先進中文語音辨認系統之發展(3/3)

Development of Advanced Mandarin Speech Recognition Systems

計畫編號：NSC-91-2219-E-009-038
執行期限：91 年 8 月 1 日至 92 年 10 月 31 日
全程計畫：89 年 8 月 1 日至 92 年 10 月 31 日

schen@mail.nctu.edu.tw

一、中文摘要

本三年計畫擬開發先進的中文語音辨認技術，研究主題涵蓋語音辨認前處理、聲學辨認單元模式、韻律模式、雜訊通道效應補償等主題，本報告說明第三年之研究成果，包括音節基週軌跡模型、說話速度及音節間耦合效應補償、電話語音之通道/語者模型與辨認。

關鍵詞：中文語音辨認、基週軌跡模型、說話速度補償、通道/語者模型

Abstract

The three-year project aims at developing advanced technologies for Mandarin speech recognition. Research topics cover pre-processing, acoustic modeling, prosodic modeling, and adverse speech recognition. This is the third-year report. Items accomplished are described as follows. Firstly, a new statistical pitch contour model which considers several major affecting factors is proposed. Secondly, new methods to compensate speaking rate and inter-syllable coarticulation effect are proposed. Lastly, an ASR method which can effectively compensate channel/speaker effect is proposed.

Keywords: Mandarin speech recognition, pitch contour modeling, speaking rate compensation, channel/speaker modeling

二、緣由與目的

近年來語音辨認技術已有長足進步，一些實用系統陸續被開發出來，發展實用系統的關鍵之一在於雜訊及通道效應的去除或補償，國外對此問題已經由蒐集大量語料來廣泛地進行研究，國內也已完成大型電話語料庫(MAT-2000, MAT-2400) 及麥克風語料庫(TCC-300) 之蒐集，亦開始深入探討此問題。本計畫之目的是要使用 MAT 及 TCC 語料庫來進行先進的中文語音辨認技術的研究。

三、結果與討論：

本年度主要進行之研究包含以下三項：音節基週軌跡模型、說話速度及音節間耦合效應補償、電話語音之通道/語者模型與辨認，分別詳述如下：

(一) 音節基週軌跡模型

中文是聲調語言，而聲調主要由基週軌跡表現出來，因此基週軌跡模式對中文語音處理相當重要，除可用於聲調辨認及 TTS 之韻律合成外，並可提供豐富的信息協助語音之了解。首先我們對基週進行語者正規化 (speaker normalization)，以 frame-based 方式進行

$$f(t)=\frac{f'(t)-\mu_k}{\sigma_k}\cdot\sigma_{all}+\mu_{all}$$

其中 $f'(t)$ 及 $f(t)$ 為原始及正規化後的基週訊號，μ_k 及 σ_k 為語者 k 的 mean 及 standard deviation，μ_{all} 及 σ_{all} 為所有語者平均的 mean 及 standard deviation。接著對基週取對數及進行音節基週之 orthogonal polynomial expansion，以取得代表 mean 的一個參數及代表 shape 的三個參數，分別建立 mean 及 shape 的模型，pitch mean model 為

$$Y_n = X_n + \beta_{t_n} + \beta_{pt_n} + \beta_{ft_n} + \beta_{i_n} + \beta_{f_n} + \beta_{p_n}$$

它考慮現在音節的 tone t_n、前後音節的 tones pt_n 及 ft_n、現在音節的 initial i_n 及 final f_n、及韻律狀態 p_n；pitch shape model 為

$$\mathbf{Z}_n = \mathbf{X}_n + \mathbf{b}_{tc_n} + \mathbf{b}_{q_n} + \mathbf{b}_{s_n} + \mathbf{b}_{i_n} + \mathbf{b}_{f_n},$$

