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(1/3)

計畫編號：97-2221-E-009-080-MY3
執行期限：97 年 8 月 1 日至 98 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，第一年進行下列研究：

具有高解析度音節及其相關的端點偵測器(syllable-level boundary detector) －計畫中首先將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，再結合語音信號在時態變化的結構特性(temporal structure information)製作一個可靠的階層式音節及其相關的端點偵測器。

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 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). In the first year, the following items were carefully studied and implemented:

Syllable-level boundary detector using temporal structure information – High-resolution sample-based landmark detectors will be developed using articulation parameters. Moreover, hierarchical syllable-level boundary detectors will also be implemented to verify the results of the landmark detectors’ using the temporal structure constrains of the speech signal.

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。

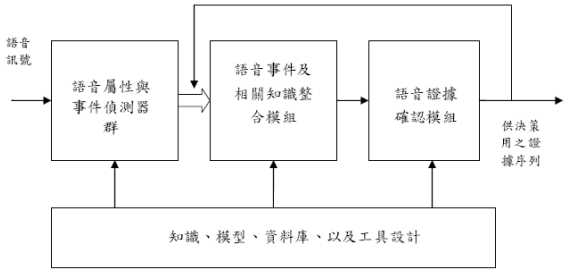

圖 1、新世代自動語音辨識技術架構圖。

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-wise 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。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-wise 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. 使用聲音樣本資訊之語音特徵

首先，我們提出一些 sample-wise 的聲學參數(acoustic feature, AP)如 signal envelope、flatness、ROR(Rate of rise)、KL distance、entropy、 ROR of signal envelope、band-signal envelope…等等)，並觀察它們在不同語音信號之變化及其它們的特性。

(1) sub-band signal envelope

在語言學家所提出的 AP 中，有許多帶通濾波器能量(band-energy)如圖 2，它們各自能用來區別不同的發音部位及方法，常見的 band energy 有以下[7]：

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 fricative、affricate 中，在高頻的成份能量極強，低頻成分能量較弱，nasal 則是在低頻的成份能量極強。Filter bank 中能量在有明顯變化的時候，可視為是語音信號中 attribution 改變的地方。

(2) ROR (Rate of rise)

語言學家所稱之 ROR(Rate of rise)事實上就是我們經常在 frame-based 的語音特徵參數中所用的 delta-term：

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

其中 $x[n+i]$為輸入參數資料，w 為做 ROR 所使用的 window-width。

(3) Spectral entropy [8-9]

Spectral entropy 可用來描述信號在頻譜

上的集中程度。假設$E_i[n]$為第 i 個 sub-band 之第 n 點正規化之後的能量，則 spectral entropy，H_S可以定義如下式表示：

$$H_S = -\sum_i E_i[n]\log\left(E_i[n]\right)$$

其中

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

(4) Sample-base spectral KL distance

將 spectral 視為一個機率分佈，我們可以以 KL distance 來描述 spectral 的相似程度。在語音信號中計算 2 點不同時間(m 與 n)的 spectral KL distance，$d_x(m,n)$，可以由下式表示：

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

上述參數（Spectral entropy、ROR of Spectral entropy、Sample-base KL distance）來觀察一段語音信號其語音特徵的變化，我們發現在語音信號變化的時候，這些語音特徵皆有容易分辨的位置來標記成邊界。

(5) Spectral flatness

我們使用正規化後 sub-band energy 的 flatness，F，表示如下式：

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

其中 S_i 為第 i 個 sub-band silence 信號正規化後能量的平均。若信號為 silence 或是 short pause，則 F 將會趨近於 1。Spectral flatness 與 envelope 參數在標記 silence 及 short pause 的切割位置時是一個有效的參數。

2. 使用 sample-wise 聲學參數之做音素端點之自動標示

由於國內現有之國語語料庫均無人工的標音位置，因此我們在製作國語語音屬性偵測器的第一步便是需要一個使用自動方法所做之切割位置。如何獲得正確標示的國語語料庫，將嚴重影響中文語音屬性（attribution）與各種語音事件（event）偵測器之效能。在第一階段計畫中，我們曾使用 HMM 所獲得之切音為製作為正確標示，雖可用來製作中文語音屬性（attribution）與各種語音事件（event）偵測器，但其效能卻一直無法提升。所以在本子計畫中做了自動標音之研究。

