高解析度之國語類音素單元端點自動標示

Sample-based Phone-like Unit Automatic Labeling in Mandarin Speech

林宥余　You-Yu Lin
國立交通大學電信工程研究所
Institute of Communication Engineering, National Chiao Tung University
rossi0927.cm97g@g2.nctu.edu.tw
王逸如　Yih-Ru Wang
國立交通大學電信工程研究所
Institute of Communication Engineering, National Chiao Tung University
yrwang@mail.nctu.edu.tw

摘要

在本論文中提出一種以取樣點爲單位(sample-based)的高時間解析度之音素端點自動標示與切割的方法，有別於傳統分析語音信號以音框爲單位(frame-based)或是音段爲單位(segment-based)的研究。本文中，我們提出了一些以取樣點爲單位的聲學參數；由實驗結果顯示，這些聲學參數在不同發音特徵之音素轉換間有明顯的變化率，有利於音素切割位置之標記。我們利用這些發音特徵變化的聲學參數特性，建立一個高時間解析度的自動音素端點標示與切割系統。由TCC-300國語語料庫進行自動端點標示之實驗結果顯示，本論文所提出的方法比傳統以音框爲單位之切割方法，亦即HMM之切割方法，更能有效切出精準的短停頓、摩擦音、塞擦音等之音素端點位置。

Abstract

This paper presents a sample-based phone boundary detection algorithm which can improve the accuracy of phone boundary labeling in speech signal. In the conventional phone labeling method adopted the frame-based approach, some acoustic features, like MFCCs, are used. And, the statistical approaches are employed to find the phone boundary based on these frame-based features. The HMM-based forced alignment method is most frequently used method. The main drawback of the frame-based approach lies in incapability of modeling rapid changes in speech signal; moreover, the time resolution of this approach is too coarse for some applications. To overcome this problem, a sample-wise phone boundary detection framework is proposed in this study. First, some sample-wise acoustic features are proposed which can properly model the variation of speech signal. The simple-based spectral KL distance is first employed for boundary candidates pre-selection in order to reduce the complexity of sample-based methods. Then, a supervised neural network is trained for phone boundary detection. Finally, the effectiveness of the proposed framework has been validated on automatic labeling of TCC-300 speech corpus.

關鍵詞：音素端點切割，帶通信號波封，sample-based 頻譜 KL 距離，監督式類神經網路

Keywords: phone boundary segmentation, sub-band signal envelope, sample-based spectral KL distance, supervised neural network

一、緒論

正確音素切割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，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。在全球有人工切割位置的語料庫不多，最著名的是 TIMIT 語料庫，但是一個大型的連續語音資料庫，使用人工標記切割位置的方式，不僅非常耗時且人工切割的標記位置也伴隨著一個缺點，就是以人工做標記的動作時，會因爲主觀上認定切割位置不同而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切割位置的語料庫是非常重要的。

在語音信號處理中，自動音素之切割是一個非常重要的問題，儘管在過去有非常多自動音素切割的研究[1]，一個具有高精準度的自動音素切割演算法，仍是一個可待持續研究的課題。在過去一些自動音素切割與偵測的研究中，主要可分爲 Model-based 及 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)做自動語音切割，其效能可在正負 20 ms 之內佔有 90%的比率(inclusion rate)，而傳統 HMM 是以整段語句所得到最大相似度函數(maximum likelihood, ML)爲訓練準則，故其自動切割之位置並非爲最佳之音節或音素邊界位置。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)爲訓練準則之 HMM[2]，就使用自動給定之已知端點間誤差最小化作爲 HMM 模型之訓練準則，在 TIMIT 語料庫中，MBE-HMM 自動切割之邊界與人工切割邊界誤差範圍 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；然而其自動切割位置只有 7.89%的邊界在人工切割位置誤差 20 ms 之外。此外，也可使用其它圖形識別的方法如支撐向量機(support vector machine, SVM)[3]、類神經網路(neural network, NN)[4]，來對 HMM 之自動切割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[5]使用頻譜轉換量測(spectral transition measure)的音素邊界偵測方法，應用在 TIMIT 語料庫其效能可達到在誤差 20ms 的容忍範圍內，只有 23.1%的音素端點位置沒偵測出來 (missed detection rate, MD)、22.0% 誤報率(false alarm rate, FA)。Kotropoulos[6]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，其效能在 NTIMIT 語料庫亦可達到 25.7% MD 與 23.3% FA 的結果。

在先前的音素切割方法中，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；且一般假設語音信號在短時間內爲穩定的特性，故使用 frame-based 的聲學參數，例如梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)。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 frame-based 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 frame-based 方法之切割位置與實際正確音素邊界位置之間產生誤差，因此對於自動語音切割之研究提昇時間解析度，必可降低大量因音框之時間解析度所造成的誤差。而語言學家就曾經提出一些用來區別發音特徵的參數，一般稱之爲 Articulation Parameter (AP)。其方法可用低解析度的頻帶，來區分像發音方式或發音位置以及偵測一些 landmark 如 voice on-set，而不是用來辨認像音素的精細分類。由以上敘述，在自動語音切割的應用，我們可以思考爲了使得自動端點標示的時間精確度能夠提昇，降低頻譜精確度的可行性。故在本文中，我們提出 sample-based 音素端點偵測方法的架構，並與 frame-based HMM 切割位置做比較。

