行政院國家科學委員會專題研究計畫成果報告

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(2/3)

計畫編號：97-2221-E-009-080-MY3
執行期限：98 年 8 月 1 日至 99 年 7 月 31 日
主　持　人：王逸如　國立交通大學電信工程系
共同主持人：廖元甫　國立台北大學電子工程系
王新民　中央研究院資訊所

一、 中文摘要

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。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。

在本計畫中將以精確的偵測語音信號中的發音特徵變化點(landmark)為起點，在地年研究中使用 sample-based 的語音參數(acoustic feature, AP)來做音素端點偵測，在第二年進行下列研究：

我們對 sample-based 的語音參數在音素端點偵測器作了些進一步研究以增進系統效能。並在國語及客語語料庫無正確音素人工標示資訊下，分別已對 TCC-300 語料庫做自動化類音素端點標示工作，並製作出國語語音之發音方式之偵測器並對客語語料庫作子母音單元的自動端點標示及觀察偵測音素邊界的結果。

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

Abstract

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 year, the proposed sample-based phone boundary detector was further improved. The performance of the sample-based phone boundary detector was checked in the TIMIT database. The sample-based phone boundary alignment algorithm was also used in TCC-300 Mandarin speech dataset and Hakka speech database. The results show the algorithm can get phone boundary with better precision and accuracy.

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

二、 緣由與目的

回顧現今自動語音辨識技術，大詞彙的連續語音辨識(large vocabulary continuous speech recognition, LVCSR)技術被開發出來，所依賴的就是大量的語音資料與語言資料。但大家發現現有的這些技術還是不夠好，仍無法與人類辨識語音的能力相比，而現有技術的進步空間有限，為了將來語音辨識技術的發展，近年國際上已不斷有學者主張，應該回頭將語音與語言的知識帶進來，建立一個以知識為基礎 (knowledge-based) 加上資料驅動的(data-driven)模式，開放測試平台，共享一個合作的設計與評量機制，將自動語音辨認推向新一代的技術[1]。在[1]中，為新一代自動語音辨識技術建立之平台及架構圖如圖 1 所示。

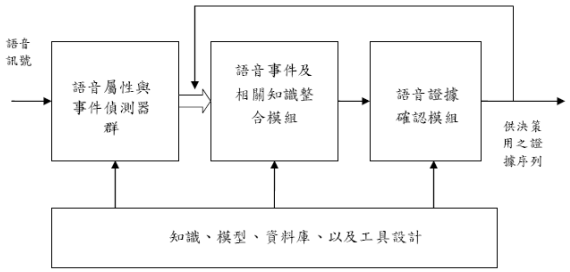

圖 1、新世代自動語音辨識技術架構圖。

在新世代語音辨認系統的語音辨認架構下，我們想由語音信號我們直接找尋一連串各階層單元及時序信號，也就是所謂的attribution及landmark；其中landmark信號可提供語音信號中各層次的部分時序資訊，如：phone、syllabic、word、prosodic phrase及utterance中的特定信號，例如：語言學家所發現的一些landmark 就是語音信號中attribution改變的地方，所以可以提供語音信號中之時序資訊。而attribution可以提供各段語音信號之articulation properties，於是語音信號就可已變成一連串的events，語音辨認就可以視為由一連串的articulation events中做解碼(decoding)。所以由語音信號中尋找一連串的時序資訊就變得十分的重要，它不像傳統的HMM架構的語音辨認架構必須在解碼後才可獲得語音信號的時序資訊(也就是各個辨認單元的時間位置)。在語音信號中phone、syllable到word等各階層單元中；較低層的phone單元的時序資訊，即使對語音學家而言，都常有不確定性。對國語及台灣常用的兩種方言－閩南語及客家語均為syllabic language，word的結構也有很高的不確定性。所以對像國語、閩南語及客家語而言，syllable就變成最可告的時序信號了。而能精確的偵測出音節階層事件如音節端點，將可進行許多關於syllabic language的temporal structure之研究。

三、研究方法

在本計畫中，我們所提出的新世代語音辨認技術中的事件偵測及語音屬性偵測，基於上述觀點，所採取的步驟就以偵測 Articulation change 也就是所謂的 landmark 做為系統的第一級。根據 landmark 的定義，我們所做的 syllable-level landmark detector 是以 sample-based 的觀點來製作，希望達到較高的精確度(1-2 msec)。第一年中，我們以 sample-based 的方法製作了 syllable boundary detector。

在過去有一些 phone boundaries segmentation/detection 研究中[2-6]都是使用傳統語音辨認所使用的語音特徵參數。但對 phone boundary 的 landmark 偵測工作不像傳統的語音辨認需要使用在頻率上有高解析度的特徵參數，如 MFCC，為了頻率高解析度而使用 frame-based 架構會使得時間解析度降低。在 landmark 偵測時所需要的是高時間解析度，因為僅需偵測發音方法或部位的變化，所以可以降低頻率上之解析度。所以計畫中我們使用了一些具有高時間解析度的 sample-based 語音參數(acoustic feature, AP)來製作高解析度音節及其相關端點之偵測器(syllable-level boundary detector)。

1. 英語語料之音素端點偵測器效能之改進

因為現有之國語語料庫均無人工標示資訊；所以先前我們使用已有人工標示資訊的英語語料庫，製作一個英語語料庫之音素端點偵測器，並證實所提出 sample-based 聲學參數對音素端點偵測結果，確實提升其音素邊界偵測準確度之效能。本年度研究改進先前所提出之 sample-based 聲學參數對音素端點偵測的架構；我們將兩個候選端點間之語音段視為穩定的語音訊號，求取 segmental-based feature 用以表達該語音段的聲學特性，依此提升音素端點的鑑別度來達到增進端點偵測效能的結果。

