畫將鼻音韻尾至靜音與韻母至靜音的轉移狀態定義為相同的目標函數

（VS）；另外，模擬音節之間無靜音的現象中，本計畫亦將鼻音韻尾至聲母與韻母至聲母的轉移狀態定義為相同的目標函數（VC）。

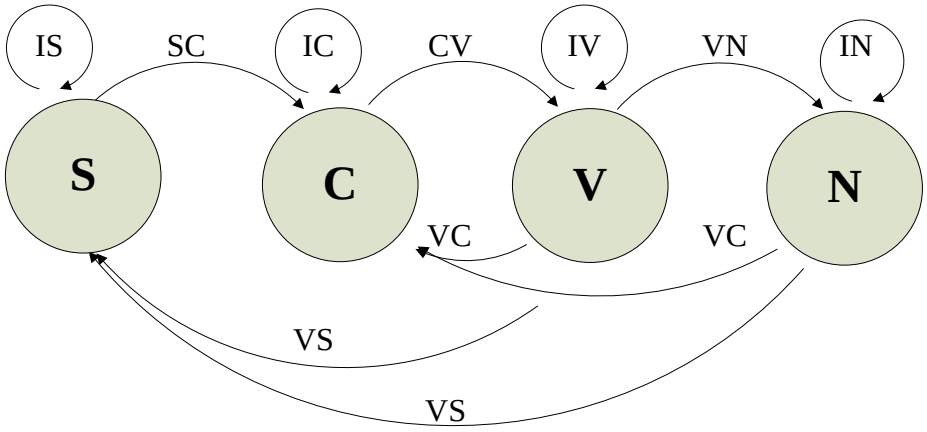


圖 4.10：類音素層級目標函數之轉移狀態圖

由上述不同層級之目標函數轉移狀態的訂定，訓練不同文字轉寫之基本語音單元使用的音素端點偵測器來達到自動語音分段的目的。

4.3 TCC300 語料庫實驗結果分析

實驗結果以音框式抽取參數的 HMM 架構，作強迫切割所獲得的類音素層級初始自動分段位置來比較，並觀察本計畫自動分段位置之精準度是否有進一步地提升。由第四章所述，在得到對應類音素層級之音素端點偵測器後，將 HMM 的類音素自動分段結果以端點偵測器所產生的概似度經正規化後作為分數，進行維特比搜尋並限制搜尋範圍在初始位置前後 100 毫秒之內，最後得到本計畫之類音素層級自動語音分段結果。

首先，以下列舉兩個語音波形比較音素端點與 HMM 的自動分段位置。由下列圖 4.11、4.12 之中，可由方圈之圈選處的音素端點位置觀察到，無論是音節與音節之間的短停頓或是聲母與韻母之間的端點位置都非常準確，尤其是母音和塞擦音、摩擦音之間的邊界端點與 HMM 之分段位置相比確實精準許多，而在聲譜上觀察這些端點位置可看出頻譜分佈差異極大，亦是正確的端點位置。圖 4.12 所示之方圈圈選處，我們亦可發現在母音轉變至鼻音韻尾的情形，其音素端點位置之準確度仍能保持良好的水準；而在爆破音前的短停頓亦能調整至適當的端點位置。由上述實驗結果在語音波形的觀察下，顯示了取樣點式聲學參數對 HMM 之自動分段位置做修正後，其自動分段之效能確有提升。

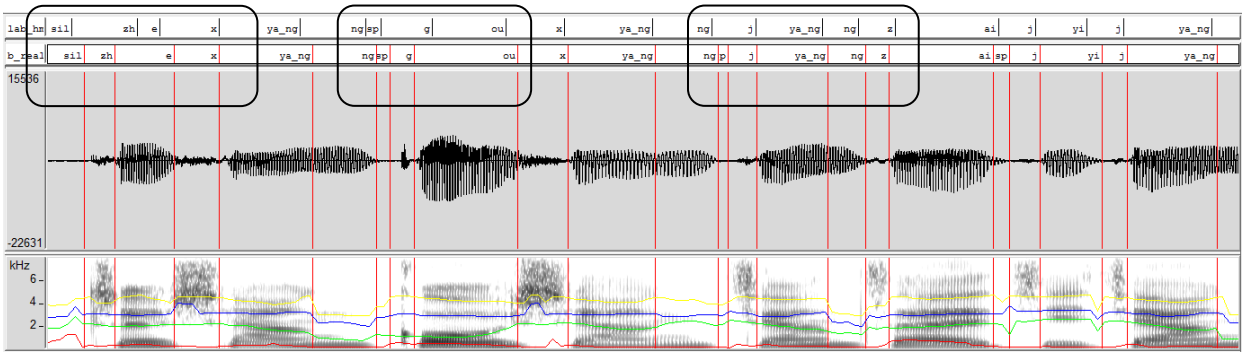


圖 4.11：國語語句自動語音分段之範例一，由上至下的圖形分別表示 HMM 分段位置及音素端點偵測之分段位置、語音波形、聲譜圖

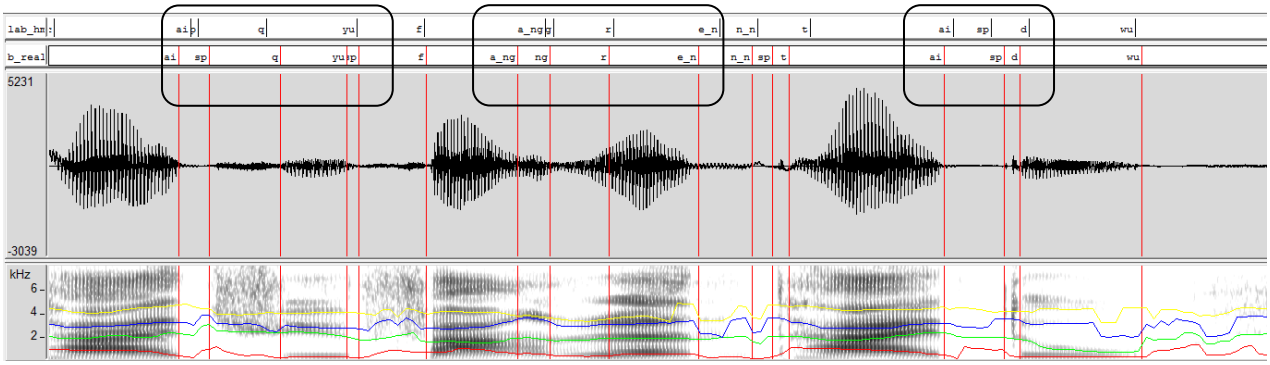


圖 4.12：國語語句自動語音分段之範例二，由上至下的圖形分別表示 HMM 分段位置及音素端點偵測之分段位置、語音波形、聲譜圖

接著，分別在成功大學與交通大學各隨機選取 7 句，共 1698 個音節，作人為標記的標準答案。統計 HMM 自動分段位置和實驗結果對人為標記的端點位置的誤差並以不同絕對偏差值之包含比率來表示，如圖 4.13。圖中以每 5 毫秒為一單位，本計畫所提出之方法在 15 毫秒內之邊界包含比率中，可明顯看出與 HMM 自動分段結果的差距，在 5 毫秒內即可達到 46%的包含比率。此數據顯示本研究方法能有效地的改正原本 HMM 的自動分段結果，提升自動語音分段的精確度。在另一方面，隨著與人為標記位置的誤差增大，兩者方法之間的差距慢慢地縮小，在絕對偏差值 30 毫秒的範圍之後仍還有約 10%的邊界誤差極大，以致於無法涵蓋其中，而本研究方法在其範圍之後效能與 HMM 相比甚至較差，其原因歸類於下列：

1. 連音現象

   連音現象易使得兩者實驗方法皆難以判斷端點位置。例如「第(/d/-/e/)一(/yi/)」中(/e/-/yi/)的端點位置，發音方法與口型上的變化都相似而頻譜亦趨於平滑變化，造成端點位置判定上的困難。

2. 語料庫錄音的背景雜訊

   語料庫中不穩定之錄音品質，造成有部分音檔的背景雜訊過大，在取樣點式聲學參數之子頻帶信號波封反映出劇烈變動的情形，因此造成端點位置標記產生偏差。

3. HMM 自動分段結果與人為標記之間誤差過大

   由第四章所述，本研究之自動語音分段方法是基於 HMM 之自動分段結果再使用端點偵

測器所提供之分數進行維特比搜尋。因此，起始分段位置之誤差過大亦難在搜尋空間找到最佳的候選端點，使得端點位置產生偏差。

4. 類音素音節結構與候選端點個數在該音段過少所引起端點位置標記誤差偏大的情形

由於本計畫是將韻母定義為介音以及韻腳除去鼻音韻尾後所組成，但是在韻母音中雙母音中的音素的變化卻是容易造成本研究方法的端點位置標記誤差增大，例如「作（/zuo/）為（/wei/）」，韻母（/wei/）中即可分為介音/wu/、主元音/ei/和韻腹/eh/以及韻尾/yi/，其中在聲譜圖內介音至主元音的變化卻是較（/o/-/wu/）變化明顯。然而候選端點在這些變化較為明顯的地方容易挑選出來，進而使標記位置錯誤。