對國語語料庫 TCC300 進行標示之更正與自動標音工作，其步驟如下：

(1) 首先我們已利用 SAT(speaker adaptation transform, feature MLLR)及 SA(speaker adaptation, MLLR)後的出語者相關 HMM 來做 TCC-300 之 phone-like unit 之自動切割位置，接著我們利用新的切割位置建立新的發音方法偵測器;以語音發音方法做分類的方式，如表一，由原本的切割位置當做 reference，找尋有效的語音參數來調整 phone-like unit 端點。由於先前 refinement 過後之切割位置，已近乎準確，但是仍有更進一步修正的空間，我們提出以 sample-based 音素端點偵測的方式可達到更為精確的切割位置。

(2) 先前在觀察 HMM 自動切音位置的標記時，發現短靜音常無法切割出來，而使得 affricate 與 stop 等音素平均音長過長的現象，經使用語者相關 HMM 切音後，各發音方法之平均長度均有下降，尤其是 stop 及 affricate 音，並且可以切割出音節間的短靜音。如圖 4，在這裡我們使用 spectral flatness、envelope 以及各 filter bank 之 energy 來判斷是否為 short pause 的產生。可以觀察到 short pause 中各個 sub-band energy 與其他有語音信號的地方相比幾乎很低，且 flatness 趨近於 1，envelope 可與 flatness 產生互補的效果來標記 short pause 的端點。

(3) 接下來我們觀察 fricative、affricate，它們在 spectrogram 中與相鄰 vowel 與 short pause 有極大的頻譜差異。故我們使用 spectral KL distance、entropy 以及 ROR of entropy 來偵測音素的端點。如圖 5，我們可以發現在音素端點頻譜分佈差異大，使得 KL distance 高，足以當做端點位置標記;且 fricative、affricate 相鄰 vowel 的端點，entropy 上升與下降速度很快，分別在 ROR of entropy 中造成極大、極小的峰值。不難看出峰值附近的位置也是音素端點。

(4) Stop 切割位置的修正時，由 waveform 與 spectrogram 觀察中，我們可以發現通常在 stop 開始的時候會有短時間 short pause 出現，envelope 接著會有急遽上升的現象，故我們使用 ROR of envelope 來描述其現象。如圖 6 所示，在 stop 結束的地方，也是音素轉換的端點。

(5) Nasal 則可使用 0.0 – 0.4 KHz filter band 的 envelope 來做調整。

(6) Vowel 端點的偵測，是利用相鄰子音及 short pause 之端點位置，當做 vowel 的邊界切割位置。

由上面可以發現我們所使用的高時間解析度的 sample-wise 語音參數對音素端點真測試具有高鑑別率的。

3. 英語語料庫之音素端點偵測器之製作

因為現有之國語語料庫均無人工標示資訊;所以在此我們先使用已有人工標示資訊的英語語料庫，來檢視所提出 sample-wise 聲學參數對音素端點偵測方法之效能。

3.1 使用 MLP 神經網路架構的 Sample-wise 端點偵測方法

此方法我們使用 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,最後我們訓練一個 supervised neural network 來做為音素端點偵測器。

我們從語音信號中抽取 sample-wise 聲學參數之後,為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(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

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

則挑選出來為候選端點值,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片段(segment),我們也可由這些片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 spectral KL distance 來評斷相鄰 2 個片段 $[c_{k-1}\ c_k]$ 、$[c_k, c_{k+1}]$，如圖 7 所示。

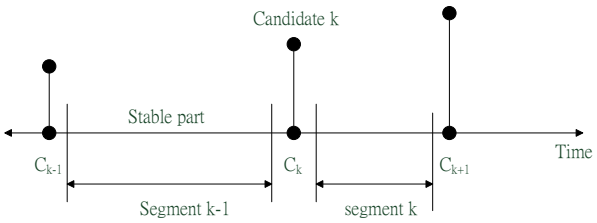

圖 7、利用候選端點將語音信號分割成片段的示意圖。

而 segmental spectral KL distance 可定義成下式

$$DS_x(k,k+1) = \sum_{i=1}^{6}\left(ES_i(k) - ES_i(k+1)\right)\log\left(\frac{ES_i(k)}{ES_i(k+1)}\right)$$