在本文中其它章節概要如下：在第二節中，我們首先說明 sample-based 音素端點偵測方法的整體架構；第三節對於本論文中所提出之一些 sample-based 聲學參數的特性做進一步地說明；第四節則是介紹利用上述 sample-based 聲學參數並使用多層感知器(multi-layer perception, MLP)類神經網路架構的 sample-based 音素端點偵測方法；第五節爲實驗結果探討，並於第六節提出簡單的結論。

二、系統架構

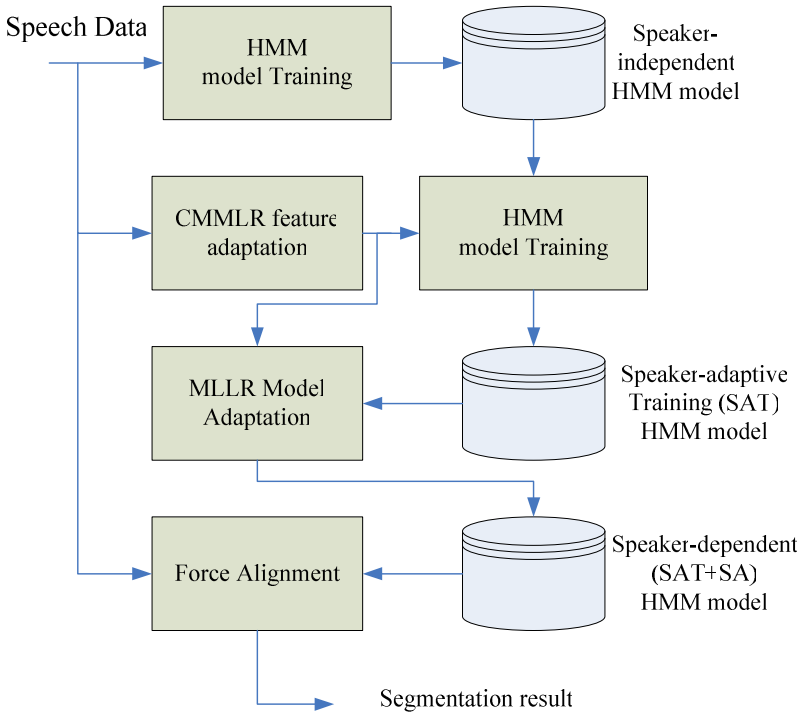
一般傳統切割的方法，主要分成兩個部份，首先利用統計模式爲基礎的方法，如 HMM-based forced alignment 當作初始切割位置，再藉由一些方法如 SVM 等，以進一步修正初始切割位置(refinement)。本研究是對 TCC-300 語料庫做切割，先使用 HMM-based forced alignment 得到初始切割位置；接著，利用 sample-based 聲學參數進一步調整該初始位置；並以 KL distance 挑選其候選端點，訓練一個 MLP 音素端點偵測器以得到最佳之切割位置。由於 TCC-300 語料庫是由不同的語者所組成，所以在取得 HMM-based forced alignment 初始切割位置時，我們使用了語者調適的技術調將 HMM 模型調適成更適合該語者之模型。接下來我們進一步介紹語者調適的流程以及 MLP 音素端點偵測器。

（一）、使用 SAT 及 SA 技術之 HMM phone-like unit alignment 流程

我們將使用下列流程做 TCC-300 語音資料庫 HMM 模型類音素層級(phone-like level)之起始切割位置，就是將一個音節區分爲聲母、介音、韻母及韻尾鼻音等部分，其方塊圖如下：

在 HMM phone model training 後，我們再使用做 speaker adaptation training(SAT)；SAT 就是使用 constraint MLLR(CMLLR)對不同語者做語音參數的轉換；使用經語者轉換(CMLLR)後之語音參數再重新訓練新的 HMM 模型將可獲得較佳之 speaker-dependent HMM 模型。做完 SAT 後，我們再做 HMM 做 model adaptation，使用 MLLR 技術來調

