行政院國家科學委員會專題研究計畫　期中進度報告

子計畫二：語音、韻律之屬性與事件偵測之研究(1/3)

計畫類別：整合型計畫
計畫編號：NSC94-2213-E-009-134-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立交通大學電信工程學系(所)

計畫主持人：王逸如
共同主持人：廖元甫
計畫參與人員：蕭希群、許見偟

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 6 月 1 日

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之研究－
子計畫二：語音、韻律之屬性與事件偵測之研究(1/3)

計畫編號：NSC94－2213－E－009－134
執行期限：94 年 8 月 1 日至 95 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
計畫參與人員：蕭希群、許見偉

一、中文摘要

在傳統語音辨認方法中，通常只使用語音的頻譜參數作辨認。但是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、韻律及語言相關的訊息，在交與後級『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（hidden Markov model, HMM）方式的困境。本計畫中即擬進行國語語音之各種語音屬性、音節邊界、基頻軌跡、韻律資訊、語者資訊之偵測研究，以做為新世代自動語音辨識系統之前端處理器。研究重點如下：

1. 中文語音屬性（attribution）與各種語音事件（event），包括偵測發音方法(articulation manner) ，發音部位(articulation position)與其他語音特徵參數（distinctive feature）。
2. 中文音節界標（boundary landmark）偵測器，提供後級正確時序訊號。
3. 中文基頻軌跡偵測器，包括新的求取方法與軌跡追蹤（tracking）方式。
4. 中文音調與韻律訊息偵測器。
5. 語者資訊（speaker profile）偵測器，包括語者性別、年齡、口音等。

關鍵詞：新世代自動語音辨識系統，語音屬性偵測，語音事件偵測，基頻軌跡偵測，語者資訊偵測。

Abstract

In this project, various speech attribution-, speech event-, syllable boundary-, pitch contour-, prosodic information- and speaker-information-detectors will be studied and act as the front-end of the next-generation automatic speech recognizer. The focuses of the research include:

(1) Mandarin speech attribution, event and other distinctive feature detectors including articulation manner and articulation position
(2) Mandarin syllable boundary detector to provide syllable timing information
(3) Mandarin pitch contour extraction and tracking
(4) prosodic information detection and tone recognizer
(5) speaker profile detectors, including gender, age, accent and speaking rate

Keywords: next generation ASR, speech attribution detection, speech event detection, pitch contour detection, speaker profile 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。各個國家都針對其所用的語言進行大量語音與語言資料的收集，就特定的一些應用領域發展語音辨識系統，例如聽寫機(voice dictation machine)、交談系統(conversational system)、口語文件擷取(spoken document retrieval)、口語翻譯(spoken language translation)等。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**錯誤! 找不到參照來源。**。在**錯誤! 找不到參照來源。**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。

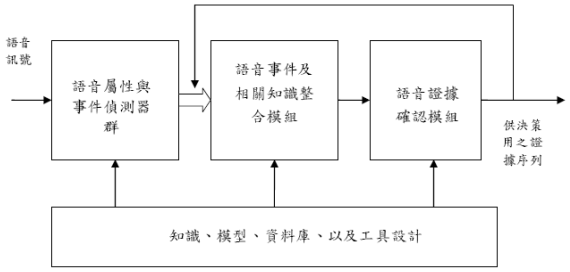

圖 1、新世代自動語音辨識技術架構圖。

在本子計畫中，我們將致力研究新世代自動語音辨識技術架構(圖)中之語音屬性與事件之偵測器。在語音屬性與事件之偵測器中對於語音訊號，不只是抽取語音特徵，而且要偵測某一時段中語音的屬性。因此除了傳統的聲學特徵參數，還要依據語音學或語言學的知識，抽取韻律相關的訊息與發音方式的訊息，協助區辨容易混淆的聲音。利用語音的特徵與屬性的發生，描述一段語音事件(event)的發生，而從事件序列來做語音辨識的決策。也就是以事件序列，判斷是否有某一段語音的發生。抽取不同的語音特徵與屬性，可以偵測出若干個事件序列，將更有助於作正確的判斷，提高語音辨識的正確率。

三、研究方法

1. 英語之發音方式及發音位置偵測器之製作及其效能

要製作語音發音方式及發音位偵測器，首先必須要有一個有標準答案的語料庫，經搜尋後發現 TIMIT 英語語料庫中有人工標示至音素單元，所以在計畫中我們先對英語語音做發音方式及發音位偵測器