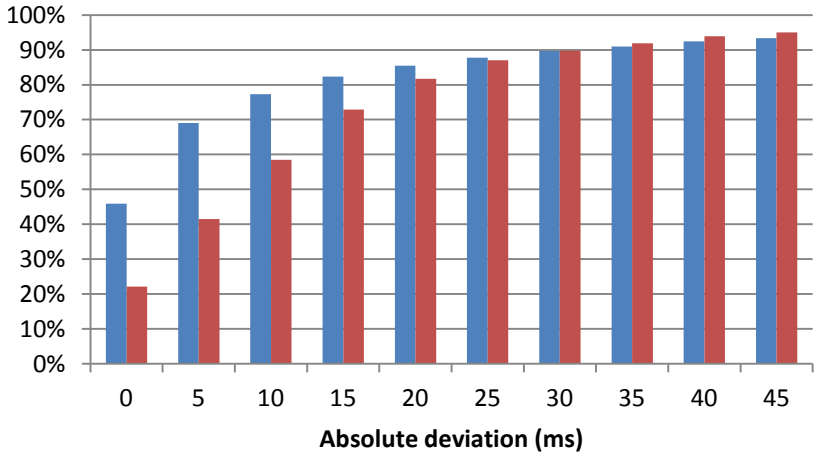


圖 4.13：實驗方法與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖，藍色線(左側)為本計畫所提出之方法，紅色線(右側)為使用 HMM 之初始自動分段位置

然而圖 4.14 為實驗結果所有邊界端點與人為標記位置的統計，以下本計畫依續節音素端點偵測應之誤報率與偵測漏失率的分析結果，將實驗結果依不同發音方法所對應的包含比率做比較，觀察自動語音分段之效能好壞。首先由圖 4.14 中在絕對偏差值為 15 毫秒的範圍內，圖(a)的整體曲線一開始便急遽拉升至近 80%以上的包含比率，但在圖(b)包含比率之整體曲線則是呈現相較緩慢速度的提高。在圖(a)中，由摩擦音與塞擦音的包含比率相較於圖(b)之結果差距逾 40%，代表著本研究方法確實有助於對此類發音方法之邊界端點來提升自動分段的準確度。然而圖(a)、(b)的結果中發音方法為靜音之曲線趨勢差異為最大，其中隱含著在 HMM 的自動分段結果中，短停頓不易標記出來抑或是不夠準確的情況，此一現象亦顯現出本研究方法對於音節間短停頓的修正，有大幅度地改進。

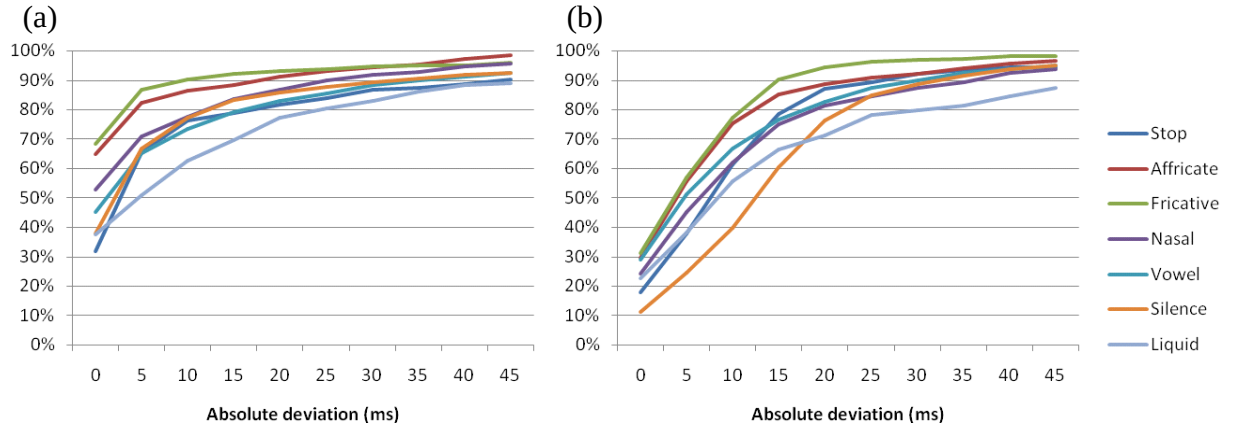


圖 4.14：實驗方法與人為標記位置之誤差以發音方法對應不同絕對偏差值的包含比率直方圖，(a) 本計畫所提出之方法，(b) HMM 之初始自動分段位置

4.4 Treebank 語料庫實驗結果分析

利用所述之自動語音分段的方法，來印證在國語 Treebank 語料庫的效能，並以音節層級和聲/韻母層級訂定目標函數進行自動語音分段，最後得到對國語 Treebank 語料庫兩種不同層級的自動分段實驗結果。

Treebank 語料庫有人為時間標記的資訊，首先針對音節層級之實驗結果統計 HMM 自動分段位置和實驗結果對人為標記的端點位置的誤差在不同絕對偏差值之包含比率，如圖 4.15。在音節層級方面，本計畫所提出之方法相較於 HMM 自動分段結果之準確率也有大幅度地提升。此數據顯示本研究方法能有效地的改正原本 HMM 音節層級的自動分段結果，提升自動語音分段的精確度效能。

另一方面，在實驗結果的分析上同樣發生與 TCC300 語料庫相同的問題，圖 4.15 以絕對偏差值為 30 毫秒做為分界點，可以觀察出分界點左邊的包含比率相較於分界點右邊的上升幅度較大，這種現象表示出在本研究方法確實能將 HMM 自動分段位置調整至更精確，但分界點右邊代表越難找到一個合適的端點位置使得偏差值的增高，造成分界點右邊包含比率上升幅度趨緩的原因。

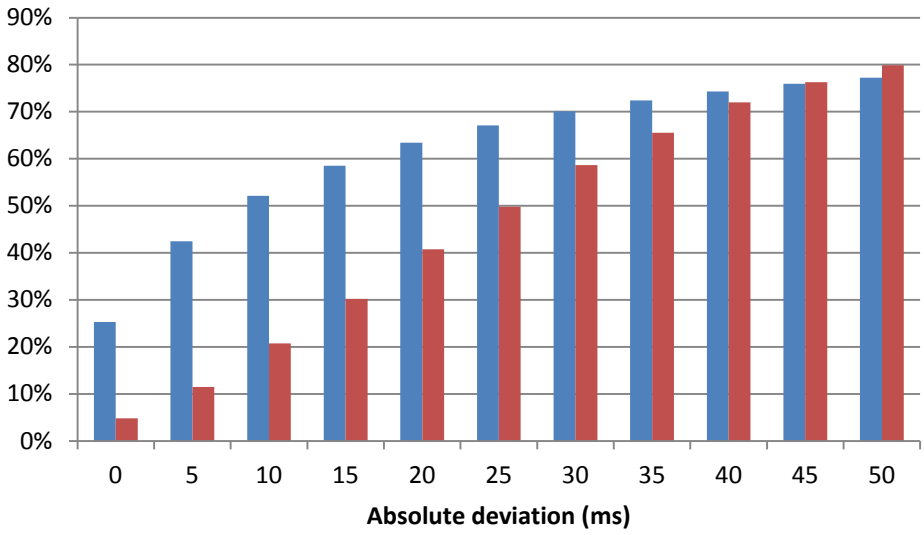


圖 4.15：實驗方法與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖，藍色線(左側)為本計畫所提出之方法，紅色線(右側)為使用 HMM 之初始自動分段位置

本計畫同樣將聲/韻母層級的端點測器以修正 HMM 自動語音分段之實驗結果顯示了與 TCC300 語料庫類音素層級以及 Treebank 語料庫音節層級實驗相似之實驗結果，在此不多作敘述。而將音節層級與聲/韻母層級之實驗結果相比，可以發現左側音節層級和右側聲/韻母層級有一段包含比率的差距，且聲/韻母層級之實驗結果較佳，如下圖 4.16 所示。此圖凸顯出不同音節結構的層級對自動語音分段效能的影響，其原因將在以下做討論。

由於音檔抽取之取樣點式聲學參數以及挑選音素候選端點的過程為相同步驟，則影響效能差異的關鍵即是對候選端點依照不同層級音節架構所標記之目標函數，而目標函數中所代表不同分類之間的轉移狀態即為自動分段所能調整的端點位置。對描述語音單元中邊界端點轉移狀態的定義分別由音素端點偵測只有一種描述音素邊界端點的轉移狀態；音節層級轉移狀態有兩種；聲/韻母層級則有四種；最後類音素層級共有五種描述語音邊界端點轉移狀態的目標函數。然而，轉移狀態的個數也象徵候選端點對描述語音單元邊界端點的分類，與只有一種描述音素邊界端點轉移狀態相比，若邊界端點的類型適當地增多便能顯現各轉移狀態統計特性的差異，減低因輸入候選端點之聲學參數特性相似讓端點偵測器產生混淆的可能性。