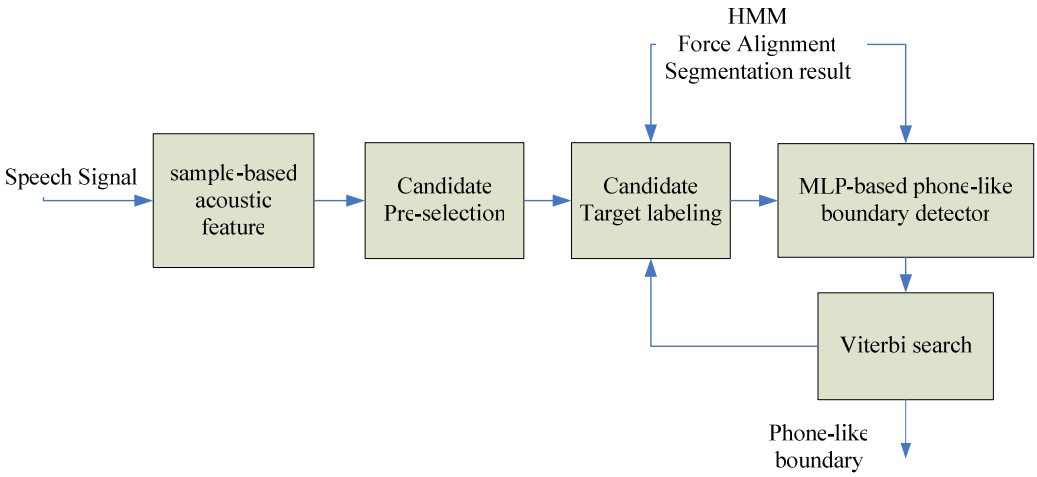
適 HMM 模型，它和 SAT 會有加成性的效果。如此就可以獲得較佳的 HMM 模型來做強制對齊(force alignment)，作爲 TCC-300 語料庫之音素的起始切割位置。



圖一、使用 SAT 及 SA 技術之 HMM　alignment 流程

（二）、MLP 音素端點偵測器訓練流程

MLP 類神經網路被廣泛地運用在各個領域當中作爲資料分類的架構，同時因爲其自我調適的能力、非線性的運算、具有學習能力等特性，故本研究使用此架構訓練一個監督式(supervised)MLP 音素端點偵測器。訓練音素端點偵測器流程，其方塊圖如圖二。



圖二、使用 MLP 架構之類音素端點偵測流程

在圖二中，因經由 HMM alignment 獲得之初始切割位置仍不夠準確，故我們利用 sample-based 的聲學參數所提供之資訊來得到較好的切割位置以作爲訓練 MLP 音素端點偵測器時之答案。且爲減少計算量，由預選擇(pre-selection)即簡單設定一個臨界值的方法來挑選較爲可能之候選端點(candidate)位置。接著將候選端點依目標函數(target function)分類後，訓練 MLP 音素端點偵測器直至收斂，最後使用 Viterbi search 演算法在候選端點中得到該語句最佳之切割位置。

三、sample-based 聲學參數的特性

首先，本研究結合語言學家所提出的 AP，利用數個頻段來區分不同發音特徵之方法，應用於切割語音信號可提高時間解析度由音框進一步精準至取樣點，並在此提出一些 sample-based 的 AP 以用於描述不同語音屬性變化時的 AP 特性，來調整音素切割位置之標記。在此節中將介紹本論文所提出之 sample-based 聲學參數及其在音素端點偵測上之特性。

（一）、Sample-based 聲學參數

我們提出一些 sample-based 的 AP 如帶通信號波封(sub-band signal envelope)、參數上升率(rate of rise, ROR)、頻譜熵(spectral entropy) 、sample-based spectral KL distance 及 spectral flatness，並觀察它們在不同語音屬性，如爆破音、鼻音、靜音等特性。以下，我們進一步介紹本研究所使用的語音特徵參數：

1、帶通信號波封[7]

在語言學家所提出的 AP 中，有許多帶通濾波器，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(filter bank)[7]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，可視爲是語音信號開始改變的地方。但語言學家所使用的 AP 爲帶通信號波封，而非現今語音辨認器中常用的能量。故我們將這六個頻段之語音信號取出它的波封來當作本研究中所使用的聲學參數。

我們在製作一個波封檢測器(envelope detector)時，爲了保持波封變化快的時候能正確地找到信號的波封，我們使用希爾伯特變換(Hilbert transform)後再經低通濾波器，求取輸入信號的波封，一個信號 $x[n]$ 的希爾伯特變換 $H(x[n])$ 的希爾伯特變換，如下式：

$$H(x[n]) = x[n] \otimes h[n] \quad \text{and} \quad h[n] = \begin{cases} 0, & n \text{ is even} \\ 1/n\pi, & n \text{ is odd} \end{cases} \tag{1}$$

2、上升率[7]

語言學家所稱之上升率，就是在 frame-based 的語音特徵參數中所用的 delta-term：

$$ROR_x[n] = \left(\sum_{i=-w}^{w} i \cdot x[n+i] \right) / \left(\sum_{i=-w}^{w} i^2 \right) \tag{2}$$

其中 $x[n+i]$為輸入參數資料，w 為求上升率所使用的音框寬度。本研究使用波封的上升率、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，來描述各 sample-based 聲學參數的變化率。

3、頻譜熵 [9-10]

頻譜熵可用來描述信號在頻譜上的集中程度，若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 6 個頻段，則頻譜熵 H_s 可以定義如下式表示：

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

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

其中 $E_i[n]$為第 i 個頻段之第 n 點正規化之後的波封。由頻譜熵對應到語音信號上，可以發現短停頓類似於雜訊，在各個頻段都會出現，所以頻譜熵值較高；而韻母在頻譜上的能量則較集中於低頻至中頻的部分，其頻譜熵值相對較低。