其中 $ES_i(k)$ 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

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

除此之外,對於每個片段中每個 sub-band envelopes 的平均值與斜率同樣可以當做端點偵測的 feature。其在 k 片段中之第 y 個 feature 可定義如下

$$S_k(y) = \left(\sum_{n=c_{k-1}+\Delta_2}^{c_k-\Delta_2} y[n]\right) / \left(c_k - c_{k-1} - 2\Delta_2\right)$$

與

$$\breve{S}_k(y) = R_y\left(\left(c_k + c_{k+1}\right)/2, \left(c_k - c_{k+1} - 2\Delta_2\right)/2\right)$$

其中 Δ_2 通常不包含候選端點轉換的部分。

接著，我們對於每個候選端點建立一個 75 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：

$\{ \left(d_x(c_j, c_j+1), DS_x(j-1,j), E_i[c_j]\right); i=1,\cdots,6$，

$R_{E_i}(c_j, \Delta_1); i=0,\cdots,6\ \}$； $j = k-1, k, k+1$，

(2) 目前片段及前、後片段之參數：

$\{ \left(S_j(e_i), \breve{S}_j(e_i)\right); i=1,\cdots,6; j=k-1,k;$

$S_k(e_0) - e_0[c_k],\ S_{k-1}(e_0) - e_0[c_k]\ \}$

(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。

最後我們在系統中訓練一個 supervised multi-layer perception(MLP)來做為音素端點器。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 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。

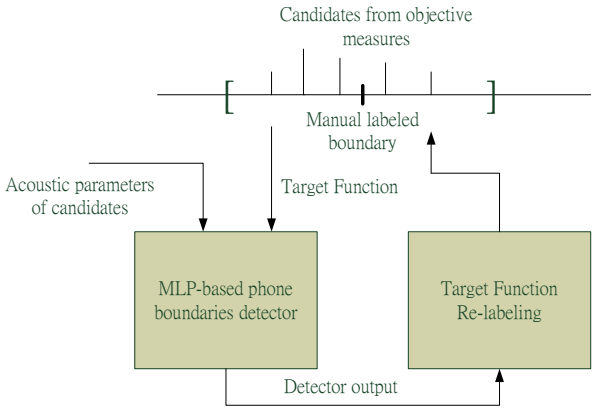

圖 8、Interactive target selection 與 MLP training algorithm 之方塊圖。

在我們的系統中，提出一個音素端點偵測之一個 interactive target selection 與 MLP 訓練演算法發展如圖 8。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，當做初始目標；
(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做 MLP-based 偵測器的新目標；
(4) 重複(2)與(3)的步驟，直至收斂。

3.2 實驗與結果

TIMIT 語料庫在此用來確認我們所提出的 sample-wise 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 有一個端點。

在實驗中，經由適當地選擇臨限值來挑選候選的端點。總體來說，對應於訓練語料與測試語料分別有 197213 個與 713236 個音素端點被挑選出來。只有 0.94%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 50。

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。在可容忍 400 samples error 下，訓練與測試資料的 MD(missing detection) vs. FA(false alarm) 曲線如圖 9 所示；其中，FA 定義為(number of false alarm)/(number of actual boundaries + number of false alarm)。結果對於訓練與測試語料分別有 20.5%與 21.4% EERs(equal error rate)。如下圖所示， 23.1% MD 與 22%FA 比率，相比於參考資料[4]之效能，大約達到下降 10% ERR。

自動端點偵測與人工標記端點之間的 MSE (mean-squared error)則如圖 10 所示。我們可以看到若自動偵測端點結果的 confidence 越高，則與人工標示值之間的差異就越小。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，可偵測出一些明顯的端點且與與人工標示值之間的差異也會越小。

用我們的系統所偵測出來的端點中，有 53.5%是與人工標記的切割位置相距不到 80 個 samples；有 93.8%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。