1.1 改良 Sample-based 音素端點偵測架構

Sample-based 音素端點偵測架構中，我們使用 6 個 sub-band signal envelope，並利用 spectral KL distance 來先挑出較適當作為端點的位置，訓練一個 supervised neural network 作為第一級音素端點偵測器；藉由調整偵測器臨限值來得到一序列偵測到之音素端點，由此抽取 segmental-based feature，輸入第二級音素端點偵測器。

首先，我們改良第一年計畫中之 AP，將 6 個 sub-band signal envelope 輸出加上一個 threshold,

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

如此預選擇時在靜音時的候選端點會減少。

我們從語音信號中抽取 sample-based 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(pre-selection)即簡單設定一個臨限值(Threshold)方法來挑選較為可能之端點位置之候選者；由於

spectral KL distance 挑選出在語音信號鄰近時間中的變化是一種很好的測量方式，故若 spectral KL distance 滿足下式

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

則代表為挑選出來的候選端點值，最後我們得到這一連串候選端點值的序列，$\{c_j; j=1\cdots,N_c\}$。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)，我們也可由這些音段語音信號求取一些 segment-based 的 feature 來協助端點偵測。首先，我們使用 segmental sub-band signal envelope 來評斷相鄰 2 個音段 $[c_{k-1},c_k]$、$[c_k,c_{k+1}]$，如圖 2 所示。

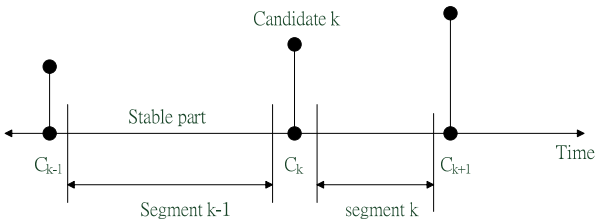

圖 2、利用候選端點將語音信號分割成片段的示意圖。

$ES_i(k)$ 為在第 k 個片段 ($[c_{k-1},c_k]$) 中 sub-band signal envelope 正規化後的平均值，也就是

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

接著，我們對於每個候選端點建立一個 38 維的 feature vector，對於第 k 個候選端點，c_k，其 feature vector 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：
$\left(d_{KL}[c_j], H[c_j], \Delta H[c_j]\right); j=k-1,k,k+1,$
$\left(E_i[c_k], \Delta E_i[c_k]; i=1,\cdots,6\right), \Delta E_0[c_k]$
，其中 $\Delta H[c_j]$ 與 $\Delta E_i[c_k]$ 為 spectral entropy 與第 i 個正規化後 sub-band signal envelope 的差量，

(2) 目前片段及前、後片段之參數：
$\left(ES_i[c_{k-1},c_k], ES_i[c_k,c_{k+1}]; i=1,\cdots,6\right), c_k - c_{k-1}, c_{k+1} - c_k$，

(3) 使用 2 個指標指出此候選端點是否是此候選端點序列之第一個或最後一個端點。

接著我們在系統中訓練一個 supervised multi-layer perception(MLP)作為音素端點偵測器。然而最重要的問題是要如何決定 MLP 分類器之目標函數。雖然 TIMIT 語料庫中有人工標記的端點位置，但與 objective measure 判斷的端點沒有一致性，故利用 MLP 來學習聲學參數的統計特性，來判斷該候選端點是否為音素端點的可能性。

在我們的系統中，應用於音素端點偵測之一個 iterative target selection 與 MLP 訓練演算法發展如圖 3。對於每個人工標記的端點位置，m_i，在此區間 $[m_i-\Delta, m_i+\Delta]$ 中，選擇其中一個候選端點當作目標，其中 Δ 代表人工標記與自動標記之中可容許的範圍。

訓練演算法的過程如下：

(1) 從此區間之候選端點中挑選出擁有最大 spectral KL distance($d_x(c_i,c_i+1)$)之端點，當作初始目標；

(2) 利用給定的目標函數來訓練 MLP-based 音素端點偵測器；

(3) 選擇由 MLP-based 偵測器輸出之最有可能的候選端點當作 MLP-based 偵測器的新目標；

(4) 重複(2)與(3)的步驟，直至收斂。

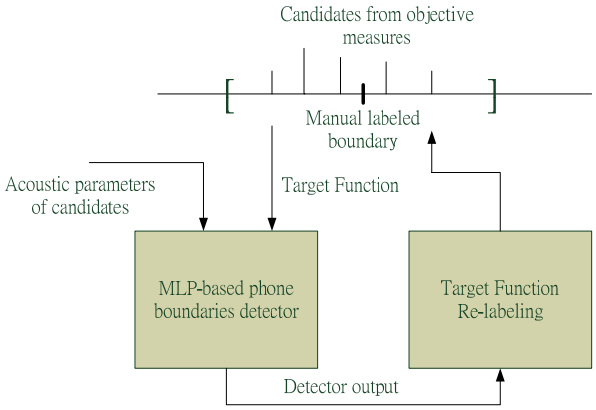

圖 3、Iterative target selection 與 MLP training algorithm 之方塊圖。

由於我們在 sample-based 聲學參數中之六個 sub-band signal envelope 均為描述語音信號短時間的特性，故在第二級我們可加入長時間的語音參數(long-term)，並於候選端點前後之音段求取線性預測倒頻譜係數(Linear Predictive Cepstral Coefficient, LPCC)來描述語音信號在語音段中頻譜更精細的部分。倒頻譜係數是將取對數頻譜的傅立葉轉換得到，然而我們可以使用簡單的遞迴運算來更有效率地利用線性預測參數(Linear Predictive Coefficient, LPC)求取 LPCC，如(3)式

$$c(n) = a(n) + \sum_{k=1}^{n-1} \left[\frac{k}{n} \right] c(k) a(n-k) \qquad (3)$$

其中 $c(n)$ 為第 n 階線性預測倒頻譜係數，$a(n)$ 為第 n 階線性預測係數，$n=1,2,...,P$.