語音之結構亦隱含前後轉移狀態之間的順序關聯性，以聲/韻母層級舉例，目前候選端點為聲母的狀態，則下一個候選端點就只能為聲母至韻母的轉移狀態或仍是聲母的狀態而不會跳過結構中的分類。在另一方面，此順序之關聯性也可能造成分段位置之絕對偏差值增大，如類音素層級之鼻音韻尾分類，欲在韻母狀態之後尋找最佳韻母至鼻音韻尾轉移狀態之候選端點，但語音信號卻有鼻音弱化的現象，使得維特比搜尋在該音段選擇轉移狀態相對較大的端點造成自動分段效能變差的情形。

因此綜合以上所述且考量語音信號所挑選出候選端點數目以及觀察候選端點之位置，自動語音分段屬聲/韻母與類音素之層級較為合適，也就是圖 4.16 聲/韻母層級之實驗結果較佳的原因。

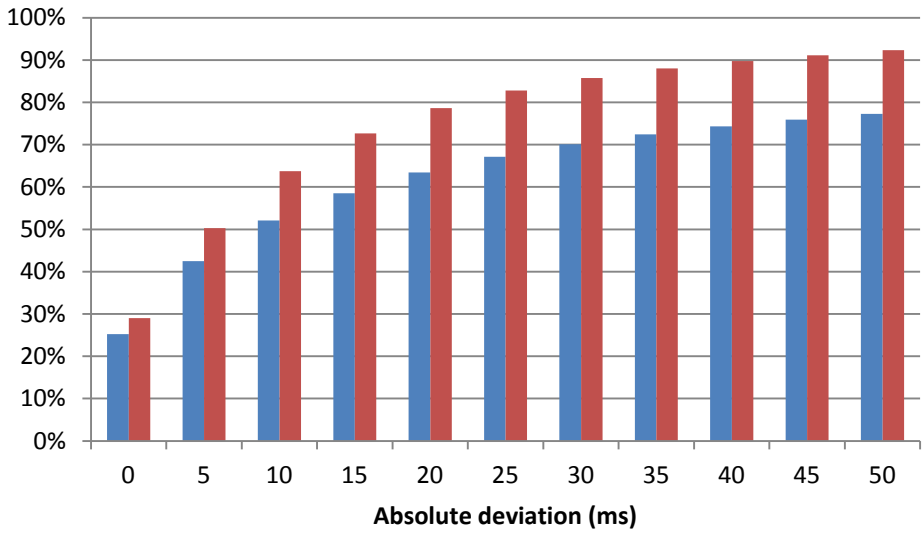


圖 4.16：不同音節結構實驗結果與人為標記位置之誤差在不同絕對偏差值的包含比率直方圖，藍色線(左側)為音節層級，紅色線(右側)為使用聲/韻母層級

在此可作延伸討論，實驗結果顯示對候選端點經過適當地分類標記，可有助於實驗效能之提升。但換句話說，候選端點間之音段也同樣經過了分類標記，那麼以音段為基礎的聲學特性來建立分類的模型，即可應用至語音屬性偵測甚至是語音辨認中。

4.5 使用客語四縣語料庫之實驗結果

4.5.1 音素端點偵測實驗結果

利用 TIMIT 英文語料庫所訓練的音素端點偵測器來偵測客語語料庫內語句的音素端點。圖 4.17 為偵測客語語句音素端點的範例，在語音波形、聲譜圖與端點偵測器輸出音素候選端點之概似度（範圍在-1~1 之間，其值為 1 則代表為音素端點可能性為最大；相反地，值為-1 可能性最小）的對應觀察中，偵測的端點位置亦可對應至語音波形轉換或是頻譜變化的位置。利用英文語料庫所訓練的模型對客語語料進行音素端點偵測之方式，由觀察實驗結果顯示音素端點偵測確實是可跨語言的。然而，由目前偵測器的輸出結果發現當語音屬性不同的轉換時如發音方法不同，則偵測器在該候選端點產生概似度較高的現象，此種現象也呼應了前後音素為不同發音方法其偵測漏失率相對較低。最後，可藉由適當地調整臨限值來達到偵測器最佳輸出之端點偵測結果，依照此結果我們可將語音信號分為一段段的音段，且這些音段即呈現語音信號中較為穩定的部分，如圖中箭頭所例，亦可提供語音屬性偵測的應用甚至是語音辨識所使用。

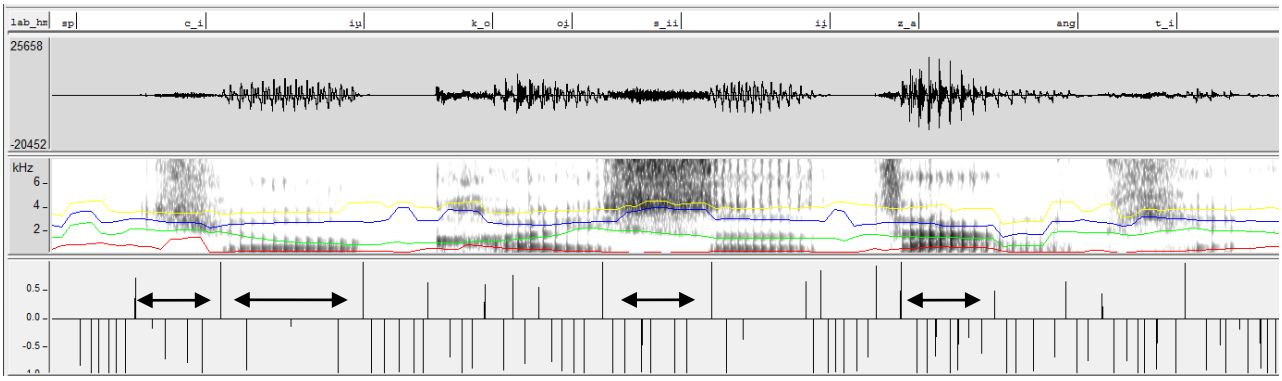


圖 4.17：偵測客語語句音素端點之範例，由上至下的圖形分別代表 HMM 分段位置、語音波形、聲譜圖以及端點偵測器輸出音素候選端點之概似度

4.5.2 自動語音分段實驗結果

利用第四章所述之自動語音分段的方法，來印證在客語語料庫的效能，相比於 TCC300 語料庫其差異在於基本語音單位不同，客語為聲/韻母層級而 TCC300 語料庫為類音素層級。因此目標函數為根據聲/韻母層級所訂定並進行自動語音分段的方法，最後得到對客語語料之自動分段位置。

客語語料庫因無人為時間標記的資訊，故以下列舉兩個語音波形的範例來比較本研究方法與 HMM 自動分段的準確度。圖 4.18、圖 4.19 之中，可由方圈之圈選處觀察到一些現象，其音節之間邊界或是聲母與韻母之間的端點位置也都能修正至較為準確的位置，在爆破音與短停頓的交界或是不同發音方式的轉換點尤其明顯，由上述實驗結果在語音波形的觀察下，顯示了本研究所提出之自動語音分段方法對 HMM 之自動分段位置做修正後，其自動分段之效能亦能有所提升。