TIMIT 英語語料內容為 2342 句平衡語料，由分佈在美國八個不同方言的地區共 630 位語者，每人錄製 10 句，共有 6300 句語音，其中 438 位男性、192 位女性。並以其中 4620 句、語料長度總和約為 3 小時 49 分 10 秒的語音訊號作為訓練語料，另外 1680 句、語料長度總和約為 1 小時 23 分 51 秒的語音，作為測試語料。語料的音訊格式為 PCM，取樣頻率為 16 kHz，位元解析度為 16 bits。表 1 為英語各音素之發音方式及發音位置表。表 2、表 3 中為 TIMIT 英語語料庫中各種發音方式及發音位之相關統計資料。

表 2、TIMIT 語料庫中發音方式之統計資料。

	Training		Test	
Manner	count	Average frame	count	Average Frame
Vowel	57463	9.57	20911	9.67
Fricative	21424	9.12	7724	9.20
Stop	25871	4.12	9176	4.11
Nasal	14157	5.68	5104	5.69
Glide	20257	6.40	7822	6.55
Silence	35877	9.48	12777	9.20
Affricate	2031	6.98	631	7.08

表 3、TIMIT 語料庫中發音位之統計資料。

	Training		Test	
Position	count	Average frame	count	Average Frame
bilabial	8796	4.57	3416	4.53
labdent	4210	8.28	1622	8.41
dental	3577	4.90	1320	4.83
alveolar	32662	6.56	11375	6.60
velar	10648	6.26	3658	6.43
glottal	4547	6.50	1600	6.67
rhotic	11992	7.62	4708	7.82
front	34883	9.07	12503	9.14
central	15684	7.61	5881	7.66
back	14204	10.30	5285	10.40
silence	35877	9.49	12777	9.20

在計畫中，我們首先使用 Bayesian detector 來做為發音方法、發音位置偵測器。Bayesian detector 利用事後機率(a Posteriori probability)來判別輸入信號是否屬於某一特定事件，也就是

$$p(\theta \mid x)=\frac{p(x \mid \theta)p(\theta)}{p(x \mid \theta)p(\theta)+p(x \mid \hat{\theta})p(\hat{\theta})} \geq \zeta$$

其中 θ 為 target model，而 $\hat{\theta}$ 為 anti-model，x 為每一個 frame 的特徵參數向量，ζ 則為 threshold 值。$p(\theta)$ 與 $p(\hat{\theta})$ 為各個 class model 與 anti-model 的事前機率。

藉由比較某一發音方法或發音位置與其 anti-model 的事後機率去偵測測試語料的每一個 frame 是否為所要偵測的種類，如圖 2 所示。

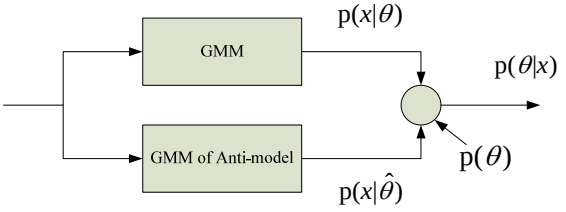

圖 2、Bayesian detector 架構圖。

而 $p(\theta \mid x)$ 及 $p(\hat{\theta} \mid x)$ 則使用 GMM (Gaussian

mixture model)來表示。也就是令

$$p(x|\theta) = \sum_{i=1}^{N} C_i \cdot N(\mu_i, \Sigma_i)$$

$$N(\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(x-\mu_i)^T (\Sigma_i)^{-1}(x-\mu_i)]$$

在每個發音方法、發音位置偵測器中的每個 model 及 anti-model 的 mixture 數分別設定為 256(發音方法)與 64(發音位置)，因此我們將看看對於測試語料，各種偵測器的偵測效果。

表 4 為 GMM-Based 發音方法偵測器與國外學者用類神經網路(multilayer perceptron, MLP)偵測器架構所做出來的 EER (Equal Error Rate)性能[2]作比較。

表 4、GMM-Based Bayesian manner detector 與其它偵測架構之效能比較。

	Frame-based detector	
EER(%)	Baseline(GMM)	MLP
Vowel	12.3	9.0
Fricative	10.0	11.3
Stop	16.7	14.5
Nasal	8.7	12.2
Glide (Approximant)	16.3	15.9
Silence	9.7	3.7
Affricate	7.2	

由表 4 可以看出，以 GMM-Based 的 Fricative 與 Nasal 偵測器，其效能較佳於 ANN，尤其是 Nasal 偵測器改善了約 3%，尤其是 silence 偵測器差了 6%。這提供了我們未來在做其他偵測器架構一個參考的依據。表 5 則為 GMM-Based 發音位置偵測器的效能。