我們利用候選端點相鄰的音段求取 16 階的 LPCC 並且計算其歐幾里得距離(Euclidean

distance)量測兩音段的頻譜相似程度。我們將此距離參數加入第二級的音素偵測器中。

1.2 改良 Sample-based 音素端點偵測架構之實驗與結果

TIMIT 語料庫在此用來確認我們所提出的 sample-based 音素端點偵測演算法的效果。TIMIT 中對應於訓練語料與測試語料的音素端點總數分別為 172460 與 62465。對應於訓練與測試的 sample 總數分別為 $2.27*10^8$ 與 $8.29*10^7$。平均來說，在一秒內會有 12.2 個音素端點，或是每 1310 samples 一個端點。

在我們實驗中，經由適當地選擇臨限值來挑選候選的端點，來減低計算量。總體來說，對應於訓練語料與測試語料分別有 377998 個與 136452 個音素端點被挑選出來。只有 0.16%語音取樣點被當作候選端點，且將這些端點輸入之後的 MLP-based 端點偵測器。MLP-based 端點偵測器之訓練過程如上節所述，且隱藏層神經元數目設成 75。

偵測器的輸出可以用一個臨限值來決定音素端點之偵測結果。訓練與測試資料的 MD vs. FA 比例曲線如圖 4 所示。圖 4 中，FA 定義為(number of false alarm)/(number of actual boundaries + number of false alarm)。且對於訓練與測試語料分別有 12.9%與 14% EERs。如下圖所示， 15.4% MD 與 22%FA 比率，相比於[4]之效能，大約達到下降 4% EER。

自動端點偵測與人工標記端點之間的 MSE (mean-squared error)如圖 5 所示。我們可以看到若自動偵測端點結果的 confidence 越高，則他們之間的差異就越小。也就是說，若將 MLP-based 偵測器的輸出使用一個較高的臨限值，則會擁有較高的端點偵測之信心區間，且可達到較小的錯誤率。

用我們的系統所偵測出來的端點中，有 42.94%是與人工標記的切割位置相距不到 80 個 samples；有 88.07%在±240 個 samples 的範圍之內。而在[4]中，有 27%的端點是在人工標記端點的同一個 frame 內，且 70%偵測出來的端點在差距 1 個 frame(10ms)內，由此可見我們所提出方法之準確度較高。

同時我們不僅已將上述自動學習及端點調整方法用於國語 TCC-300 語料庫做自動 phone-like unit 端點標記工作，也用於客家話語料庫來進行實驗結果分析。

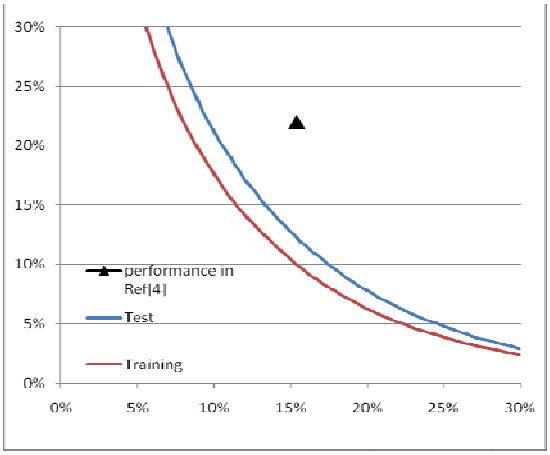

圖 4、所提出之端點偵測器效能圖（MD vs. FA），其中(15.4%,22%)為參考資料[4]之效能。

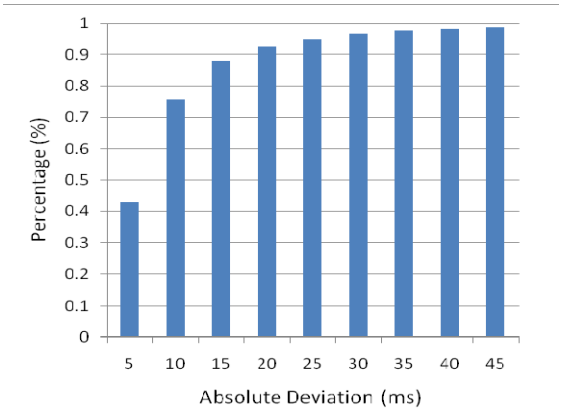

圖 5、對應於自動偵測與人工標記的端點之間差異絕對值之直方圖。

1.3 英語語料庫之音素端點偵測器之實驗結果分析

計畫中，使用 TIMIT 語料庫來驗證我們所提出的 Sample-based boundary detection 演算法的偵測效能。首先，統計了訓練語料與測試語料所處的語音取樣點數(Sample)、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)，並透過修正濾波器係數與數值錯誤影響，得到最後偵測邊界的 EER 效能，如表一。