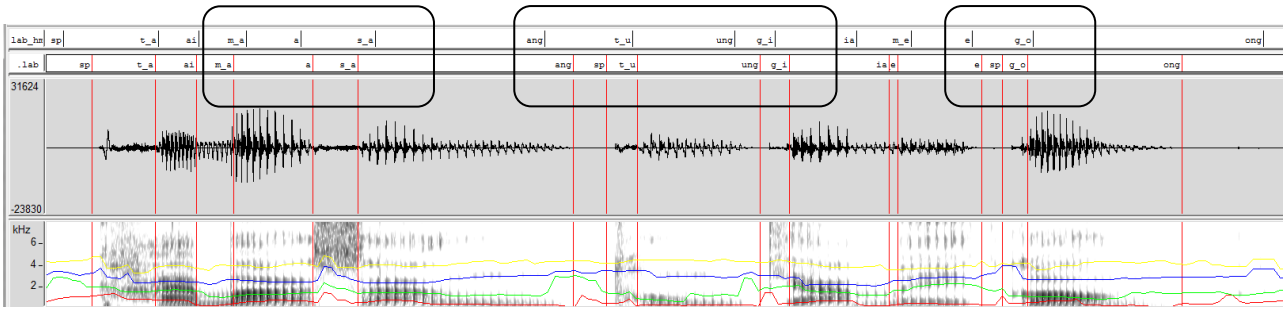


圖 4.18：客語語句自動語音分段之範例一，由上至下的圖形分別表示 HMM 分段位置及音素端點偵測之分段位置、語音波形、聲譜圖

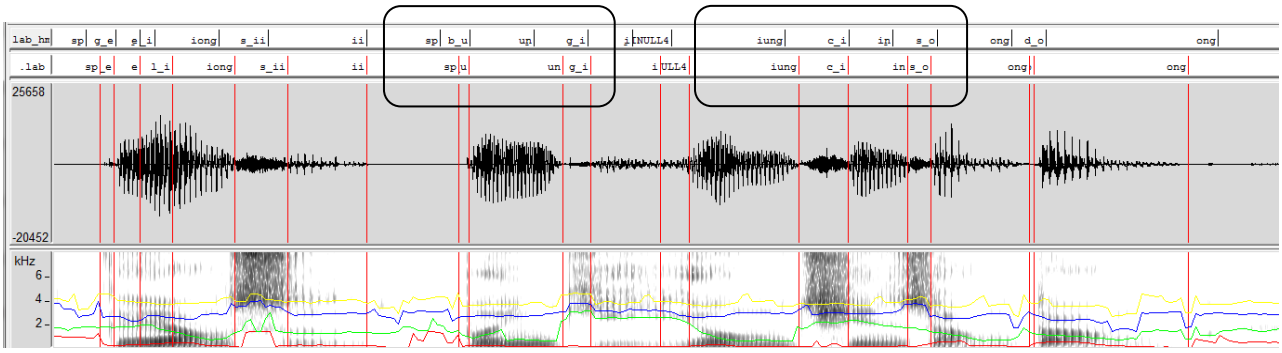


圖 4.19：客語語句自動語音分段之範例二，由上至下的圖形分別表示 HMM 分段位置及音素端點偵測之分段位置、語音波形、聲譜圖

第五章 使用音段式語音發音方法辨認器

在取樣式聲學參數之音素端點偵測器之應用方面，在計畫中我們也做了在發音方法(pronunciation manner)辨認的實驗。

語音信號經取樣式聲學參數之音素端點偵測器後，會被切割成一連串語音段(speech segment)；我們可以讓取樣式聲學參數之音素端點偵測器操作在低偵測漏失率的工作點，當然會有較高的假警報率，但是每段的平均長度會大於音框式系統中的音框移動值(10 msec)。接著，計畫中將以音段作為發音方法的辨認單元，而非傳統系統中使用固定長度及距離的音框(frame)為單位做辨認。而我們所使用的音段長度約略與音素的長度相當，所以對同一語音信號平均音段數會小於傳統的音框數。但在語音信號變化快的地方，我們所使用的音段長度會隨之變短，所以可以將較短之音節描述得較精確。

5.1 音段式發音方法辨認器之參數抽取

在語音信號被切割成音段後，每一個取樣式聲學參數在此音段會形成一個時間函數(或曲線)；例如最基本的語音信號波封與子頻段信號波封，如圖 5.1 所示。我可可將這些曲線取出固定為度的參數來做為該音段發音方法辨認器之輸入參數。在此，我們使用 discrete legengre 多項式的參數[16]來表示一段曲線。

discrete legengre 多項式就是對一段參數曲線，$f(i); i = 0, \cdots, N-1$，用下列基底函式來表示，

$$
\begin{aligned}
\phi_0\left(\frac{i}{N}\right) &= 1 \\
\phi_1\left(\frac{i}{N}\right) &= \left[\frac{12N}{(N+2)}\right]^{1/2}\left[\left(\frac{i}{N}\right)-\frac{1}{2}\right] \\
\phi_2\left(\frac{i}{N}\right) &= \left[\frac{180N^2}{(N-1)(N+2)(N+3)}\right]^{1/2}\left[\left(\frac{i}{N}\right)^2-\left(\frac{i}{N}\right)-\frac{N-1}{6N}\right] \\
\phi_3\left(\frac{i}{N}\right) &= \left[\frac{2800N^3}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right]^{1/2} \\
&\quad \left[\left(\frac{i}{N}\right)^3-1.5\left(\frac{i}{N}\right)^2-\frac{6N^2-3N+2}{10N^2}\left(\frac{i}{N}\right)-\frac{(N-1)(N-2)}{20N^2}\right]
\end{aligned}
\tag{5.1}
$$

也就是一個時間函數(或曲線)可以用下列基底函式來表示

$$f(i) \cong \sum_{k=0}^{3} a_k \phi_k \left(\frac{i}{N} \right); i = 0, \cdots, N-1 \qquad (5.2)$$

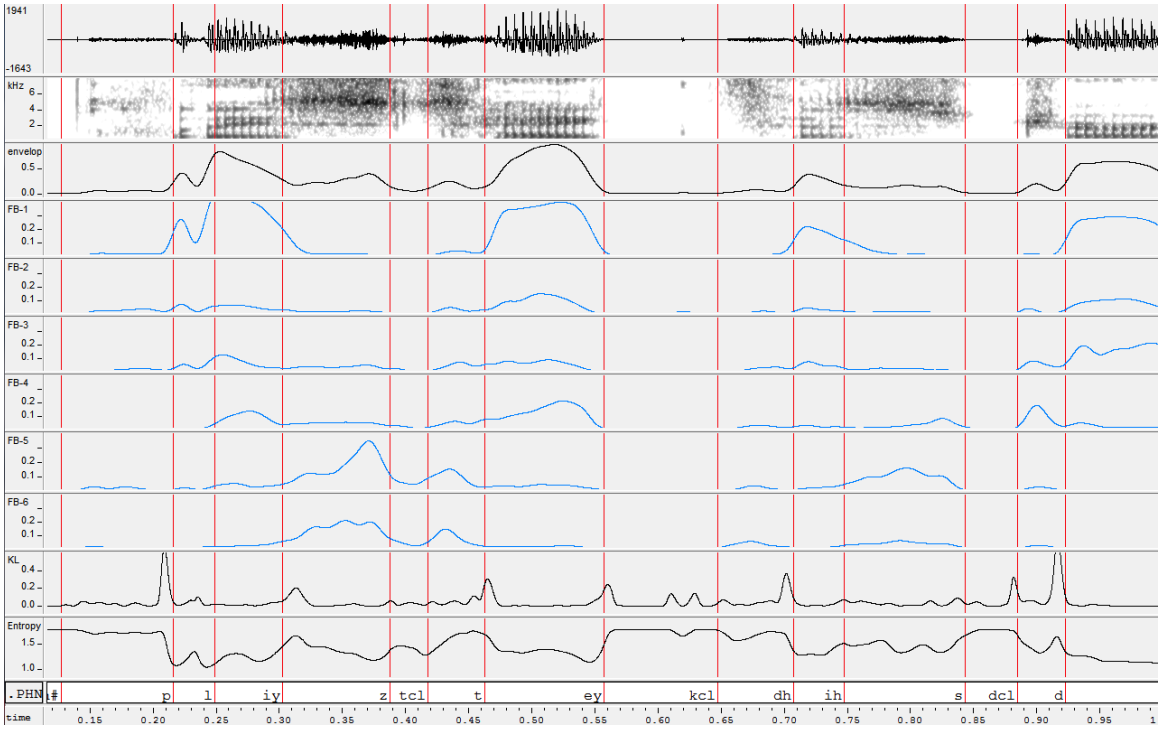


圖 5.1:音段式發音方法辨認器之參數抽取之示意圖。

在此我們使用第二章所求得之正規化語音信號波封與子頻段信號波封之對數值，

$$\log(E_i), i = 0, \cdots, 6$$

為參數。對每個音段，正規化語音信號波封與子頻段信號波封之對數值之曲線各求 4 個 discrete legengre 多項式參數$(a_i), i = 0, \cdots, 6$，總共為 28 個參數。

在音段式語音發音方法辨認器中，使用待辨認音框、前後各 1 個音段的音段參數及各音段的長度，合計共 87 維參數。

5.2 使用音段式發音法法辨認器辨認結果

在實驗中，我們使用 TIMIT 語料庫，讓第一級取樣式聲學參數之音素端點偵測器操作在低偵測漏失率的工作點(在下列實驗中設定為 3%)，取樣式聲學參數之音素端點偵測器會的到總音段數為 80123，也就是平均一個音素會被分成 1.25 段。

我們使用 RNN 作為音段式語音發音辨認器，輸入層有 87 個神經元，所使用的隱藏層神經元數目為 90 個，輸出為 6 種發音方式。所獲得語音發音方法辨認率如表 5.1 所示。若與李錦輝教授論文[17]中使用音框式參數之語音發音辨認器比較，在[17]中實際上是使

用待辨認音框及前後個 4 個音框(共 9 個音框)之 MFCC 參數(共 108 維參數)作為發音方法辨認器之輸入參數，使用 MLP 類神經網路做辨認器。其以音框為單位(frame-based)之發音方式辨認結果也並列於表 5.1 中。

表 5.1：使用音段式語音發音辨認方法之辨認結果

Pronunciation manner	**Segment-based Recog. Rate (%)**	**Frame-base Recog. Rate(%)[17]**
Fricative	79.4	85.2
Stop	78.8	72.5
Glide	68.0	56.5
Vowel	90.5	89.0
Nasal	75.4	77.5
Silence	89.2	92.2
Total	83.3	82.1