其中 tc_n 為考慮前後 tones 之組合、q_n 為韻律狀態。

我們使用 EM algorithm 來估計 model 之參數以解決韻律狀態為 hidden 之問題，對一個 5 人的 TL-database 做 simulation，此 model 之 RMSE 為 0.362 and 0.373 ms/frame for 訓練及測試語料，所求出之 compressing-expanding 係數符合中文語言學上的知識，且它所標示的韻律狀態可用以判斷語音在無標點符號處的停頓，表 1 列出所標示的 major break、minor break、non-break 和標點符號間之關係，由表中可看出語音中的 major break、minor break 和標點符號有密切關聯性，但並不完全相等；另外，我們這個 model 也提供 quantitative

inter-tone correlation，而非僅是傳統的 qualitative 式的 sandhi rules，圖 1 畫出著名的 Tone pairs 3-3 and 4-4 變調規則，圖中顯示 Tone3 除受右接 3 聲影響而變為 2 聲外，其基週軌跡亦會受左接聲調影響；4 聲除受後接 4 聲影響基週軌跡後段會下彎外，亦會受左接聲調影響。由上討論可知此 pitch contour model 是一極有效之 model。

(二) 說話速度及音節間耦合效應補償

(1) 說話速度補償

說話速度對語音辨認有很大影響，過去研究發現速度快及慢的語音辨認率較差，改進的方式主要有將語音依速度分成快、中、慢三類各自訓練一組 HMM models，辨認時合併所有 models 找最佳的結果；另外也有人提出增加 pronunciation network 以描述速度快的語音。我們採用前一 approach，以音節為單位對語音做分類，以考量在一句話中的速度不一致問題 (例如 prosodic phrase 結尾之音節拉長效應)。以 MAT-2000 及 MAT-2500 的 90%語者的語料來訓練 context-independent (CI) 100 個聲母及 40 個韻母 HMM models，以剩餘的 10%語者的語料做測試，表 2 列出實驗結果，由表中可看出正常速度的語音有最好的辨認率，而語音依速度分類可微幅改進辨認率。

(2) 音節間耦合效應補償

對 inter-syllable coarticulation 嚴重之音段建立額外的連音 final-initial HMM models，其 state 數目降為 6 個，由 MAT-2000 加 MAT-2500 的 90% 語者訓練語料，經 forced-alignment 後，訂定 continuity measure 來檢驗相鄰音節 coarticulation 程度，我們將較常出現嚴重 coarticulation 的 219 個 final-initial pairs 挑選出來建立 HMM models (包括 state duration model)，被挑選的音節 pair 主要是聲母為 sonorant (無聲母、nasal、liquid 等)。辨認時將所有 HMM models 拿來辨認，但為防止發音速度快的音節耦合 models 引致較大的音節插入錯誤，我們在辨認時加入了音節轉移處罰，使用 MAT 語料之實驗結果列於表 3，其中一般模型採用 context-independent (CI) 100 個聲母及 40 個韻母 HMM models，表中顯示此方法可微幅 (2%) 增進辨認率。

(三) 電話語音通道/語者模型與辨認

我們考慮通道/語者模型為

$$\mathbf{y} = A\mathbf{x} - \mathbf{b}$$

其中 $\mathbf{y}$ 和 $\mathbf{x}$ 分別為正規化和原始的語音特徵向量，A 及 $\mathbf{b}$ 為 affine 轉換關係的參數。要估計 A 及 $\mathbf{b}$ 可定義一個客觀的目標函數如下式所示：

$$Q_k = \sum_{t=1}^{T_k} (A_k\mathbf{x}_t - \mathbf{b}_k - \mu_{s_t m_t})^T (A_k\mathbf{x}_t - \mathbf{b}_k - \mu_{s_t m_t})$$

藉由 minimize Q_k 可求出最佳的 A_k 及 $\mathbf{b}_k$ for speaker k。在去除通道/語者效應後，我們可以重新估計 HMM model，和原來的 model 比較，圖 1 顯示 F-ratio 的改變，由圖中可看出對男女/聲韻母之 models，F-ratio 均有顯著增加，顯示此方法效果良好。

我們進一步檢驗此方法對辨認率的改進，以 MAT 的 90%語者的語料來訓練 CI 及 context-dependent (CD) HMM models，以剩餘的 10%語者的語料做測試。比較之方法包括：(1)使用傳統 signal bias remover (SBR) 的 baseline system；(2)以整個語音段求每一語者一組 A_k 及 $\mathbf{b}_k$ 之通道/語者效應補償 (-NF)；及 (3)更進一步將語音信號的有聲 (voiced) 音、無聲 (unvoiced) 音、靜音部分區分開，分別求取此三類型資料的矩陣 $A_{k,(v,u,s)}$ 及向量 $\mathbf{b}_{k,(v,u,s)}$ 之通道/語者效應補償(-NFI)。圖 3 顯示辨認測試之方塊圖，其中使用了測試語料之切割資訊，這在真正測試時並無此資訊，因此測試結果只能視為 performance 之 upper bound，留待以後再解決此問題。