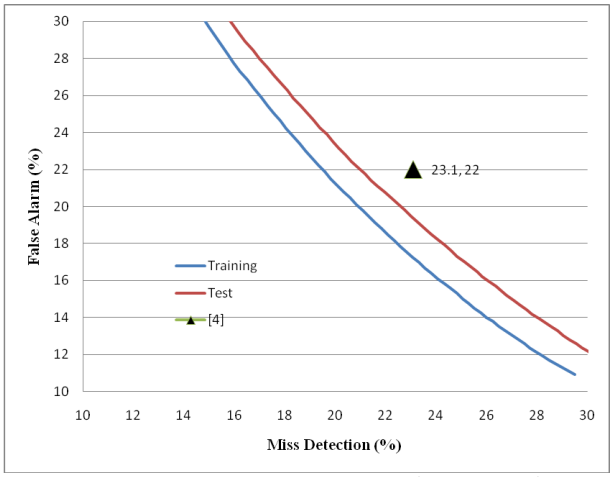

圖 9、所提出之端點偵測器效能圖（MD vs. FA），其中 (23.1%,22%)為參考資料[4]之效能。

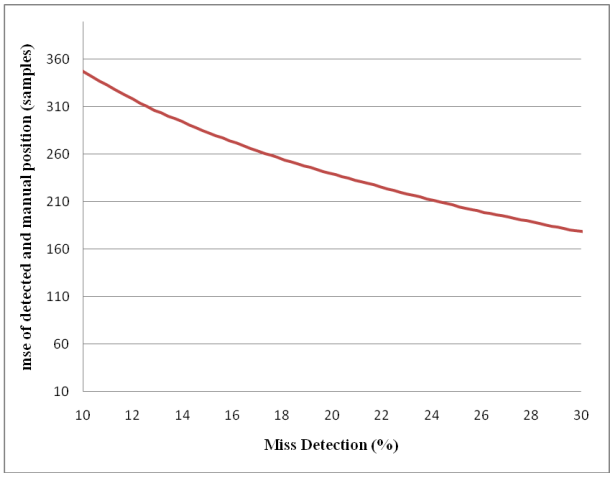

圖 10、對應於自動偵測與人工標記的端點之間的 MSE 與 Miss Detection 比率之關係圖。

同時我們也正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。

4. 結論

本計畫提出一些 sample-wise 的語音參數(acoustic feature, AP)，並在國語語料庫無正確音素人工標示資訊下，已對 TCC-300 語料庫做自動化 phone-like unit 端點標示工作。我們也在英文語料庫上證實所提出之 sample-wise

的語音參數在音素端點偵測器中之優秀效能。也正在製作國語語音之發音方式之偵測器。

四、 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，並提出高解析度之自動 phone-like 單元位置之標示演算法，對 TCC-300 語料庫所獲得之自動 phone-like 單元位置之標示將會透過總計畫之平台提供給國內語音相關研究者使用。所提出之方法也在具人工標示資料之英文語料庫 TIMIT 獲得效能驗證。

五、 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.