在表 5.1 中，可以發現對長度較短的音素如：stop 及 glide，使用音段式語音發音方法辨認方法，其辨認率可以大幅改善。

若將音段換算為 frame-based 的辨認率，使用音段式語音發音方法辨認器其總音框辨識率為 83.15%，也較[17]中的結果為佳。

另我們驚訝的是我們使用較低的頻率解析度所求的知參數還能獲得較佳的結果；所以在音段式語音發音方法及位置辨認或偵測器上，也就是 detection-based ASR 上之應用，將還有進一步探討的空間。

計畫成果自評

本計畫以獲得一個精確的自動音素端點偵測系統為目標，進行音素端點偵測及自動音素分段的實驗，以提升對應於語音信號之文字轉寫時間標記之精準度，降低人為標記語料庫費時費力的繁雜過程，並期望作為語音辨識或是語音合成系統中處理文字轉寫的標準流程。 同時也利用自動音素端點偵測後所得到的語音音段，建立了音段式發音方法之辨認器，發現其效能會優於音框式發音方法之辨認器。

本計畫提出數個取樣點式聲學參數如各頻段信號波封、聲學參數之上升率、頻譜熵以及頻譜 KL 距離，以描述語音信號中各種不同音素之語音特性，加入音素端點偵測以及自動語音分段的系統架構中，針對音素端點偵測是以音素之邊界端點作為目標函數，另外自動語音分段的架構是根據文字轉寫所使用不同基本語音單位（音節、聲/韻母、類音素）的層級來分類，並依照各個分類彼此之間可能的轉換狀態訂定目標函數，再分別使用類神經網路之多層感知器架構，以半監督式之模型訓練方法建立起音素端點偵測器模型。

在音素端點偵測的效能及錯誤分析的過程中，對於前後音素為相同發音方法時，因為兩者在頻域與時域上的變化不明顯，使偵測器之偵測漏失率會有增高的現象。

在自動語音分段的效能分析中，實驗以隱藏式馬可夫模型以文字轉寫中不同層級的語音單位，來得到初始的分段位置，再加入端點偵測器的模型進行修正而獲得更好的分段位置。此方式和初始分段位置一同和人為之時間標記位置相比後，準確率確實有效地提升，且根據音節不同層級之實驗結果，分析後顯示聲/韻母、類音素層級之效能較音節高。另一方面，也歸納了數個會造成自動語音分段效能降低的現象。

本計畫中也提出了使用音段參數之語音發音方法辨認器，其效能亦優於音框式參數之發音方法辨認器。

完成工作項目

(1) TIMIT 語料庫之音素端點偵測器及使用音段資訊之兩階段式音素端點偵測器
(2) 在國語語料方面，計畫中完成了 TCC300 及 Treebank 語料庫之音素端點自動分段(alignment)
(3) 台灣方言語料方面，也完成客語語料庫之音素端點自動分段(alignment)
(4) 也完成利用音素端點偵測器將語音信號分為音段後，完成一個語音辨認之應用：使用音段式之語音發音方法偵測

產出物

(1) TCC-300 語料庫之類音素端點標示資料 – TCC300 是國內從事語音辨認研究及

公司須會購買之語料庫(中華民國計算語言學會)，TCC-300 語料庫之類音素端點標示資料可以授權發行，將對 TCC-300 語料庫使用者有極大的助益。

已發表之論文

(1) You-Yu Lin, Yih-Ru Wang, "Sample-based Phone-like Unit Automatic Labeling in Mandarin Speech, ", Proc. of *ROCLING* 2009, Taichung, ROC. pp. 137-149, Sept. 2009.

(2) You-Yu Lin, Yih-Ru Wang and Yuan-Fu Liao, "Phone Boundary Detection using Sample-based Acoustic Parameters, ", Proc. of *INTERSPEECH*-2010, Makuhari, Japan, pp. 1397-1400, Spet., 2010.

(3) Yih-Ru Wang, "A Two-stage Sample-based Phone Boundary Detector using Segmental Similarity Features, ", Proc. of *INTERSPEECH*-2011, Florence, Italy, pp. 413-416, Aug., 2011. (本篇論文之內容未詳列於本報告，故列於附件，本論文基本上在偵測出音素端點後，如同第五章中一樣在使用音段參數來幫助，可以做到更好的音素端點偵測。)

已投稿之論文

(4) Yih-Ru Wang, You-Yu Lin, "High-Resolution Phone Boundary Detection using Sample-based acoustic Parameters, ", submitted to IEEE Trans. on Audio, Speech and Language and Processing.

參考文獻

[1] F. Malfrère, O. Deroo, and T. Dutoit, “Phonetic alignment: Speech synthesis based vs. hybrid HMM/ANN,” in *Proceedings of the International Conference on Spoken Language Processing*, vol. IV, Sydney, NSW, Australia, 1998, pp. 1571–1574.

[2] Toledano, D.T.; Gomez, L.A.H.; Grande, L.V., “Automatic phonetic segmentation,” *Speech and Audio Processing, IEEE Transactions on* , vol.11, no.6, pp. 617-625, Nov. 2003.

[3] Jen-Wei Kuo and Hsin-min Wang, “Minimum Boundary Error Training for Automatic Phonetic Segmentation,” *The Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)*, September 2006.

[4] J.-W. Kuo, H.-Y. Lo, and H.-M. Wang, “Improved HMM/SVM methods for automatic phoneme segmentation,” in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2057-2060.

[5] K.-S. Lee, “MLP-based phone boundary refining for a TTS database,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 981–989, 2006.

[6] Sorin Dusan and Lawrence Rabiner, “On the Relation between Maximum Spectral Transition Positions and Phone Boundaries,” in *Proc. Interspeech 2006*, pp. 17–21.

[7] J. Garofolo et al., “Documentation for the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM,” Feb. 1993.

[8] Almpanidis, G., Kotti, M., Kotropoulos, and C., “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, Feb. 2009.

[9] Sharlene A. Liu, “ Landmark detection for distinctive feature-based speech recognition,” J. Acoust. Soc. Am. 100 (5), November 1996, pp. 3417-3430.

[10] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, “Spectral entropy based feature for robust ASR,” *in Proc. ICASSP 2004*, pp. 193–196.

[11] Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee, “Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments,” *in Proc. ICSLP 1998*.

[12] Nico Tool Kit : Available: http://nico.nikkostrom.com

[13] Li Lao, X Wu, L Cheng, X Zhu, “Maximum weighted entropy clustering algorithm,” *Proceedings of the 2006 IEEE International conference on Networking, Sensing and Control*, pp. 1022-1025.

[14] B.-H. Juang, and S. Katagiri, “Discriminative learning for minimum error classification,” *IEEE Trans. Speech and Audio Processing*, vol. 40, no. 12, pp.

3043-3054, Dec., 1992.

[15] B.-H. Juang, W. Hou and C.-H. Lee, “Minimum classification error rate Methods for Speech Recognition,” *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 257-265, May 1997.

[16] Sin-Horng Chen and Yih-Ru Wang, ' Vector Quantization of Pitch Information in Mandarin Speech, ', IEEE Trans. on Communications, Vol. 38, No. 9, pp. 1317-1320, Sept., 1990.

[17] Sabato Marco Siniscalchi, Jinyu Li, Chin-Hui Lee, "A study on lattice rescoring with knowledge scores for automatic speech recognition", *INTERSPEECH-2006*, pp. 1319-1322.

[18] 林宥余，”使用取樣點式語音聲學參數之音素端點偵測”，交通大學碩士碩文，民國 99 年。

附錄一

TIMIT 語料庫音素分類表

發音方法	標音符號	範例字詞	音素層級之文字轉寫
爆破音	*b*	bee	*BCL B iy*
	d	day	*DCL D ey*
	g	gay	*GCL G ey*
	p	pea	*PCL P iy*
	t	tea	*TCL T iy*
	k	key	*KCL K iy*
	dx	muddy,	*dirty m ah DX iy, dcl d er DX iy*
	q	bat	*bcl b ae Q*
塞擦音	*jh*	joke	*DCL JH ow kcl k*
	ch	choke	*TCL CH ow kcl k*
摩擦音	*s*	sea	*S iy*
	sh	she	*SH iy*
	z	zone	*Z ow n*
	zh	azure	*ae ZH er*
	f	fin	*F ih n*
	th	thin	*TH ih n*
	v	van	*V ae n*
	dh	then	*DH e n*
鼻音	*m*	mom	*M aa M*
	n	noon	*N uw N*
	ng	sing	*s ih NG*
	em	bottom	*b aa tcl t EM*
	en	button	*b ah q EN*
	eng	washington	*w aa sh ENG tcl t ax n*
	nx	winner	*w ih NX axr*
半母音與流音	*l*	lay	*L ey*
	r	ray	*R ey*
	w	way	*W ey*
	y	yacht	*Y aa tcl t*
	hh	hay	*HH ey*
	hv	ahead	*ax HV eh dcl d*