4、Sample-based spectral KL distance

將頻譜視爲一個機率分佈，因此可用 KL distance 來描述頻譜上的相似程度。在語音信號中計算兩點不同時間(m 與 n)的 spectral KL distance，$d_x(m,n)$，可以由下式表示：

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

以上所敘述的參數頻譜熵、頻譜熵的上升率、sample-based KL distance 來觀察一段語音信號其語音特徵的變化，這些語音特徵證實可以分辨不同語音屬性的邊界。

5、Spectral flatness[11]

使用正規化後之帶通信號波封計算的 flatness，F，表示如下式：

$$F = \frac{\left(\prod_{i=1}^{6}\frac{E_i[n]}{s_i}\right)^{1/6}}{\frac{1}{6}\left(\sum_{i=1}^{6}\frac{E_i[n]}{s_i}\right)} \tag{6}$$

其中 S_i 爲第 i 個頻段靜音信號正規化後波封的平均。若信號爲靜音(silence) 或是短停頓(short pause)，則 F 將會趨近於 1。若 spectral flatness 與波封等參數經過設定適當的臨界值(thresholds)，對於標記靜音及短停頓的切割位置時是一個有效的參數。

（二）、Sample-based 聲學參數之語音特徵

在此我們將觀察 sample-based 聲學參數語音特徵之特性與類音素端點間之關係，並以實例證實先前我們所提出之 sample-based 聲學參數具有正確偵測類音素端點間端點的能力與特性。

以下使用 TCC-300 麥克風語料庫來做爲觀察 sample-based 聲學參數之語音特徵的

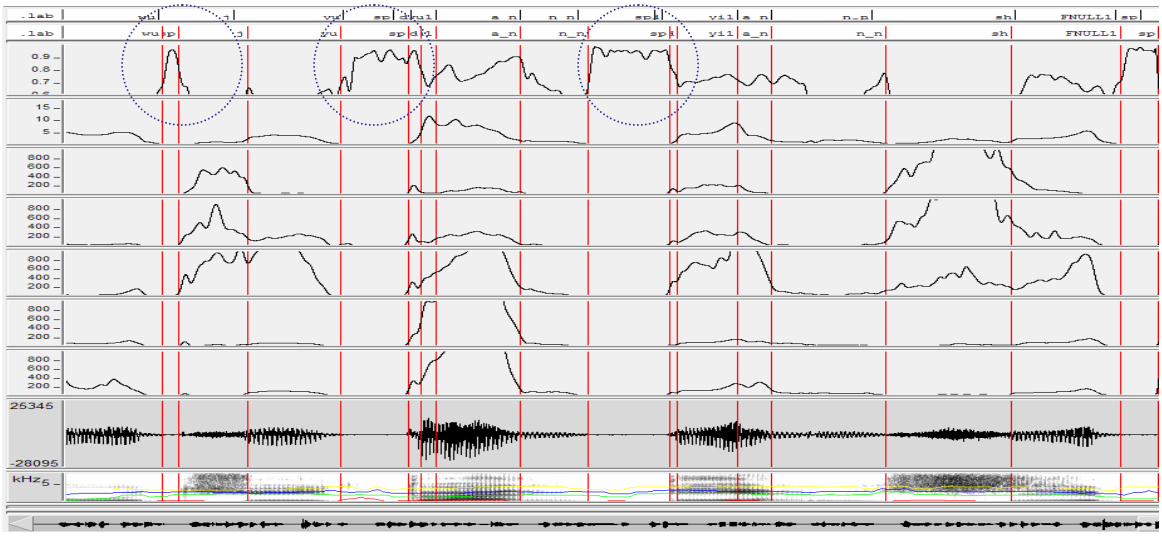
語料，在觀察中我們利用先前由 SAT 及 SA 技術之 HMM phone-like unit alignment 所獲得之切割位置作爲比較對象。

首先，利用 SAT(speaker adaptation transform, feature MLLR)及 SA(speaker adaptation, MLLR)後的語者調適 HMM 模型來得到 TCC-300 的類音素單元之切割位置，接著我們利用此新切割位置以語音屬性的不同做分類，如表一。由新切割位置當做參考位置利用 sample-based 的聲學參數特性觀察是否可用來調整音素端點的位置。由於先前語者調適 HMM 之切割位置，已近乎準確，但是仍有更進一步修正的空間，故我們提出以 sample-based 音素端點偵測的方式以期達到更爲精確的切割位置。