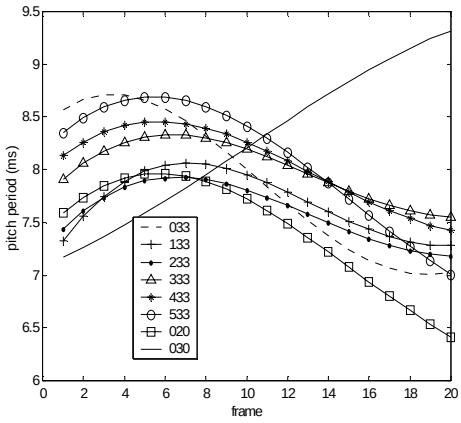
實驗結果列於表 4，由表中可看出此方法較傳統的 SBR 有效，可有效去除語音之通道/語者效應。

四、計畫成果自評：

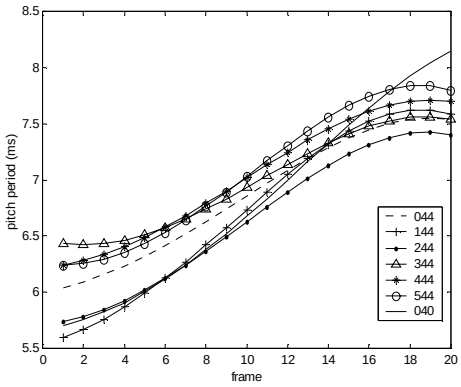
我們在計畫中對中文語音辨認的多個重要問題進行深入探討，對發展先進辨認系統所需的技術已有很好的掌握，執行成果尚稱良好。

表 1: 使用韻律狀態標示停頓之實驗結果。Major PM={，,。,！,；,？}, Secondary Major PM={、,：} and Minor PM={brace, bracket, dot}

Break / PM	Non-boundary	Minor boundary	Major boundary
Non-PM	89.18%	9.80%	1.02%
Minor PM	57.73%	33.48%	8.80%
Secondary Major PM	30.52%	44.65%	24.83%
Major PM	19.31%	31.66%	49.02%



(a)



(b)

圖 1: (a) A comparison of the patterns of Tone 3 preceding another Tone 3 with canonical patterns of Tone 2 and Tone 3. (b) A comparison of the patterns of Tone 4 preceding another Tone 4 with canonical pattern of Tone 4.

表 2：以音節為單元分快中慢速模型之音節辨識率，()中之值為不分速度之基本系統辨認率

unit:%

語者速度	插入率	遺失率	替代率	正確率
SLOW	4.07	0.37	34.4	61.21 (60.3)
NORMAL	1.44	0.78	32.8	65.00 (64.2)
FAST	0.87	2.02	37.7	59.38 (57.5)
TOTAL	2.14	1.03	34.9	62.00 (60.8)

表 3：考慮音節間嚴重耦合模型之音節辨識率

unit:%

語者速度	插入率	遺失率	替代率	正確率
SLOW	2.7	0.6	33.8	62.8
NORMAL	0.9	1.2	32.8	65.2
FAST	0.5	2.5	36.8	60.2
TOTAL	1.4	1.4	34.4	62.8