	el	bottle	*bcl b aa tcl t EL*
母音	*iy*	beet	*bcl b IY tcl t*
	ih	bit	*bcl b IH tcl t*
	eh	bet	*bcl b EH tcl t*
	ey	bait	*bcl b EY tcl t*
	ae	bat	*bcl b AE tcl t*
	aa	bott	*bcl b AA tcl t*
	aw	bout	*bcl b AW tcl t*
	ay	bite	*bcl b AY tcl t*
	ah	but	*bcl b AH tcl t*
	ao	bought	*bcl b AO tcl t*
	oy	boy	*bcl b OY*
	ow	boat	*bcl b OW tcl t*
	uh	book	*bcl b UH kcl k*
	uw	boot	*bcl b UW tcl t*
	ux	toot	*tcl t UX tcl t*
	er	bird	*bcl b ER dcl d*
	ax	about	*AX bcl b aw tcl t*
	ix	debit	*dcl d eh bcl b IX tcl t*
	axr	butter	*bcl b ah dx AXR*
	ax-h	suspect	*s AX-H s pcl p eh kcl k tcl t*
其他	SYMBOL	DESCRIPTION	
	pau	pause	silence
	epi	epenthetic	silence
	h#	begin/end	marker (non-speech events)
	1	primary	stress marker
	2	secondary	stress marker

附錄二

中文音素分類對照表

表一、國語 21 類聲母表

編號	拼音	注音	編號	拼音	注音	編號	拼音	注音
1	*zh*	ㄓ	8	*g*	ㄍ	15	*t*	ㄊ
2	*ch*	ㄔ	9	*k*	ㄎ	16	*n*	ㄋ

3	*sh*	ㄕ	10	*h*	ㄏ	17	*l*	ㄌ
4	*r*	ㄖ	11	*j*	ㄐ	18	*b*	ㄅ
5	*z*	ㄗ	12	*q*	ㄑ	19	*p*	ㄆ
6	*c*	ㄘ	13	*x*	ㄒ	20	*m*	ㄇ
7	*s*	ㄙ	14	*d*	ㄉ	21	*f*	ㄈ

其中，關於空聲母（INULL）為當音節只有韻母發音時給予的預設聲母。

表二、國語 18 類韻母表

編號	拼音	注音	編號	拼音	注音	編號	拼音	注音
1	*FNULL1*	Φ1	7	*a_ng*	ㄤ	13	*e_ng*	ㄥ
2	*FNULL2*	Φ2	8	*o*	ㄛ	14	*e_n*	ㄣ
3	*a*	ㄚ	9	*ou*	ㄡ	15	*er*	ㄦ
4	*ai*	ㄞ	10	*e*	ㄜ	16	*yi*	ㄧ
5	*ao*	ㄠ	11	*eh*	ㄝ	17	*wu*	ㄨ
6	*a_n*	ㄢ	12	*ei*	ㄟ	18	*yu*	ㄩ

其中，關於注音中ㄢ、ㄤ、ㄥ、ㄣ，在本計畫中之類音素層級將韻母細分，使鼻音韻尾自成一類。

表三、國語 2 類鼻音韻尾

編號	拼音
1	*n_n*
2	*ng*

其中，/*n_n*/和/*ng*/分為ㄢ、ㄣ及ㄥ、ㄤ的鼻音韻尾。

附錄三

Yih-Ru Wang, “A Two-stage Sample-based Phone Boundary Detector using Segmental Similarity Features, “, Proc. of *INTERSPEECH*-2011, Florence, Italy, pp. 413-416, Aug., 2011.
(本篇論文之內容未詳列於本報告，故列於附件，本論文基本上在偵測出音素端點後，如同第五章中一樣在使用音段參數來幫助，可以做到更好的音素端點偵測。)

A Two-Stage Sample-based Phone Boundary Detector using Segmental Similarity Features

Yih-Ru Wang

Institute of Communication, National Chiao Tung University, Hsinchu, Taiwan, ROC

yrwang@mail.nctu.edu.tw

Abstract

In this paper, a two-stage sample-based phone boundary detection algorithm is proposed. In the first stage, some local sample-based acoustic parameters are used to pre-select some phone boundary candidates. Then, in the second stage, some high-order statistics of the log-likelihood differences of two adjacent speech segments around each boundary candidate are calculated to serve as similarity measure for candidate verification. Experimental results on the TIMIT speech corpus showed that EERs of 8.6% and 7.6% were achieved for one-stage and two-stage sample-based phone boundary detections, respectively. Moreover, for the two-stage system, 42.1% and 81.9% of boundaries detected were within 5- and 15-sample error tolerance from manual labeling results.

Index Terms: phone boundary detection, similarity measure

1. Introduction

Automatic phonetic segmentation is a historic and basic problem in speech signal processing. Although a lot of researches had been done in the past [1], an automatic phonetic segmentation algorithm with high accuracy and precision is still a state-of-the-art work. Without knowing the text of the speech signal, it becomes a phone boundary detection problem which is more difficult than the phone boundary alignment problem. An accurate phone boundary detector is important and essential for speech processing engineering and linguistics.

In automatic boundary detection without knowing the text of the speech signal, the rate of acoustic signal change is the most important cue for decision making. In [2], the spectral transition measure, which is in fact the norm of delta MFCC, was used to find the phone boundaries. 15.4% miss detection (MD) and 22.0% false alarm (FA) rates were achieved on the TIMIT training data set. In [3], the model selection technique, DISTBIC, was used to perform the phone boundary detection. The DISTBIC first used the Kullback-Leibler (KL) distance to find the boundary candidates, and then employed the Bayesian information criterion (BIC) to further verify those candidates. 25.7% MD and 23.3% FA rates were achieved on the NTIMIT database. In our previous work [4], some sample-based acoustic features were proposed to model the rapid spectral changes in speech signal. Both the precision and accuracy of the sample-based phone boundary detector were shown to be better than those of frame-based algorithms, such as the system shown in [3].

In this paper, a two-stage sample-based phone boundary detector is proposed. It is a modification of our previous system [4]. In the first stage, some sample-based phone boundary candidates are found. Then, in the second stage, each candidate is verified by using a new similarity measure with features extracted from the neighboring speech segments. For obtaining the similarity measure, a more precise signal modeling method, the common component Gaussian mixture model (CCGMM) [5], is employed to model the speech signal. Some high-order statistics of the log-likelihood difference functions of the two neighboring segments, like mean, variance and skewness, can then be represented in terms of CCGMM coefficients [6]. These high-order statistics are used to calculate the similarity measure for improving boundary candidate verification.

The paper is organized as follows. In Section 2, the proposed sample-based phone boundary detection algorithm is discussed in detail. The performance of the two-stage system is examined by simulations discussed in Section 3. Some conclusions are given in the last section.

2. Two-stage Sample-based Phone Boundary Detector

In the proposed two-stage sample-based phone boundary detection algorithm, speech signal is first processed sample-by-sample to extract some sample-based acoustic parameters. Then, those local acoustic parameters are used in the first stage to detect some candidates of phone boundary. The speech signal is accordingly segmented into lots of acoustic segments. In the second stage, the high-order statistics of the log-likelihood difference of two neighboring segments are calculated to serve as the similarity measure of the two segments for verifying the boundary candidate. In the following subsections, we discuss these two stages in detail.

2.1. First-stage Boundary Candidate Detection

It is known that the spectrum of a speech signal is an effective cue for phone boundary detection. In this study, six sub-band signal envelopes are used. The input speech signal firstly passes through six band-pass filters with cutoff frequencies shown below

0.0– 0.4 KHz, 0.8 – 1.5 KHz, 1.2 – 2.0 KHz,
2.0 – 3.5 KHz, 3.5 – 5.0 KHz, 5.0 – 8.0 KHz.

The energies of the above six sub-band signals were shown to be effective in speech landmark detection [7]. In the sample-based approach, the envelopes of those sub-band signals, $x_i[n]$, are extracted instead of their energies. Detection of each sub-band signal envelope is realized by passing the complex analytic signals, $x_i[n] + j \cdot y_i[n]$, through a low-pass filter. The Hilbert transformed signals, $y_i[n]$, in analytic signals can be produced by

$$y_i[n] = x_i[n] \otimes h[n] \quad \text{for } i = 1, \cdots, 6 \tag{1}$$

where

$$h[n]=\begin{cases}1/|n-N|\pi, & n \text{ is odd and } 0<n<2N\\ 0, & \text{otherwise}\end{cases}$$

The envelope of the *i*-th sub-band signal is denoted by $e_i[n]$. Besides, the envelope of the original speech signal, $e_0[n]$, is also extracted. The cutoff frequency of the low-pass filter is set to 30 Hz.