表 1、國語語音發音方法的分類表。

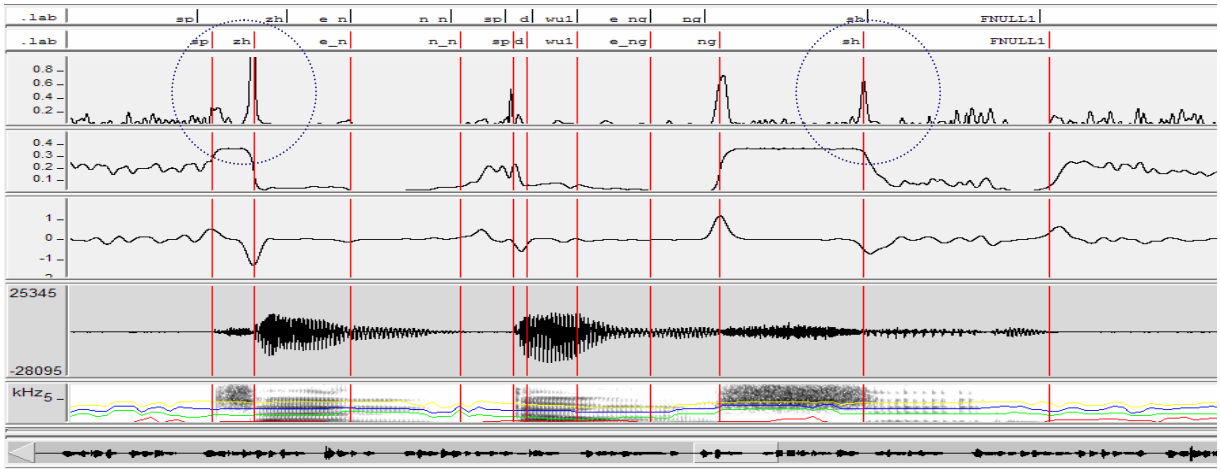
爆破音 Stop	b	p	d	t	g	k
鼻音 Nasal	m	n	(n_n)	(ng)		
摩擦音 Fricative	f	s	x	h	sh	
塞擦音 Affricate	q	j	c	z	zh	ch
流音 Liquid	l	r				
韻母音 Vowel	others					

先前觀察 HMM 自動切割位置的標記時，發現短停頓常無法標記出來，而使得塞擦音與爆破音等音素平均音長過長的現象。如圖一，在這裡我們使用 spectral flatness、波封以及各頻段之信號波封來判斷是否爲短停頓的狀態。在短停頓與爆破音及塞擦音的交界處，短停頓在各個頻段之信號波封與其它有語音信號的地方相比幾乎很低，且 flatness 趨近於 1，波封可與 flatness 產生互補的效果來標記短停頓的端點。



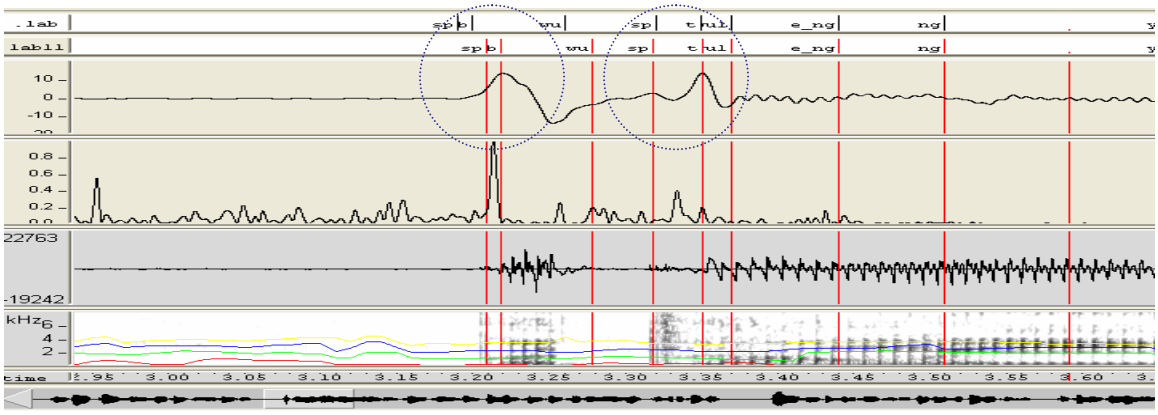
圖三、國語語句觀察短停頓切音位置之例子，由上至下的圖形分別表示 Spectral flatness、波形之波封、第 6 個至第 2 個頻段的信號波封、波形、頻譜。（最上方兩列標音位置分別表示是原語者調適 HMM 切割位置及修正後之切割位置）

接下來我們觀察摩擦音、塞擦音等聲母，它們在頻譜中與相鄰韻母與短停頓有極大的頻譜差異。在此我們使用 spectral KL distance、頻譜熵以及頻譜熵的上升率來偵測音素的端點。如圖四所示之圓圈圈選處中，我們可以看到與摩擦音及塞擦音相鄰頻譜之差異非常大，而 spectral KL distance 在摩擦音、塞擦音等聲母接續至韻母或是韻母轉換至摩擦音、塞擦音之情形有相對其他部分有較高的峰值。且摩擦音、塞擦音相鄰韻母的端點，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我們所期望的正確端點位置差距不遠，我們可以了解頻譜熵、 KL distance 等已知在 frame-based 偵測信號變化量是非常有用之聲學參數，同樣在 sample-based 的效果一樣明顯，且標記之切割位置更精準。



圖四、國語語句觀察摩擦音、塞擦音切音位置之例子，由上至下的圖形分別表示 sample-based KL distance、頻譜熵、頻譜熵上升率、波形、頻譜。（最上方兩列標音位置分別表示是原語者調適 HMM 切割位置及修正後之切割位置）