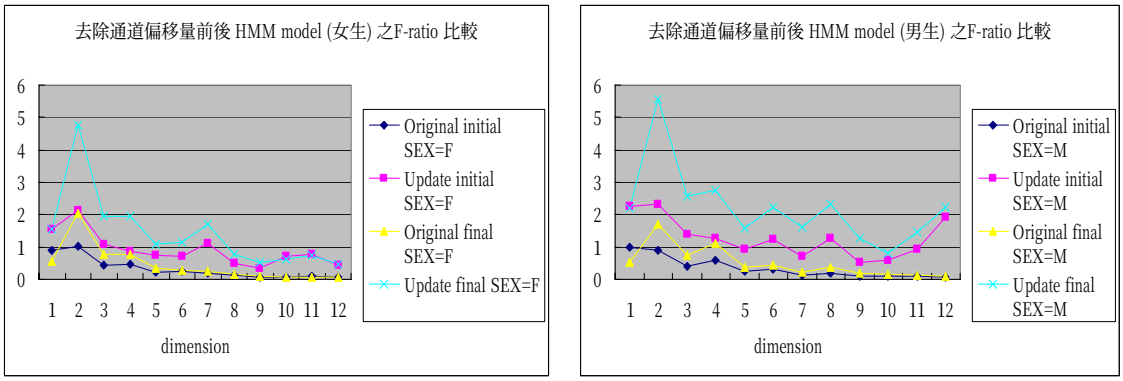


圖 2：去除通道/語者效應前後 HMM model 之 F-ratio 比較圖

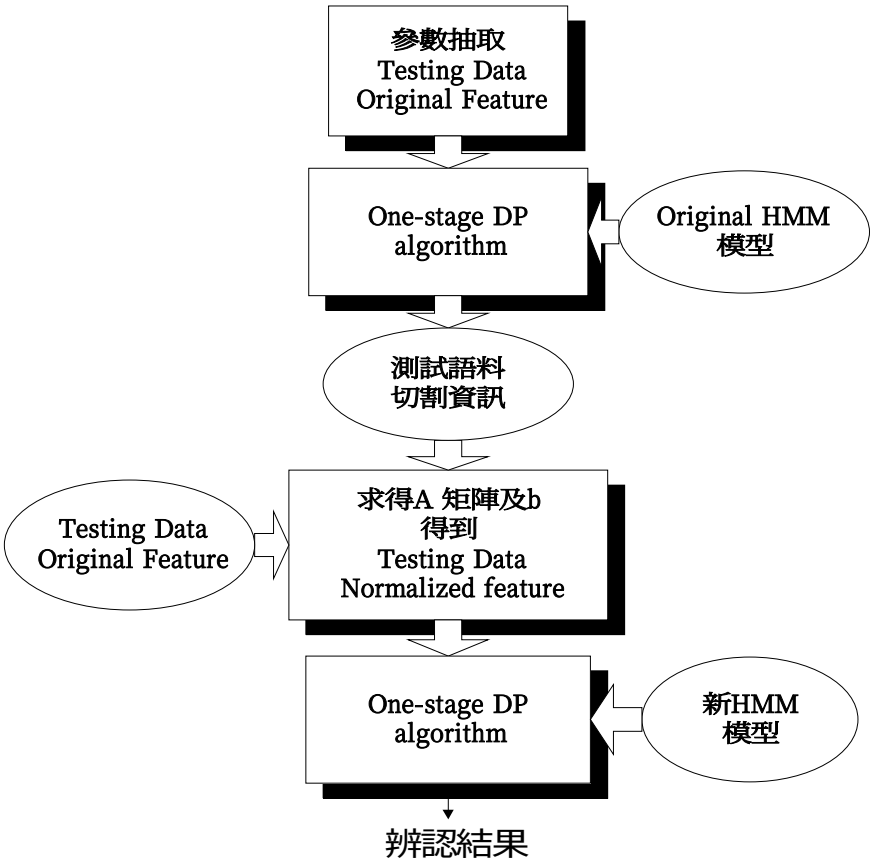


圖 3：移除通道/語者效應之語音辨認測試方塊圖

表 4：特徵參數 affine 轉換(-NF)、改進版本(-NFI)與基本系統之辨認率比較

unit:%

HMM 模型種類 - 偏移量移除方法	音節 插入率	音節 遺失率.	音節 替代率	音節 正確率
CI-SBR (Baseline)	2.57	1.34	35.9	60.2
CD-SBR (Baseline)	3.53	0.81	33.1	62.6
CI-NF (Upper bound)	2.92	1.01	30.4	65.7
CD-NF (Upper bound)	3.98	0.66	27.4	68.0
CI-NFI (Upper bound)	0.77	0.22	25.7	73.3
CD-NFI (Upper bound)	0.96	0.10	22.4	76.6