表一、TIMIT 語料庫的統計資料結果

TIMIT corpus	Sample	Candidate	Phone boundary	EER
Training part	226727341	377998	172461	12.9%
Test part	82786737	136452	62466	14.0%

接著，為了能與傳統 frame-based 方法比較，我們統計每 5ms 所包含到偵測音素邊界的比率，並計算被偵測到的邊界端點落在相同音框或是相鄰之音框內的比率。其中表二所示，音素邊界包含比率分別在相同的音框為 42.94%，相鄰音框為 88.07%，皆優於 Rabiner

(27%/ 10ms, 70%/ 20ms)。

表二、音素邊界偵測結果使用 frame 的方式計算，音框率為 10ms。

Test				
5ms	10ms	15ms	In the same frame	In ±1 frame
42.94%	75.84%	88.07%	41.38%	87.01%

(1) Miss Detection 分析

研究方法為利用音素變化時其語音信號在頻譜之間的變化程度來進行參數的萃取並進行邊界偵測;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。由表三,可以看到相鄰音素是相同與不同的發音方式對照下，發現大部分的 MD rate 都大幅降低的現象。而不同發音方式的相鄰音素邊界，其偵測效能較相同發音方式相比能有效地提升。

表三、TIMIT 測試語料中相鄰音素在相同與不同的發音方法之 MD rate。符號*代表無此組合方式

Test	MD rate (current and next)	
Manners	Same	different
Stop	19.5%	17.5%
Affricates	*	6.7%
Fricatives	31.6%	7.6%
Nasals	58.2%	18.9%
Glides	18.9%	20.4%
Vowels	33.3%	10.2%
Silence	69.2%	11.2%

(2) Test - False Alarm 分析

語音信號之頻譜變化快速的地方容易在不同的發音方式產生 False Alarm 的情形，如表四中，塞擦音、摩擦音以及母音等發音方式之邊界。

表四、TIMIT 測試語料中不同的發音方法之 FA rate

Manners	FA rate	Manners	FA rate
Stop	13.59%	Nasals	10.20%
Affricates	13.35%	Glides	16.58%
Fricatives	15.01%	Vowels	17.00%
		Silence	13.35%

2. 國語語料庫之類音素端點偵測器之實驗結果分析

本研究使用此架構訓練一個監督式(supervised)MLP 類音素端點偵測器。訓練類音素端點偵測器流程，其方塊圖如圖 6。

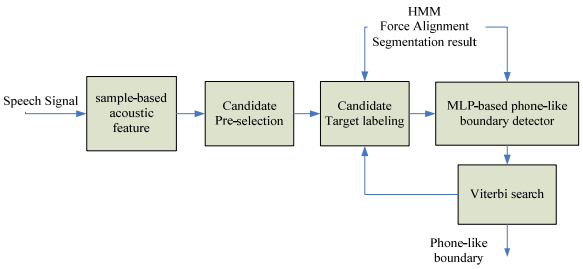

圖 6、使用 MLP 架構之類音素端點偵測流程

實驗結果以 frame-based HMM 架構之切割位置來比較，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。所使用的國語語料庫為 TCC-300 麥克風語音資料庫之長文語料，是由國立交通大學、國立成功大學所錄製的部分，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。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，作為 TCC-300 語料庫之類音素起始切割位置。

獲得 MLP 端點偵測器後，將 sample-based MLP 類音素端點偵測器偵測之端點範圍限制於 HMM 初始切割結果之正負 100 ms 範圍內進行 Viterbi search 的到最佳的切割結果，以下列舉兩個語音波形比較音素端點偵測結果與 HMM 的切割位置。由下列圖 7、圖 8 之中，我們同樣可由圓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子音與母音之間的端點位置都非常準確。圖 7 所示之橢圓形圈選處，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，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。

在此我們統計了 HMM 各發音方法之平均音長，並與我們提出的方法作比較(表五)。而由各發音方法之平均音長，觀察發現 HMM 之結果較音素端點偵測器之平均音長平均結果皆多出 10-20ms 以上的範圍，其原因在於 HMM 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子音之平均音長下降，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。

3. 客語語料庫之 initial-final 端點偵測器之實驗結果分析

在計畫中，新世代語音辨認技術中之屬

性偵測器是利用語言學中的一些發音特徵所以應該是可以是用於所有語言的。因此，我們將英文語料庫上提出之 sample-based 的語音參數所訓練的音素端點偵測器(phone boundary detector)的優異效能，應用於客家話語料中並觀察其特性。

所測試之客語語料為四縣客家話的語料庫，文章來源主要由苗栗退休教師龔萬灶老師所主筆，語料發音人也是龔老師。音檔共有660句，均為20kHz的取樣頻率及16-bit之單聲道pcm格式。語音檔是由發音人在普通房間依照文稿唸出，錄音軟體為Adobe Audition 1.0，並使用單一指向性麥克風。此客語語料庫並無任何人工標記的切割位置可供使用，故以HMM force alignment之切割位置來作說明。

我們利用由 TIMIT 英語語料庫所訓練的音素端點偵測器來偵測其客語語料。為提升在語音屬性偵測器的效能，如圖 9 由上至下的圖形分別代表 HMM force alignment 的切割位置、波形、spectrogram 以及偵測器輸出音素邊界轉換點之 likelihood(範圍在-1~1 之間，其值為 1 則可能性為最大；相反地，值為-1 可能性最小)。首先觀察語音信號的變化可對應至我們證實其音素邊界偵測確是可跨語言的，並可藉由適當調整臨限值(threshold)來達到最佳偵測器所輸出之偵測結果。依照此結果我們可將語音信號分為一段段的音段，這些音段邊界呼應了先前我們提出之 sample-based 的語音參數可描述在語音信號變化時的特性。