爆破音切割位置的修正時，由波形與頻譜觀察中，我們可以發現通常在爆破音開始的時候會有短停頓出現，波封接著會有急遽上升的現象，故我們使用波封之上升率來描述其現象。如圖 6 所示，在爆破音結束的地方，也是音素轉換的端點。



圖五、國語語句觀察爆破音切音位置之例子，由上至下的圖形分別表示波封上升率、sample-based KL distance、波形、頻譜。（最上方兩列標音位置分別表示是原語者調適 HMM 切割位置及修正後之切割位置）

鼻音部分我們可發現信號在頻譜上多集中在0.0 – 0.4 KHz 與0.8 – 1.5 KHz 的低頻頻段，與相鄰的音素皆有頻譜上的差異，在此我們也使用 spectral KL distance 來判斷。而韻母端點的偵測，是利用相鄰聲母及短停頓之端點位置，當做韻母的邊界切割位置。

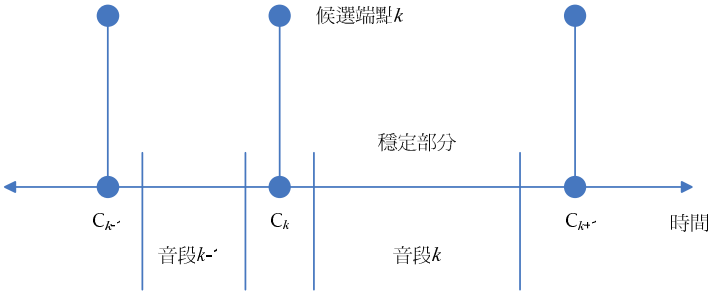
四、使用 MLP 類神經網路架構的 Sample-based 音素端點偵測方法

在前一節的觀察中已證實本論文中所提出之 sample-based 聲學參數有精確偵測類音素端點的能力。在這一節我們將使用這些參數來製作一個監督式(supervised)的 MLP 類神經網路模型來作爲音素端點之偵測器。

我們從語音信號中抽取 sample-based 聲學參數之後，爲了減少在端點偵測器中過於龐大的資料計算量，經由預選擇即簡單設定一個臨界值方法來挑選較爲可能之候選端點位置；而當語音信號在頻譜中的變化量大時，spectral KL distance 是一種很好的測量方式，故若 spectral KL distance 滿足下式：

$$d_x(n-1,n) < d_x(n,n+1),\ d_x(n,n+1) > d_x(n+1,n+2) \text{ and } d_x(n,n+1) > Th_d \tag{7}$$

則代表爲挑選出來的候選端點，最後我們得到這一連串候選端點的序列，$\{c_j; j=1,...,N_c\}$。經過預選擇步驟後，候選端點會將語音信號分割成很多音段(segment)，我們也可由這些語音信號的音段求取一些 segment-based 之聲學參數來協助端點偵測，如圖六所示。



圖六、利用候選端點將語音信號分割成音段的示意圖

在此，我們使用音段中正規化後的各個頻段之信號波封平均值來評斷相鄰 2 個音段$[c_{k-1},c_k]$、$[c_k,c_{k+1}]$之語音特性。其中 $ES_i(k)$爲在第 k 個音段($[c_k,c_{k+1}]$)中正規化後的各個頻段之信號波封平均值，可定義成下式：

$$ES_i(k) = \left(\sum_{n=c_{k-1}+1}^{c_k-1} E_i[n] \right) / \left(c_k - c_{k-1} - 2 \right) \tag{8}$$

接著，我們對於每個候選端點建立一個 30 維的參數向量(feature vector)，對於第 k 個候選端點，c_k，其參數向量包括以下聲學參數，

(1) 目前候選端點之參數：

$$\{fb_i[k],\ ror\_fb_i[k],\ d_k(n,n+1),\ F(k),\ env(k),\ ror\_env(k),\ H_s(k),\ ror\_H_s(k)\}; i=1,...,6 \tag{9}$$

(2) 目前候選端點前($ESp_i(k)$)、後($ESn_i(k)$)音段之參數：

$$\{ESp_i(k)=\left(\sum_{n=c_{k-1}+1}^{c_k-1}E_i[n]\right)/\left(c_k-c_{k-1}-2\right),\ ESn_i(k)=\left(\sum_{n=c_k+1}^{c_{k+1}-1}E_i[n]\right)/\left(c_{k+1}-c_k-2\right)\};i=1,...,6 \tag{10}$$