The low-passed KL distance was used in probability theory to measure the similarity of two distributions. In this study, we use it to measure the similarity of two adjacent speech samples represented by six sub-band envelopes, $\{e_i[m];i=1,\cdots,6\}$ for $m=n$ and $n+1$. The KL distance is implemented by first normalizing the six sub-band signal envelopes [8] by

$$E_i[n]=\frac{e_i[n]}{\sum_{j=1}^{6}e_j[n]} \tag{2}$$

Then, the sample-based KL distance is calculated by

$$d_{KL}[n]=\sum_{i=1}^{6}\left(E_i[n]-E_i[n+1]\right)\log\left(\frac{E_i[n]}{E_i[n+1]}\right) \tag{3}$$

Spectral entropy is commonly used in measuring the flatness of a speech power spectrum in a frame-based system [7]. In this study, it is extended to the sample-base spectral entropy defined by

$$H[n]=\sum_{i=1}^{6}\left(E_i[n]\right)\log\left(E_i[n]\right) \tag{4}$$

Its value will be small in the fricative/affricate and nasal parts of speech signal.

The similarity of the signals around the boundary candidate can also be a useful measure of signal change. For each boundary candidate, c_j, the feature vectors $\left(E_i[n];i=0,\cdots,6\right)$ in its two neighborhood windows B_j^- and B_j^+ are assumed to be normal distributed. Here, B_j^- and B_j^+ are defined as

$$B_j^-=[c_j-r_j^-,c_j-1],\ B_j^+=[c_j,c_j+r_j^+],$$

where the lengths, r_j^- and r_j^+, are defined by

$$r_j^-=\begin{cases}r_{\min} & ,c_j-1-c_{j-1}<r_{\min}\\ c_j-1-c_{j-1} & ,r_{\min}\le c_j-1-c_{j-1}\le r_{\max}\\ r_{\max} & ,r_{\max}<c_j-1-c_{j-1}\end{cases}$$

and

$$r_j^+=\begin{cases}r_{\min} & ,c_{j+1}-c_j<r_{\min}\\ c_{j+1}-c_j & ,r_{\min}\le c_{j+1}-c_j\le r_{\max}\\ r_{\max} & ,r_{\max}<c_{j+1}-c_j\end{cases}.$$

The maximum and minimum window lengths, $r_{\max}$ and $r_{\min}$, are set to 5 ms and 10 ms, respectively in this study. The KL distance at boundary candidate, c_j, can be defined as the KL distance of the *pdfs* of the feature vectors in B_j^- and B_j^+, i.e.,

$$D_{KL}[c_j]=\frac{1}{2}\left\{\begin{array}{l}tr[(\Sigma_- -\Sigma_+)(\Sigma_-^{-1}-\Sigma_+^{-1})]+\\ tr[(\mu_- -\mu_+)^T(\Sigma_-^{-1}+\Sigma_+^{-1})(\mu_- -\mu_+)]\end{array}\right\}, \tag{5}$$

where μ_- and μ_+ are the means of feature vectors in B_j^- and B_j^+; and Σ_- and Σ_+ are the covariance matrices.

The normalized sub-band signal envelope, sample-based KL distance and spectral entropy and their delta terms are effective parameters for modeling the short-term spectral changing rate. They are used as the input features of the first-stage phone boundary pre-selection.

2.1.1. Sample-based boundary detection by neural networks

A boundary candidate pre-selection procedure is first used to reduce the number of data needed to be processed in the following boundary detection. The selected boundary candidates are those samples having larger speech signal changing rate. Thus, the sample-based KL distance is employed for boundary candidate pre-selection. A simple peak picking method with threshold is used to select all samples which satisfy the following constrains as candidates

$$d_{KL}[n]>d_{KL}[n-1],\ d_{KL}[n]>d_{KL}[n+1],\ \text{and } d_{KL}[n]\ge Th_d, \tag{6}$$

where Th_d is a threshold. The sequence of boundary candidates is denoted as $\{c_j;j=1\cdots,N_c\}$.

The average normalized sub-band envelope of the segment, $[c_{k-1},c_k]$, is defined by

$$ES_i[c_{k-1},c_k]=\left(\sum_{n=c_{k-1}+1}^{c_k}E_i[n]\right)/\left(c_k-c_{k-1}-1\right) \tag{7}$$

Then, a 27-dimensional feature vector is constructed for each boundary candidate. For the candidate at time c_k, its feature vector includes the following acoustic parameters,

(1) Features from current boundary candidates :

$$d_{KL}[c_k],\ D_{KL}[c_k],\ H[c_k],\ \Delta H[c_k],\ \left(E_i[c_k];i=0,\cdots,6\right),$$

where $\Delta H[c_j]$ is the delta term of the sample-based spectral entropy.

(2) Features from adjacent segments $[c_{k-1},c_k]$ and $[c_k,c_{k+1}]$:

$$\left(ES_i[c_{k-1},c_k],ES_i[c_k,c_{k+1}];i=0,\cdots,6\right),c_k-c_{k-1},c_{k+1}-c_k$$

Lastly, two neural network-based classifiers, a multi-layer perceptron (MLP) and a recurrent neural network (RNN), are used to screening these phone boundary candidates.

2.2. Similarity measure of acoustic signal segments

After pre-selecting some boundary candidates in the first stage, we then verify each of them in the second stage. A new similarity measure based on CCGMM representation is introduced to calculate the distance of the two acoustic segments around a candidate for its verification.

For a speech segment *k*, the *pdf* of its acoustic feature vectors, $o[n]=\left[E_1[n],\cdots,E_6[n]\right]$, can be modeled by

$$p_k(o[n])=\sum_{l=1}^{L}c_{lk}\cdot N(o[n];\mu_{kl},\Sigma), \tag{8}$$

where $\{N(o[n];\mu_{kl},\Sigma);\ l=1,\cdots,L\}$ is a set of Gaussian distributions with common covariance matrix which is used as

the basis of signal *pdf*; and $\{c_{kl};\ l=1,\cdots,L\}$ are the coefficients of CCGMM [5].

Then, the un-symmetric KL (KL1) distances are calculated and used as the similarity measures of the two adjacent segments O_1 and O_2:

$$D_1(O_1 \mid O_2) = \int p_1(o) \ln \frac{p_1(o)}{p_2(o)} do = \int p_1(o) \mathbf{P}_{1-2}(o) do;$$

$$D_2(O_2 \mid O_1) = \int p_2(o) \ln \frac{p_2(o)}{p_1(o)} do = \int p_2(o) \mathbf{P}_{2-1}(o) do \ , \qquad (9)$$

where $\mathbf{P}_{1-2}(o)$ and $\mathbf{P}_{2-1}(o)$ are log-likelihood difference functions of the two segments. They can be approximated by [5]

$$D_{1-2}(O_1 \mid O_2) = E_1\left(\mathbf{P}_{1-2}(o)\right) \simeq \sum_i c_{1i} \cdot \ln\left(\frac{c_{1i}}{c_{2i}}\right) \triangleq \mu_{1-2}$$

$$D_{2-1}(O_2 \mid O_1) = E_2\left(\mathbf{P}_{2-1}(o)\right) \simeq \sum_i c_{2i} \cdot \ln\left(\frac{c_{2i}}{c_{1i}}\right) \triangleq \mu_{2-1} \qquad (10)$$

Instead of using the means of $\mathbf{P}_{1-2}(o)$ and $\mathbf{P}_{2-1}(o)$ only, some high-order statistics of $\mathbf{P}_{1-2}(o)$ and $\mathbf{P}_{2-1}(o)$ are also used for modeling the similarity of two adjacent speech segments. The variances and skewnesses of $\mathbf{P}_{1-2}(o)$ and $\mathbf{P}_{2-1}(o)$ can be approximated by

$$\sigma_{1-2} \simeq \left(\sum_{i=1}^{N} c_{1i} \left(\ln \frac{c_{1i}}{c_{2i}} \right)^2 - \left(\mu_{1-2}\right)^2 \right)^{1/2};$$

$$S_{1-2} \simeq \left(\sum_{i=1}^{N} c_{1i} \left(\ln \frac{c_{1i}}{c_{2i}} - \mu_{2-1} \right)^3 \right)^{1/3} \left(\sigma_{1-2}\right)^{-1};$$

$$\sigma_{2-1} \simeq \left(\sum_{i=1}^{N} c_{2i} \left(\ln \frac{c_{2i}}{c_{1i}} \right)^2 - \left(\mu_{2-1}\right)^2 \right)^{1/2};$$

$$S_{2-1} \simeq \left(\sum_{i=1}^{N} c_{2i} \left(\ln \frac{c_{2i}}{c_{1i}} - \mu^1 \right)^3 \right)^{1/3} \left(\sigma_{2-1}\right)^{-1}. \qquad (11)$$

Besides these first- to third-order statistics of the log-likelihood differences, some segment-based acoustic features calculated in the first stage are also used. In summary, the 30-dim features used for determining whether the candidate at time c_k is a phone boundary are listed below:

(1) Output of first-stage RNN boundary detector,

(2) Features from two adjacent segments:

$$\left(E_i[c_k], ES_i[c_{k-1}, c_k], ES_i[c_k, c_{k+1}]; i = 0, \cdots, 6\right),$$
$$(c_k - c_{k-1}), (c_{k+1} - c_k)$$

(3) Statistics of the log-likelihood differences of two adjacent segments, $c_k^+ = [c_k, c_{k+1}]$, $c_k^- = [c_{k-1}, c_k]$:

$$(\mu_{c_k^+ - c_k^-}, \sigma_{c_k^+ - c_k^-}, S_{c_k^+ - c_k^-}, \mu_{c_k^- - c_k^+}, \sigma_{c_k^- - c_k^+}, S_{c_k^- - c_k^+}) .$$

Finally, a recurrent neural network (RNN) is used as the second-stage phone boundary detector.