表 5、GMM-Based mixture64 的發音位置偵測效能比較。

EER(%)	GMM-based Bayesian detector
Bilabial	12.2
Lab-dent	11.0
Dental	12.7
Alveolar	12.0
Velar	12.4
Glottal	18.3
Rhotic	9.4
Front	13.5
Central	17.7
Back	17.8

由表 5 可以看出幾乎全部的發音位置偵測器的 EER 均大於 10%以上、除了 Rhotic 偵測器、但也很接近 10%。其中以 Glottal、Central、Back 偵測器錯誤率皆大於 17%以上為最差。在計畫中，我們也初步嘗試結合兩個 detector 的輸出，首先我們假設所有 phoneme 其所屬的發音方法與發音位置是獨立的關係

$$p(\text{manner} \cap \text{position} \mid x) = p(\text{manner} \mid x) \times p(\text{position} \mid x)$$

其中 $p(\text{class} \mid x)$ 為這個 frame 屬於這一個 class 的機率值。

最後我們可以得到每一個音框其發音方法結合發音位置的機率值，並且藉由設定 threshold 可以畫出 False Alarm 以及 False Reject 圖。由圖 3 可以看出對一些音素發音方式及位置偵測結果結合後的結果可以改善 EER，某些音素則否；這是因為 manner 及 position 這兩種屬性事實上並不是獨立。如何結合不同偵測器之結果是個有趣的課題，我們將會在繼續在往後兩年計畫中加以探討。

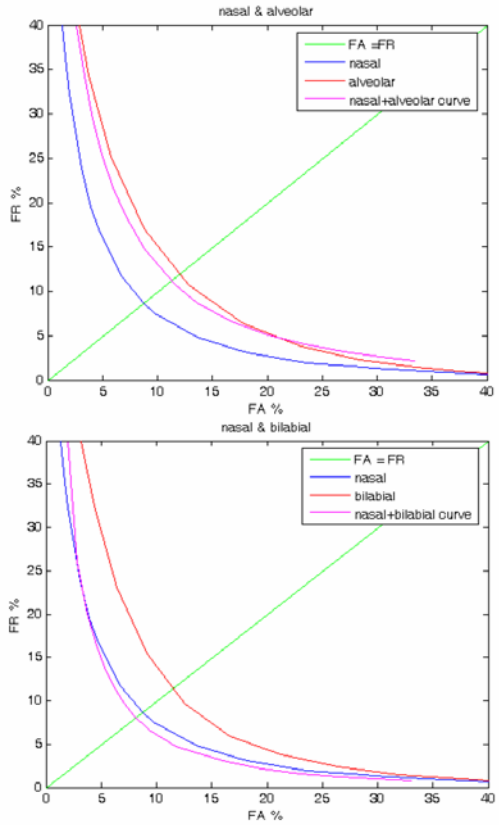

圖 3、manner、position detector 之初步結合之結果。

2. 國語語音資料庫之預處理

由於國內現有之國語語料庫均無正確（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。在計畫中，我們使用國內發行的國語語料庫 TCC300，利用 HMM 模型做 force alignment 以自動求取 TCC300 語料庫的標音位置。在 HMM 模型訓練時 frame size 為 32ms 、 frame shift 為 10ms 的 38 維 MFCC 參數，訓練以 Phone-Based 為架構的 HMMs ，而其每個 phone-like model 狀態數皆

設為 3 。並將每一個 model 的每一個 state 的 mixture 數平均值設為 128。接著我們將訓練好的 HMMs 拿來對 TCC300 的訓練語料以及測試語料作已知字串的切割，因此我們可以得到一個有粗略位置資訊的 phone-level 的 labeling file，最後我們可以利用中文發音方法的分類表(表 6)，將 phone-level 的 labeling file 轉為 manner-level 的 TCC300 的訓練語料的 labeling file。而此 labeling file 便當作是我們在製作國語發音方法偵測器的初始切割位置。表 7 為 TCC300 訓練語料初始切割位置的統計資料。

manners	times	Frame amount	Min frame	Average frame
Vowel	418337	4088079	3	9.77
Fricative	74276	829482	3	11.17
Stop	76291	632948	3	8.30
Nasal	119535	692825	3	5.80
Liquid	14653	100047	3	6.83
Silence	350316	1456902	0	4.16
Affricate	75889	781293	3	10.30