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				
母音 Vowel	others					

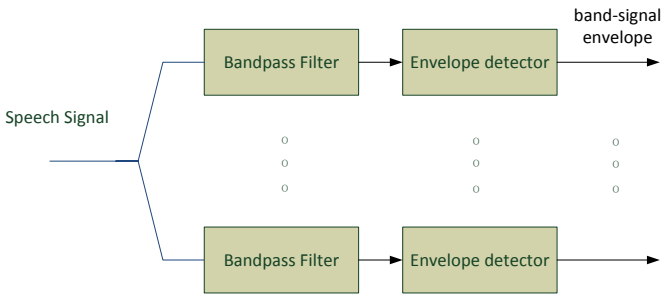

圖 2、band-signal envelope detector 之方塊圖。

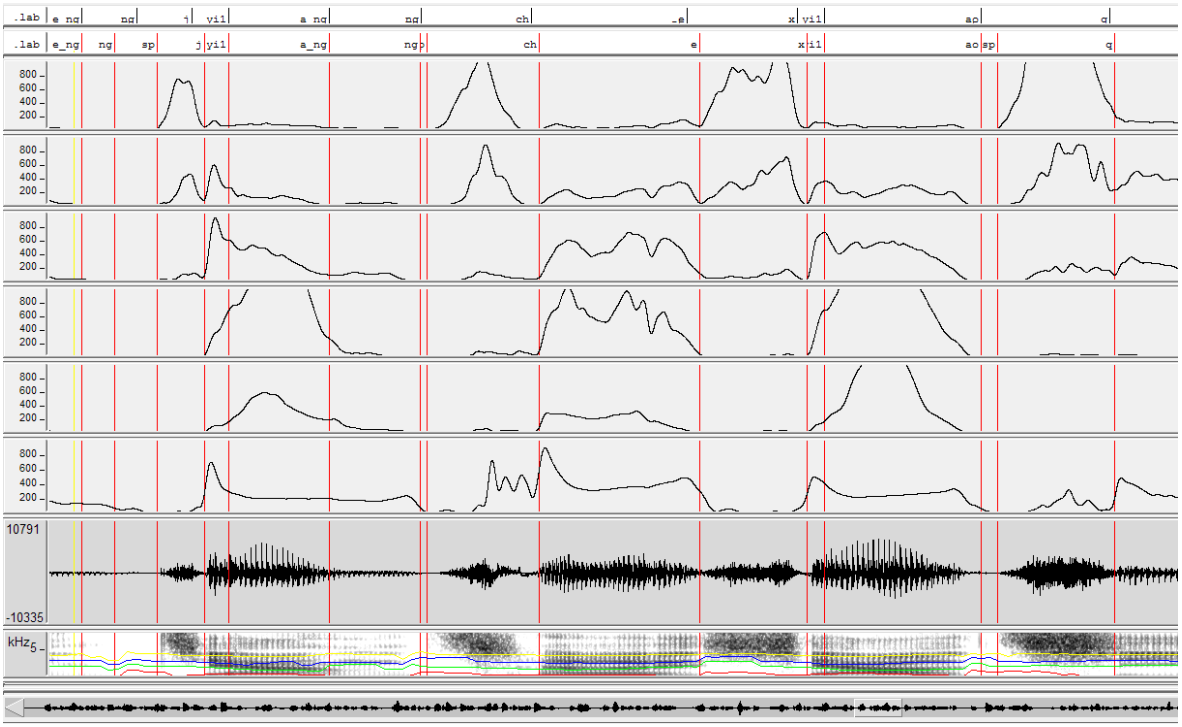

圖 3、國語語句切音位置自動調整(fricative、affricate)演算法則之例子，由上至下的圖形分別代表第 6 個至第一個 sub-band energy、波形、spectrogram。（最上方兩行標音是原端點及修正後之端點）