觀察目前偵測器的輸出結果，我們可以發現在語音屬性不同的轉換時，偵測器的偵測效能明顯，而利用這些音段來提供該語音信號的特性以用於語音屬性偵測器的測試。

語句切割位置之準確度與精準度的提高，可幫助製作語音屬性偵測器的效能提升。若訓練語句之切割位置越精準則表示該音段屬性之統計特性越明顯越可能提高其準確度。圖 9 中之紅色方框內，其子音(g_e)包含了 short pause、stop(g)和部分母音(e)等 3 種語音分類，若我們依照 force alignment 的切割位置去製作一個語音屬性偵測器，會使得訓練模型的可靠度降低。同樣的情形，我們也可在圖 9 藍色方框內發現。

利用前述 TCC-300 語料庫所建立之自動標示的方法，我們利用其架構來印證在客語語料庫的效能，兩者不同的差異在於切割端點之最小單位不同，客語為 initial-final 單位，較類音素單位來的大。

以下同樣列舉兩個語音波形的例子來比較音素端點偵測結果與 HMM 的切割位置。圖 11、圖 12 之中，可由圓圈之圈選處觀察到一些現象，其音節之間邊界或是子音與母音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，若我們依照此切割位置來訓練語音屬性偵測器，可預估其語音信號特性整體表現會較 HMM 切割位置好，進而提升訓練模型的可靠度。上述結果不只證實其發音特徵是可應用於所有語言且利用本研究所提出之自動標示方法，能大幅度地提升自動標記的精準度。

4. Sample-based weighted spectral KL distance

由先前所述，我們所提出之 sample-based 聲學參數是以六個 sub-band signal envelope 為基礎。然而這些 sub-band 的頻帶寬度不盡相同，是否對頻寬作正規化。故計畫中將此問題作延伸探討。

本研究所使用的聲學參數中最能指出語音信號在頻譜間的差異為 spectral KL distance，若語音信號在該候選端點之頻譜間的差異越大，其參數值則越高，代表著該候選端點越有可能是不同音素之間的轉換，如下式。

$$d_{KL}[n]=\sum_{i=1}^{6}\left(E_i[n]-E_i[n+1]\right)\log\left(\frac{E_i[n]}{E_i[n+1]}\right) \quad (4)$$

其中 $E_i[n]$ 為第 n 點之第 i 個 sub-band signal envelope 正規化之後的能量波封。

因此，假定我們能增進 spectral KL distance 的鑑別程度，使得屬於音素端點的候選端點能夠更容易挑選出來並壓抑非音素端點的候選端點，即可達到音素邊界偵測的效能。參考 weighted entropy 的概念[16]，我們將 spectral KL distance 進一步改為 weighted spectral KL distance, weighted spectral KL distance 定義為每個 sub-band 乘上對應之加權值 w_i 的總合，如下式：

$$\begin{aligned} d_{KL}[n] &= \sum_{i=1}^{6} w_i\left(E_i[n]-E_i[n+1]\right)\log\left(\frac{E_i[n]}{E_i[n+1]}\right) \\ &= \sum_{i=1}^{6} w_i x_{n,i} \end{aligned} \quad (5)$$

本研究使用最小分類錯誤(Minimum Classification Error, MCE)[17]方法來估測由 sample-based 聲學參數之六個 sub-band signal envelope 計算所得到 weighted sub-band

spectral KL distance $x_{n,i}$ 之加權值，w_i。期望藉由加權計算後所得到的 weighted spectral KL distance 對於音素端點偵測之鑑別度能有更進一步地提升。

而在 MCE 鑑別式訓練方法內主要可分為三個部分，首先我們定義鑑別性訓練的錯誤分類量測(Misclassification measure)，接著利用損失函數(loss function)來表示其分類的正確率與錯誤率，最後是以最小化分類錯誤的目標來估測模型參數。由上述我們將數學式表達如下式：

$$Min\left\{\sum_{x_n\in NB} f(d_n\sum_i w_i x_{n,i})+\sum_{x_n\in B}\left[1-f(d_n\sum_i w_i x_{n,i})\right]\right\} \quad (6)$$

其中我們定義欲辨別之類別分別為音素端點之類別，B；非音素端點類別，NB。依照音素端點偵測的假設檢定(hypothesis)定義 d_n，其虛無假設(null hypothesis)該候選端點為音素端點；而對立假設(alternate hypothesis)表示該候選端點並非音素端點，

$$\begin{aligned} &H0: d_n = +1 \\ &H1: d_n = -1 \end{aligned} \quad (7)$$

且對於所有加權值皆滿足 $\sum_{i=1}^{6} w_i = 1$，$0 \le w_i \le 1$ 之限制。

另外，定義損失函數為 sigmoid function 所近似的 0-1 損失函數，意即

$$f(X) = \frac{1}{1+\exp(-c(X-x_0))} \quad (8)$$

其中 sigmoid function 對應至值域範圍為[0,1]的連續性函數，且具有可微分之特性；而 c 可依照參數 X 的動態數值範圍大小來設置，其反映出參數輸入 sigmoid function 所處理的數值範圍；x_0 表示參數 X 欲處理之數值範圍的參數偏權值。