3. Experiment Results

The TIMIT speech corpus was used to evaluate the effectiveness of the proposed sample-based phone boundary detection algorithm. The numbers of phone boundaries in the training and testing parts of TIMIT were 172460 and 62465, respectively. The total numbers of samples were $2.27*10^8$ and $8.29*10^7$ for training and testing data sets. In average, there were 12.2 phone boundaries in 1 sec, or one boundary per 1310 samples for training data.

First, the envelopes of six sub-band signals were found from the speech signal. Then, the sample-based KL distance and spectral entropy were extracted. The threshold value of KL distance was properly chosen empirically to preselect the boundary candidates. The numbers of the resulting boundary candidates were 534189 and 194201 for the training and test data sets, respectively. In other words, only 0.85% speech samples, or one out of 116 samples, were preselected as boundary candidates for the training data.

These boundary candidates were then screened by an MLP classifier and an RNN classifier in which all neurons in the hidden layer were fully feedback to themselves. The numbers of hidden neurons were empirically set to 75 and 80 for MLP and RNN classifiers, respectively. These two neural networks were trained using the iterative target selection algorithm [4]. Lastly, the NN output was compared with a threshold for the first-stage decision.

The curves of MD (Miss Detection) rate vs. FA (False Alarm) rate for both MLP and RNN classifiers are shown in Figure 1. Note that the FA was defined as (number of false alarms)/(number of true boundaries + number of false alarms). EERs of 11.6% and 8.6% were achieved for MLP and RNN classifiers, respectively. Compared with the performance of [4], which is 15.4% MD and 22% FA rates for the TIMIT training data set, about 50% EER reduction was achieved. In order to check the accuracy of those detected boundaries, the normalized cumulative histogram of the absolute deviation between the automatically detected boundary and the manually labeled one was shown in Figure 2. As shown in the figure, about 70% phone boundaries were correctly detected within 10-ms error tolerance.

In order to compare the performance of the proposed systems with the state-of-the-art model-based approach, an HMM phone recognition system was implemented. 61 3-state phone models were built. Besides, the minimum mean absolute boundary error (MMAE) criterion was used to realign the recognition result of the HMM system. The EER of the HMM system was 6.3%. The EER of the sample-based RNN phone boundary detector is about 20% higher than that of the HMM system. However, as shown in Figure 2, the inclusion rate of the RNN detected boundaries is much higher than those of the HMM recognizer when the error tolerance is less than 30-msec. So, the RNN boundary detector is of higher precision.

In order to perform the 2nd-stage verification, a low threshold value was adopted for the first-stage RNN classifier. The resulting MD rate is only 3.9%, while the FA rate is 17.7%. To calculate the similarity measure of speech segments found from stage 1, 64 Gaussian mixtures were used in the CCGMM. An RNN with 80 hidden neurons was used as the second-stage verification. The curve of MD rate vs. FA rate for the two-stage phone boundary detector was also shown in Figure 1. An EER of 7.6% was achieved. The performance was much better than that of the first-stage RNN detector. It can also be found from Figure 2 that 42.1% and 81.9% of boundaries detected at the EER operating point were within 5- and 15-ms error tolerance from manual labeling results. They were slightly higher than those of the first-stage RNN detector.

4. Conclusions

In this paper, a two-stage sample-based phone boundary detection algorithm has been discussed. An EER of 7.6% was

reached by the proposed method tested on the TIMIT database. The performance is only 1.3% higher than the HMM phone recognizer with MMAE criterion realignment. Nevertheless, the accuracy of the proposed system is much higher than the HMM recognizer. 42.1% and 81.9% of boundaries detected were within 5- and 15-ms error tolerance from manual labeling results.

5. Acknowledgements

This work was supported by the National Science Council, Taiwan, ROC, under the project with contract NSC 97-2221-E-009-080-MY3.

6. References

[1] D. T. Toledano, L. A. H. Gomez and L. V. Grande, "Automatic phonetic segmentation," *Speech and Audio Processing, IEEE Transactions on* , vol.11, no.6, pp. 617-625, Nov. 2003.

[2] S. Dusan, L. Rabiner, "On the Relation between Maximum Spectral Transition Positions and Phone Boundaries," in *Proc. Interspeech 2006*, pp. 17–21.

[3] 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, Feb. 2009.

[4] You-Yu Lin, Yih-Ru Wang and Yuan-Fu Liao, "Phone Boundary Detection using Sample-based Acoustic Parameters, " Proc. of INTERSPEECH-2010, Makuhari, Japan, pp. 1397-1400, Spet., 2010.

[5] Yih-Ru Wang and Chi-Han Huang, ' Speaker-and-environment Change Detection in Broadcast News using the Common Component GMM-based Divergence Measure, ' , Proc. of ICSLP 2004, Jeju island, Korea, pp. 1069-1072, Oct. 2004.

[6] Yih-Ru Wang, " The Signal Change-point Detection using the High-order Statistics of Log-likelihood Difference Functions, ", Proc. of ICASSP 2008, Las Vegas, USA, pp. 4381-4384, April, 2008.

[7] Sharlene A. Liu, "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am. 100(5), pp. 3417-3430, Nov. 1994.

[8] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *Proc. ICASSP 2004*, pp. 193–196.

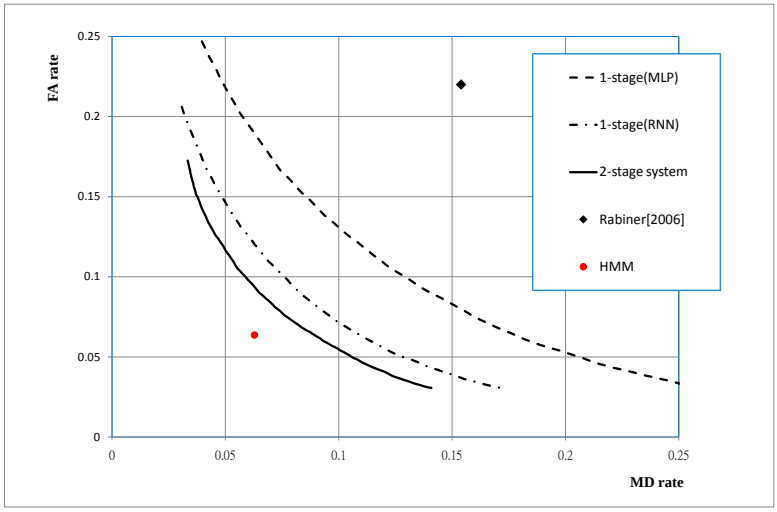


Figure 1 : The MD rate vs. FA rate curves for sample-based phone boundary detectors, two-stage sample-based phone boundary detector and HMM system.

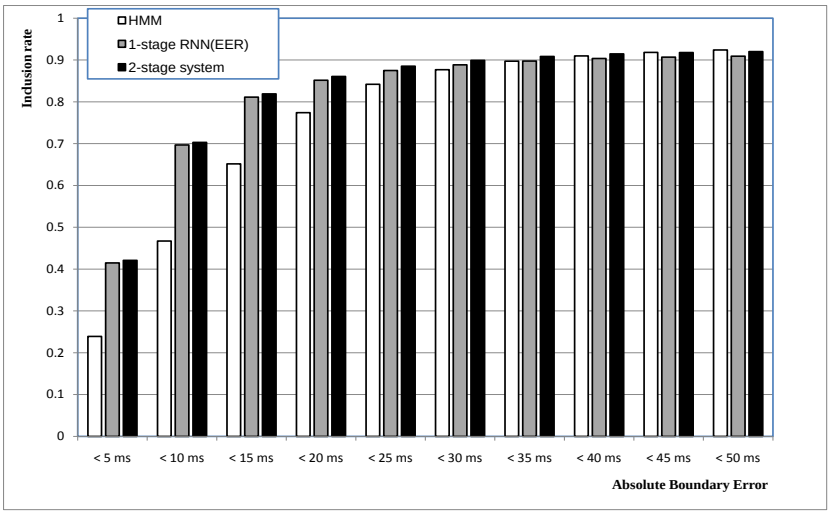


Figure 2 : The normalized cumulative histogram of the absolute deviation between automatically detected boundaries and manually labeled result for difference methods.