最後使用 2 個指標指出此候選端點是否爲此候選端點序列之第一個或最後一個端點。

由先前所述的方法，我們已利用 sample-based 聲學參數的調整得到較佳之切割位置以訓練 MLP 端點偵測器。然而最重要的問題是要如何決定 MLP 音素端點偵測器之目標函數。由於所使用的參數不但能描述波形變化，也能使用帶通信號波封來辨別語音之發音特性。因此，我們定義 9 大類的目標函數，分別表示候選端點出現在短停頓(short pause, IS)、聲母(consonant, IC)、韻母(vowel, IV)、韻尾鼻音(nasal endings, IN) 4 種分類與彼此分類的轉換點，短停頓變化至聲母或是韻母(SC)、聲母至韻母(CV)、韻母接韻尾鼻音(VN)、韻母或韻尾鼻音變化至短停頓(VS)和可略過短停頓轉換點(CP)，而目標函數之間的轉移機率由目標函數所產生的 likelihood 作正規化後計算。

在我們的系統中，應用於 MLP 音素端點偵測之訓練演算法的流程。我們利用程式調整過後的 HMM 切割位置，再經過預選擇挑選出來的所有候選端點，進行分類標記來訓練 MLP 音素端點偵測器。

訓練演算法的過程如下：

(1) 經過分類標記的所有候選端點，當做初始目標函數；
(2) 利用給定的目標函數來訓練 MLP-based 音素端點偵測器直至收斂；
(3) 由 MLP-based 偵測器輸出目標函數的 likelihood 來計算候選端點之轉移機率，並依照其轉移機率使用 Viterbi search 得到最佳路徑，再重新進行候選端點的分類標記並將其標記結果當作 MLP 端點偵測器新的目標函數；
(4) 重複(2)與(3)的步驟，直至收斂。

五、Sample-based 音素端點偵測方法實驗結果

本章節主要是將我們所提出的 sample-based 音素端點偵測方法，運用 MLP 類神經網路的架構訓練一個音素端點偵測器，觀察並分析其切割位置之結果。同時以 frame-based HMM 架構之切割位置來比較其結果，觀察本研究所提出的方法切割位置之精準度是否有進一步地提升。

（一）、語音資料庫簡介

本文中的自動語音切割實驗所使用的 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。其中台灣大學語料庫主要包含詞以及短句，文字經過設計，考慮了音節及其相連出現機率，由 100 人錄製而成；成功大學及交通大學主要包含長文語料，文章由中研院提供之 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 3-4 段，每段至多 231 字，由 200 人朗讀錄製，每人所朗讀之文章皆不相同。語音的取樣頻率爲 16kHz，取樣位元數爲 16 位元。

（二）、實驗驗環境與實驗架構設定

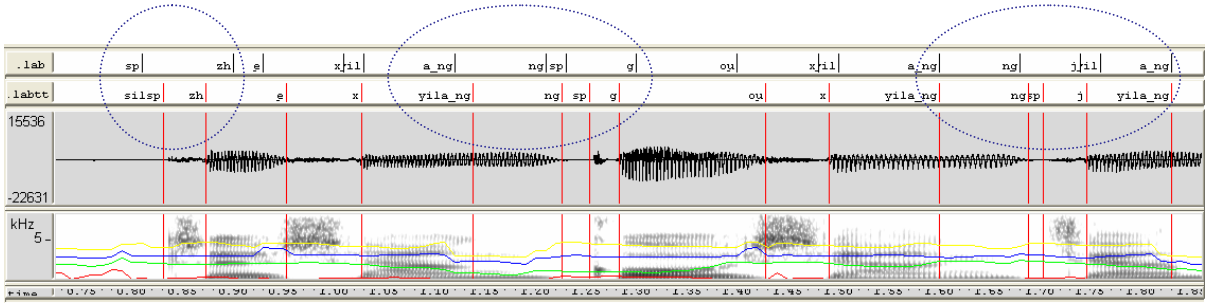
在 TCC-300 語音資料庫之語料選取方面，本論文使用交通大學與成功大學所錄製的長文語料，並隨機選取六分之五的部份當作訓練語料，其它部分爲測試語料。首先使用 SAT 及 SA 技術之 HMM phone alignment 流程，獲得較佳的 HMM 模型後進行強迫對齊之切割結果，作爲 TCC-300 語料庫之類音素起始切割位置。

在本研究使用 NICO Toolkit[12]來訓練我們的 MLP 端點偵測器，並採用 30×50×9 的 MLP 類神經網路架構分爲 3 層，包含一個輸入層、一個隱藏層、一個輸出層，輸入層點數共 30 點，分別輸入 6 個頻段之信號波封及其波封之上升率、頻譜熵及其上升率，波形之波封及其上升率、sample-based spectral KL distance、spectral flatness，與前、後音段正規化 6 個頻段信號波封後的平均值等參數。