欲得到最佳的加權值 w_i 與參數偏權值 x_0，我們利用最陡坡降法(Steepest Descent Method)來迭代求取最佳加權值 w_i，

$$\nabla w_i = c\cdot d_n \cdot f(dw_{KL}[n])\left(1-f(dw_{KL}[n])\right) \cdot w_i \cdot \left(x_{n,i} - dw_{KL}[n]\right) \quad (9)$$

$$\nabla x_0 = -c\cdot d_n \cdot f(dw_{KL}[n])\left(1-f(dw_{KL}[n])\right) \quad (10)$$

$$\begin{aligned} w_i^{(k+1)} &= w_i^{(k)} + \mu\cdot\nabla w_i^{(k)} \\ x_0^{(k+1)} &= x_0^{(k)} + \mu\cdot\nabla x_0^{(k)} \end{aligned} \quad (11)$$

其中，μ 為學習速率常數 (Step size)。

我們觀察原本 spectral KL distance 的分佈，將 c 設定為 50 且學習速率常數之初始值為 0.2，每個 sub-bandspectral KL distance 之加權值設為等機率。

最後之實驗結果我們得到對應至每個 sub-band 的加權值。如圖 13 所示，我們可以看出第一個頻帶其加權值較大，依序為第三個頻帶及第六個頻帶，而第四及第五個頻帶其加權值相對於其他頻帶小非常多；然而觀察語音信號的轉變，若其靜音接至母音、摩擦音接至母音或是塞擦音接至母音，可由看出其頻譜差異最大的是較低頻的部分，這與圖 13 的分布一致。我們將會將此 weighted spectral KL distance 進一步用於音素端點偵測器中。

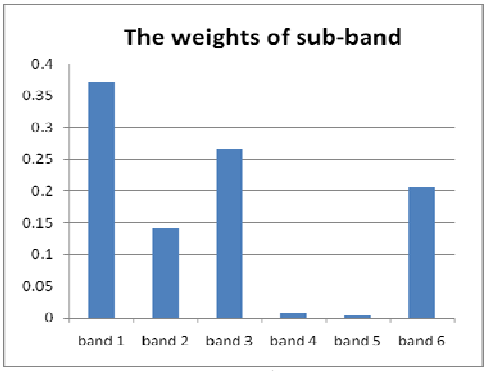

圖 13、加權值收斂後，對應於每個 sub-band 的加權值

四、 結論

本計畫提出一些 sample-based 的語音參數(acoustic feature, AP)，我們也增進 sample-based 的語音參數在音素端點偵測器中之效能。並在國語及客語語料庫無正確音素人工標示資訊下，分別已對 TCC-300 語料庫做自動化類音素端點標示工作，並製作出國語語音之發音方式之偵測器與對客語語料庫作子母音單元的自動端點標示及觀察偵測音素邊界的結果。

五、 計畫成果自評

在計畫書中所列舉之項目均已執行並獲得初步之結果。sample-based 的語音參數在音素端點偵測器中之效能以進一步改善。並在國語及客語語料庫無正確音素人工標示資訊下，分別已對 TCC-300 語料庫做自動化類音素端點標示工作，並製作出國語語音之發音方式之偵測器與對客語語料庫作子母音單元的自動端點標示及觀察偵測音素邊界的結果。同時國語 TCC-300 語料庫做自動 phone-like unit 端點標記結果也發表於 ROCLING-2009[18]。

六、 參考文獻

[1] C.-H. Lee, “From knowledge-ignorant to knowledge-rich modeling: A new speech research

paradigm for next generation automatic speech recognition," *Proc. ICSLP2004*, Keynote speech, 2004.

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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

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

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

[10] C.H. Lee, M.A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.H. Juang, L. R. Rabiner, " An overview on automatic speech attribute transcription (ASAT), " , *InterSpeech 2007*.

[11] Paul Mermeilsteinu, " A Phonetic-Context Controlled Strategy for Segmentation and Phonetic labeling of Speech, ", IEEE Trans. On ASSP, Vol. ASSP-23, No. 1, pp. 79-82, Feb. 1975.

[12] H. Hermansky, N. Morgan, " RASTA processing of speech, ", IEEE Trans. On SAP, Vol. 2, No. 4, pp. 578-589, Oct., 1994.

[13] Mari Ostendorf, Salim Roukos, " A Stochastic Segment Model for Phoneme-based Continuous Speech Recognition, " IEEE Trans. On ASSP, Vol. 37, No. 12, pp. 1857-1869, Dec., 1989.

[14] Eric A. Wan, " Neural Network Classification : A Bayesian Interpolation, ", IEEE Trans. On Neural Network, Vol. 1, No. 4, Dec. 1990.

[15] Yih-Ru Wang, " The signal change-point detection using the high-order statistics of log-likelihood difference functions, ", ICASSP 2008, pp. 4381-4384, April, 2008.

[16] Li Lao, Xiaoming Wu, Lingpeng Cheng, Xuefeng Zhu. "Maximum weighted entropy clustering algorithm, ", Proceedings of the 2006 IEEE International conference on Networking, Sensing and Control, 2006, 1022-1025.

[17] 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.

[18] 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.