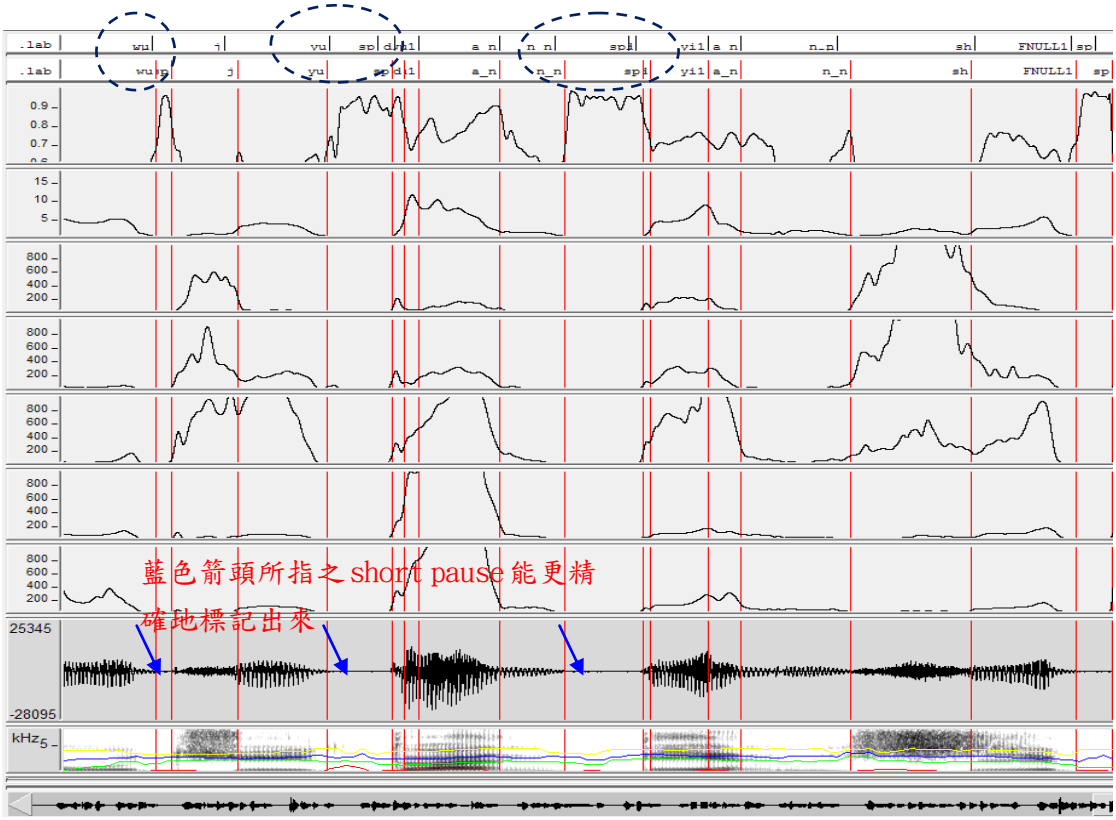

圖 4、國語語句切音位置自動調整(stop、affricate)演算法則之例子，由上至下的圖形分別代表 spectral flatness、envelope、代表第 6 個至第 4 個 sub-band energy、波形、spectrogram。（最上方兩行標音位置是原端點及修正後之端點）