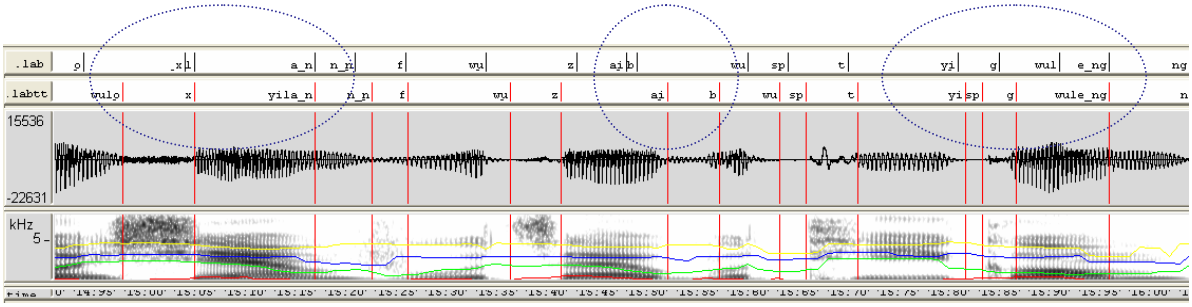
獲得 MLP 端點偵測器後，我們可以使用 HMM 切割位置作爲初始切割位置再以 sample-based MLP 類音素端點偵測器來做更精確之切割。我們將 sample-based MLP 類音素端點偵測器偵測之端點限制於 HMM 初始切割之正負 100 ms 範圍內，亦即起始切割位置不須十分精確，我們都可以找到正確端點。因此我們利用一個加上限制範圍的 Viterbi search 方法來獲得新的切割位置。

（三）、MLP 音素端點偵測器實驗結果比較與分析

在此實驗中我們觀察 Viterbi search 限制範圍是 100 ms 之音素端點偵測結果，並列舉數個圖形比較音素端點偵測結果與 HMM 語者調適模型強迫對齊的切割位置。可以由下列圖八、圖九之中看到 HMM 強迫對齊切割位置之結果，對於音素的切割位置經常有誤差存在，尤其是聲母之前的短停頓之切割效果不好或是沒有切出來，造成聲母長度普遍變長之現象，同時其聲母與韻母之間的切割位置亦不甚理想。我們同樣可由圓圈圈選處之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。圖八所示之橢圓形圈選處，我們亦可發現在韻母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，由上述結果皆可證實其 sample-based 的聲學參數具有偵測發音特徵變化之效能。



圖八、國語語句音素端點偵測之例子，由上至下的圖形分別表示原語者調適 HMM 切割位置及音素端點偵測之切割位置、波形、頻譜



圖九、國語語句音素端點偵測之例子，由上至下的圖形分別表示原語者調適 HMM 切割位置及音素端點偵測之切割位置、波形、頻譜

在此我們統計了 HMM 模型強迫對齊切割各發音方法之平均音長，並與我們提出的方法作比較，且本研究之音素端點偵測器將介音與韻母合併視爲單一韻母偵測，故韻母之平均音長不作比較。而由各發音方法之平均音長，觀察發現 HMM 語者調適模型之結果(表二)較音素端點偵測器之平均音長(表三)平均結果皆多出 10-20ms 以上的範圍，其原因在於 HMM 音素端點之切割位置皆有誤差，而 sample-based 音素端點偵測器皆能準確地將短停頓的位置標記出來使得聲母之平均音長下降，特別是爆破音與流音之平均音長下降 20-30ms 以上，明顯地較 HMM 切割位置之平均音長更加符合合理的範圍。

表二、HMM 語者調適模型切割位置各發音方法平均音長

單位： 10ms 發音方法	各發音方法平均音長
爆破音 Stop	4.96
鼻音 Nasal	5.95
摩擦音 Fricative	11.13
塞擦音 Affricate	8.92
流音 Liquid	6.23

表三、MLP 自動端點標示各發音方法平均音長

單位： 10ms 發音方法	各發音方法平均音長
爆破音 Stop	2.62
鼻音 Nasal	4.46
摩擦音 Fricative	8.75
塞擦音 Affricate	7.13
流音 Liquid	2.70

六、結論

本篇論文在 TCC-300 語音資料庫無正確音素人工標示資訊下，第一階段使用 SAT 及 SA 技術之 HMM phone alignment 流程，獲得較佳的 HMM 之切割位置資訊，作爲

TCC-300 語料庫之音素起始切割位置；在第二階段我們提出 MLP-based 音素端點偵測器的架構並加入數個 sample-based 的聲學參數對語料庫做自動化類音素單元之端點標示工作。實驗結果顯示，由於 HMM 切割位置的不準確會造成聲母過長或是無法正確切割出短停頓的情形均有大幅改善，證實這些以往使用於 frame-based 之聲學參數在 sample-based 的應用上也確實有顯著的效果。語音信號對於發音方式的不同會有不同的特性，而語音屬性應該是與語言無關的，因此在未來我們即可利用此性質對國內經常使用的聲調語言類型如閩南語、客語等方言來進行跨語言的自動端點標示的工作，並將此架構應用於有人工切音位置之 TIMIT 語料庫以評估此方法之效能。

七、參考文獻

[1] 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.

[2] J. -W Kuo and H.-M Wang, "Minimum Boundary Error Training for Automatic Phonetic Segmentation," *The Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)*, September 2006.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] Hasegawa-Johnson, etc. "Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop," *Acoustics, Speech, and Signal Processing, 2005. ICASSP 2005.* vol.1, no., pp. 213-216, March 18-23, 2005

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

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

[11] J. D. Markel and A. H. Gray, "A spectral-flatness measure for studying the autocorrelation method of linear prediction speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 207-217, June 1974.

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