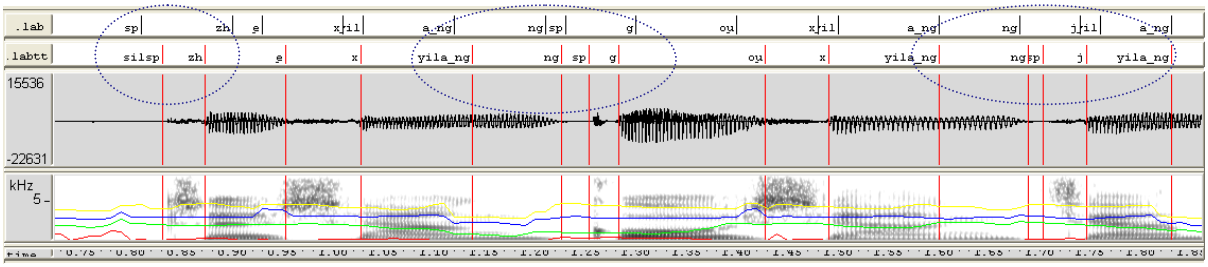

圖 7、國語語句音素端點偵測之例子，由上至下的圖形分別表示原語者調適 HMM 切割位置及音素端點偵測之切割位置、波形、頻譜。

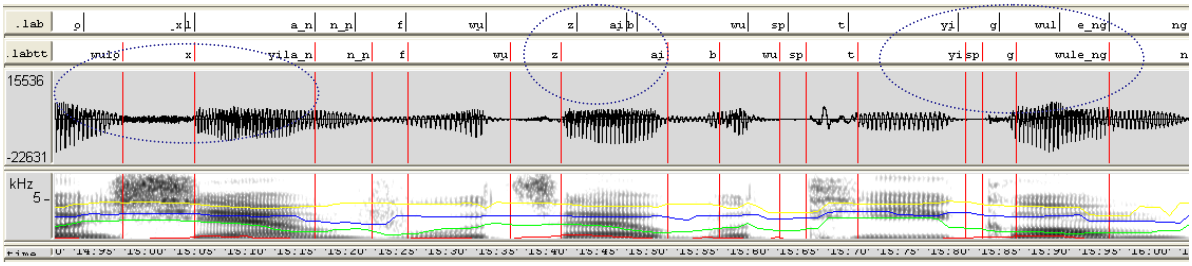

圖 8、國語語句音素端點偵測之例子，由上至下的圖形分別表示原語者調適 HMM 切割位置及音素端點偵測之切割位置、波形、頻譜。

表五、HMM frame-based 與 MLP sample-based 切割位置之發音方法平均音長

發音方法	HMM frame-based	MLP sample-based
爆破音 Stop	4.96	2.62
鼻音 Nasal	5.95	4.46
摩擦音 Fricative	11.13	8.75
塞擦音 Affricate	8.92	7.13
流音 Liquid	6.23	2.70

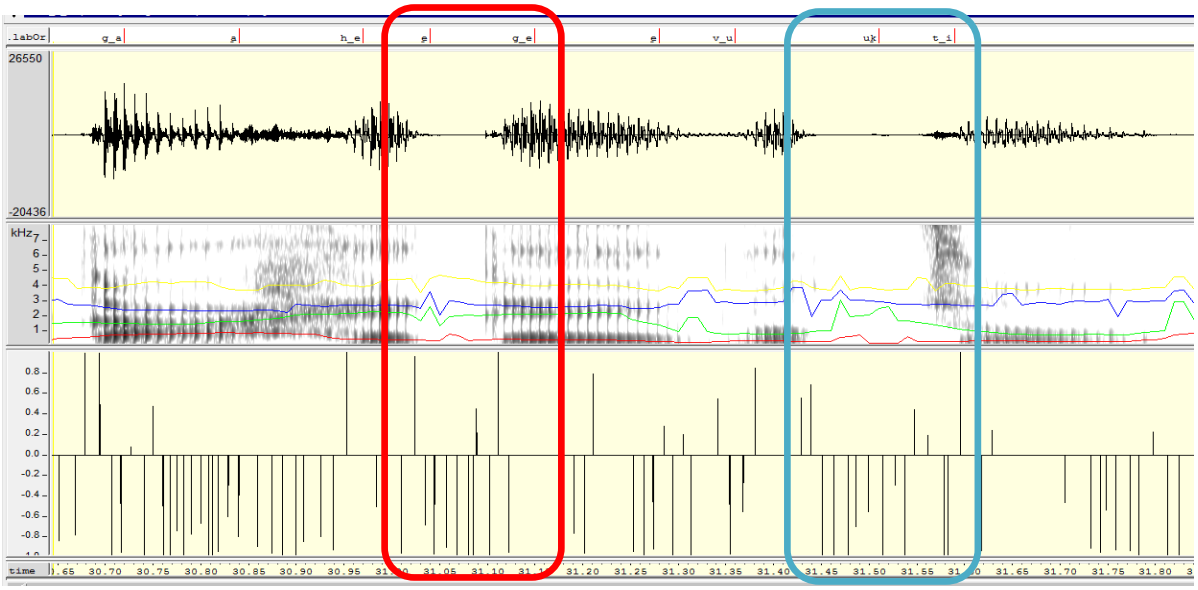

圖 9、客語語句偵測音素邊界之例子，由上至下的圖形分別代表 HMM force alignment 的切割位置、波形、spectrogram 以及偵測器輸出音素邊界轉換點之 likelihood。