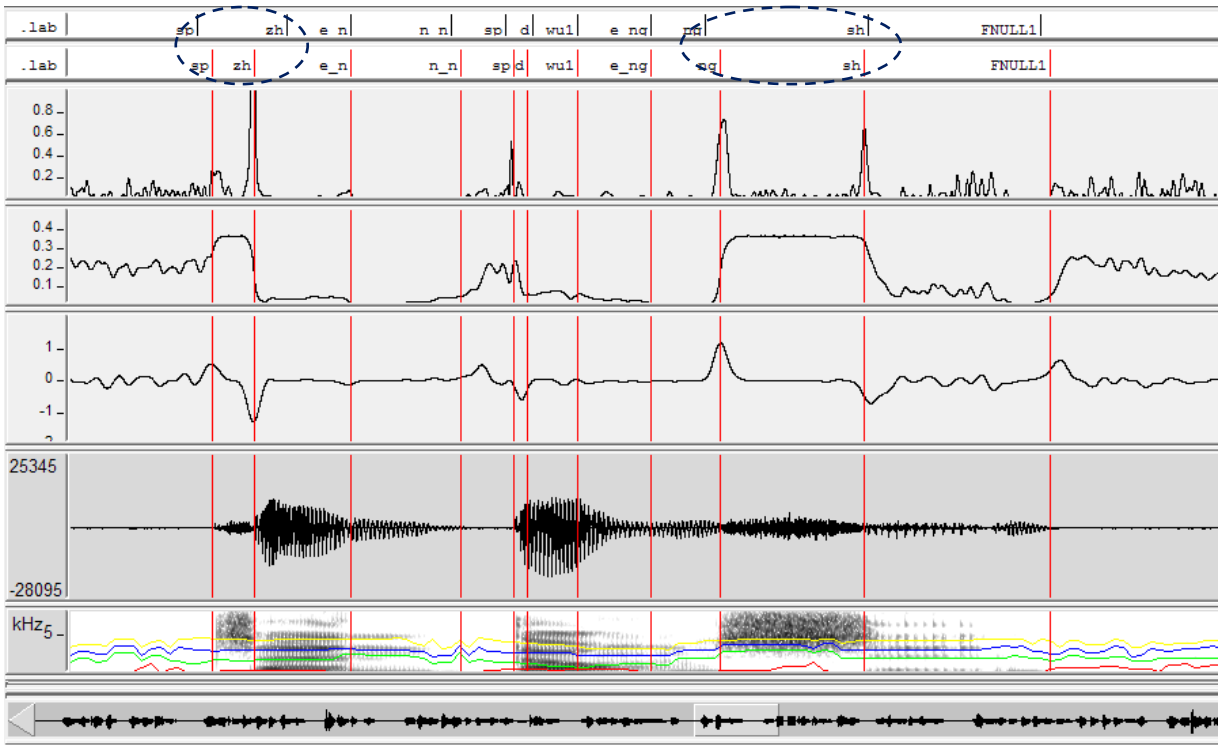

圖 5、國語語句切音位置自動調整(fricative、affricate)演算法則之例子，由上至下的圖形分別代表 sample-base KL distance、spectral entropy、ROR of Spectral entropy、波形、spectrogram。（最上方兩行標音位置是原端點及修正後之端點）

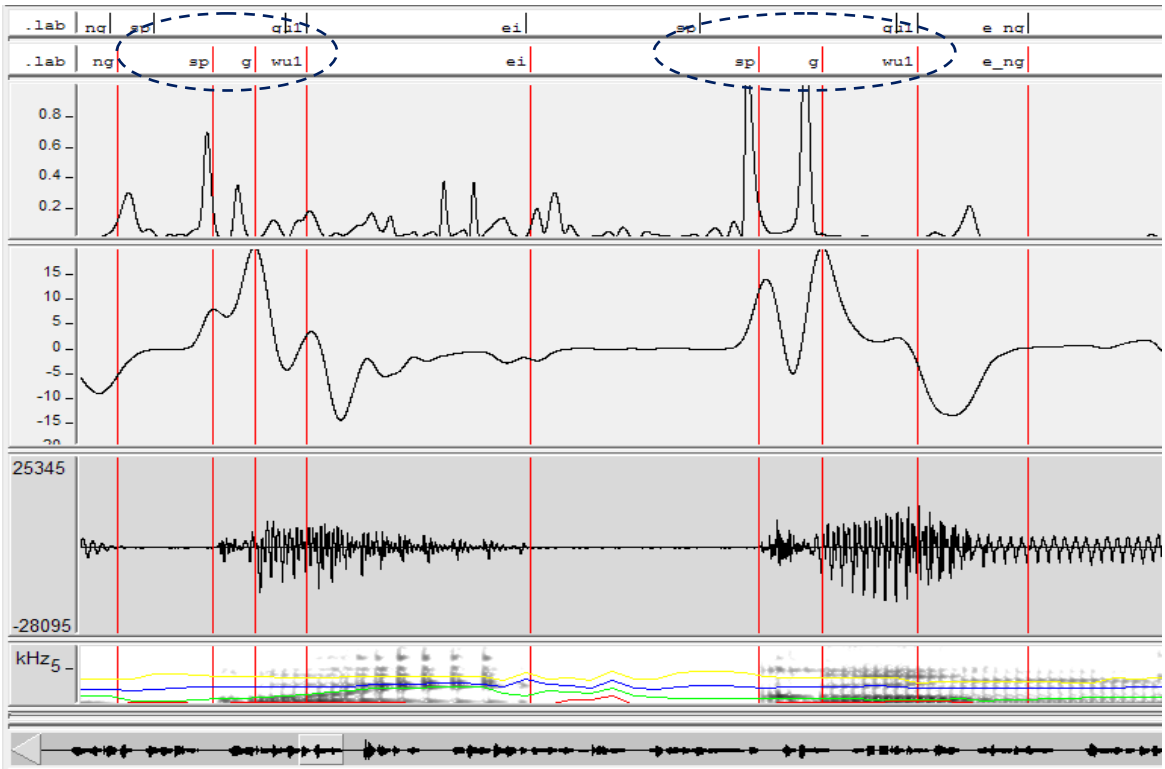

圖 6、國語語句切音位置自動調整(stop)演算法則之例子，由上至下的圖形分別代表 sample-base KL distance、ROR of envelope、波形、spectrogram。（最上方兩行標音位置是原端點及修正後之端點）