表 7、TCC300 訓練語料初始切割位置的統計資料。

3. 國語之發音方式偵測器之製作及其效能

我們使用與英語屬性偵測器一樣的 Bayesian detector 架構來做國語發音方式的偵測器。各種發音方式的偵測器之 EER 值則如表 8 所示。

表 8、 英文與中文發音方法偵測器各自對其測試語料所作的偵測結果。

	Frame-based Bayesian detector	
EER(%)	**English**	**Mandarin**
Vowel	12.3	10.70
Fricative	10.0	15.7
Stop	16.7	11.5
Nasal	8.7	11.5
Glide/Liquid	16.3	9.2
Silence	9.7	8.0
Affricate	7.2	11.5

圖 4 中顯示國語語音發音方法偵測器偵測中文語句實例，我們利用 wavesurfer 來製作一個語音、韻律之屬性與事件偵測器之標準圖形輸出介面[4]。其中 column3-9 為各發音方式 detector 輸出，圖中 column10 為使用 HMM 獲得之 phone-level 切割位置，column11 為 detector hard-decision 輸出。

4. 國語語料庫 HMM 切割資訊之改善

由觀察許多國語語音發音方法偵測器之偵測結果發現，silence 與 fricative、nasal 常會混淆，。事實上，HMM 切割資訊與人工切割在音節起點與終點會有 30-50msec 的誤差這也是已知的事實。由表 7 中我們可以發現 Stop 及 Affricate 的平均音長過長，所以我們必須設法獲得較佳之切割位置。事實上，使用 HMM 所獲得之的靜音模型都較差，那是因為在語音資料之 transcription 中並無是否有靜音存在的資訊；我們正利用下列方法來改善使用 HMM 所求得的音素切割位置

(1) 在 stop initial 前加入 non-skipable silence，以其獲得更正確的靜音切割位置；
(2) TCC300 語料庫中有許多背景雜訊存在：如呼吸聲、麥克風撞擊聲等，我們利用 UBM (universal background model) 來 model 這些背景雜訊存，如此可自動的將這些信號從 silence model 中移除，如此將可製作出較佳的 silence detector；
(3) 將 GMM-based detector 視為 1-state HMM model，使用 GMM-based detector 去重新作語音信號的切割。

上述工作證進行中，其結果亦顯示靜音部分之切割位置也獲得改善，其結果亦將在用於製作新的國語之發音方式偵測器。

5. 結論

本計畫在第一年以製做出基本的英語發音方式及發音位偵測器並對其效能進行評估。在無音素人工標示資訊下，也已製作出國語語音之發音方式之偵測器；並對國語語音音素之自動且正確標示也做了許多工作，才可以對無人工標音之國語語料庫進行語音、韻律之屬性與事件偵測之研究。

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，而正確音素切割位置之國語語料庫之建立將必須與總計畫及其他子計畫通力合作完成，始可提供國內進行語音、韻律之屬性與事件偵測之研究專家學者做效能評估之依據。

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] C.-H. Lee, “A Study on Separation between Acoustic Models and Its Applications,” *Proc. ICASSP2005.*

[3] D. Povey & P.C. Woodland. *An Investigation of Frame Discrimination for Continuous Speech Recognition.* Technical Report

CUED/F-INFENG/TR.332, Cambridge University Engineering Dept., 1999.

[4] Wavesufer Homepage : http://www.speech.kth.se/wavesurfer/

[5] G. Flammia, P. Dalsgaard, O. Andersen, B. Lindberg, " Segment based variable Frame Rate Speech Analysis and Recognition using spectral variation function," , ICSLP1992, Banff, Vol. 1, pp. 671-674.

[6] S. Dusan, *Statistical Estimation of Articulatory Trajectories from the Speech Signal Using Dynamical and Phonological Constraints*, Ph.D. Thesis, Dept. of Electrical and Computer Engineering, Waterloo University, Canada, 2000.

表 1、英語各音素之發音方式及發音位置表。

	Bilabial	Lab-dent	Dental	Alveolar	Velar	Glottal	Rhotic	Front	Central	Back
Stop	b p			d t dx	g k	q				
Nasal	m em			n en nx	ng eng					
Fricative		f v	th dh	s z	sh zh					
Glide						hh	r	y	l hv el	w
Affricate				jh ch						
Vowel							er axr	iy ih eh ey ae ay ix	aa aw ax ax-h	ah ao oy ow uh uw ux

表 6、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	r	f	s	x	h	sh
塞擦音 Affricate	q	j	c	z	zh	ch
流音 Liquid	l					
母音 Vowel	others					

圖 4、國語語音發音方法偵測器偵測一國語語句實例。