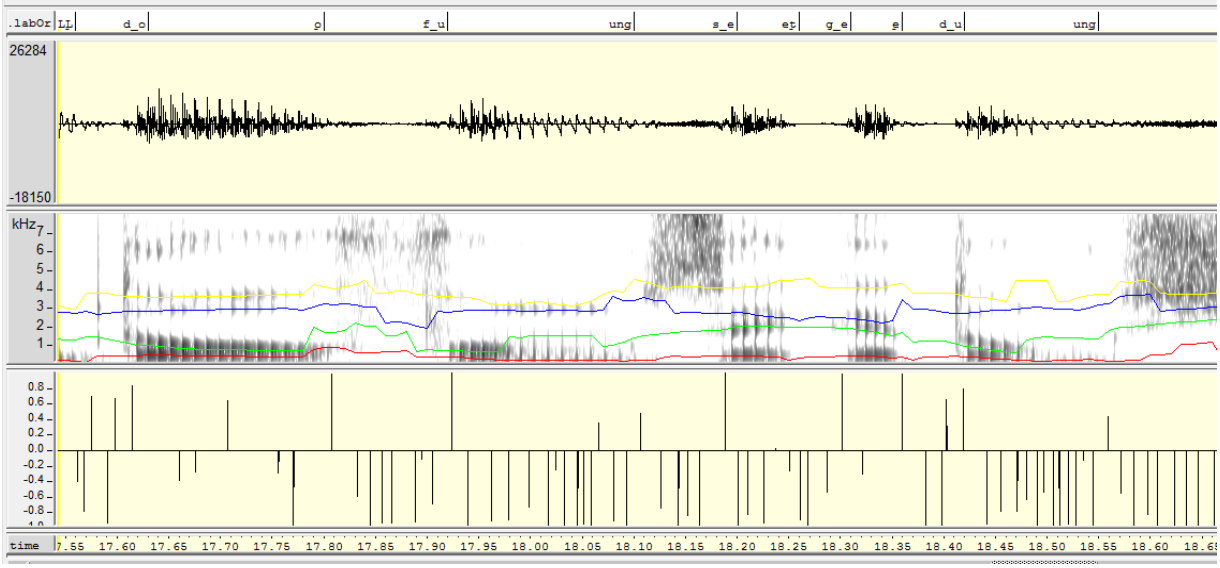

圖 10、客語語句偵測音素邊界之例子，由上至下的圖形分別代表 HMM force alignment 的切割位置、波形、spectrogram 以及偵測器輸出音素邊界轉換點之 likelihood。

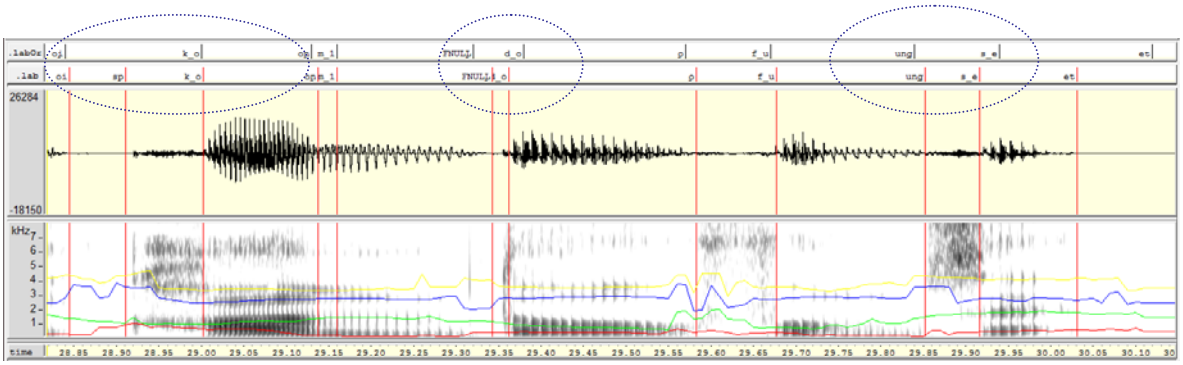

圖 11、客語語句音素端點偵測之例子，由上至下的圖形分別表示 HMM 切割位置及音素端點偵測之切割位置、波形、頻譜。

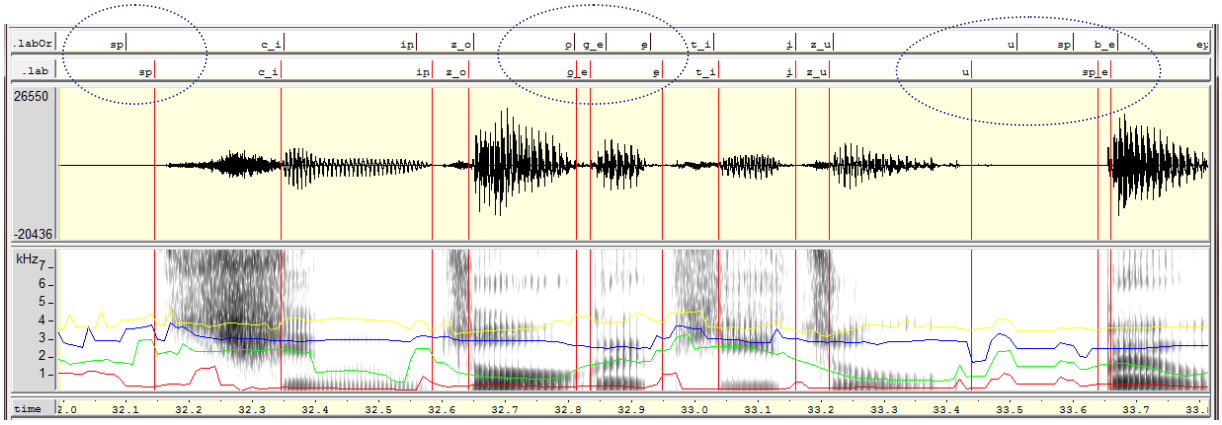

圖 12、客語語句音素端點偵測之例子，由上至下的圖形分別表示 HMM 切割位置及音素端點偵測之切割位置、波形、頻譜。