先進中文語音辨認系統之發展(2/3)

Development of Advanced Mandarin Speech Recognition Systems

期中報告

計畫編號：NSC-90-2213-E-009-041

執行期限：90 年 8 月 1 日至 91 年 7 月 31 日

主持人：陳信宏　國立交通大學電信工程學系

schen@cc.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、雜訊通道補償等主題，本報告說明第二年之研究進展，包括語音切割進一步研究、電話語音之通道效應補償、連音聲學模型之建立、韻律片語邊界之偵測、音節長度模型進一步研究，研究進行順利。

關鍵詞：中文語音辨認、語音切割、通道效應補償、韻律片語邊界、音節長度模型。

Abstract

The three-year project aims at developing advanced technologies of Mandarin speech recognition. Research topics cover pre-processing, acoustic modeling, prosodic modeling, and adverse speech recognition. This is the progress report of the second year. Items that have been accomplished are described as follows. Firstly, a further study on RNN-based speech segmentation is performed. Secondly, channel bias compensation in telephone-speech recognition is discussed in detail. Thirdly, a new final-initial acoustic model for high inter-syllable coarticulation is proposed. Fourthly, a preliminary study on the detection of prosodic phrase boundary is given. Fifthly, a further study on syllable duration modeling is given.

Keywords: Mandarin speech recognition, speech segmentation, channel bias compensation, inter-syllable coarticulation, prosodic phase boundary, syllable duration 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，國內在最近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：

(一) 語音切割之進一步研究

我們對以前提出的使用遞迴式類神經網路(RNN)進行語音切割的方法做進一步改進及分析其功能，以期使用它作為先進的國語語音辨認系統的 sophisticated 前處理器，主要的研究項目包括：輸出入 feature 的調整、切割狀態的粗及細分類、對 MAT-2000 語料之 performance 分析等。

首先，我們調整 RNN 輸入的 features 成為包含前後共五個 frames 的 MFCC 及 log-energy，以及此 utterance 中 silence 音段的平均 MFCC，前者是將 context information 加入，後者則為了對背景雜訊/channel 進行調適。輸出則調整成為包含聲母(I)、韻母(F)、韻尾鼻音(N)、及靜音(S)。

接著，我們設計兩個 finite state machines (FSMs)由 RNN 輸出來判定音段的粗及細分類。音段的粗分類是為了區分 speech/silence，其設計原則是要能identify 語音間的長silence 而忽略音段內音節間的短 silence，同時對 silence 音段中短的雜音不起反應，圖一為一實驗結果例句，顯示其效果良好。音段的細分類是為了進一步區分 speech 音段內的聲母、韻母、韻尾鼻音、聲母-韻母 transition、及音節間的短 silence，以便在辨認前了解 speech 音段之大致的音節結構，其設計原則是要能 identify 穩定 I/F/N/S 音段及可靠的聲母-韻母 transition，而將其餘的歸於一 unknown 狀態，圖二為一實驗結果例句，顯示其效果大致良好。

最後，我們檢驗 RNN 切割對 MAT-2000 語料之 performance 分析，主要檢驗項目為：(1) 無法區分音節邊界之連音(i.e., FSM2 之輸出中相鄰兩音節的 final segments 連成一個)，(2) 音節之刪除錯誤 (i.e., FSM2 之輸出中音節之 final segment 不見了)。對 MAT-2000 語料的 209,641 個音節邊界及 TCC-300 語料的 276,856 個音節邊界之實驗結果如下表所示，連音發生之原因除了後一音節以 sonorant 為起始音素外，主要因素為：(1)後接音節為空聲母，(2)前一音節有鼻音韻尾，(3)兩個相同

母音相連接。

unit:%

語料庫	MAT	TCC
二個音節連音之機率	6.37	5.46
三個以上音節連音之機率	0.74	0.73
音節無 final seg. 之機率	1	0.9
音節有多個 final seg.	0.59	0.13

(二) 電話語音之通道效應補償

電話通道效應之補償是電話語音辨認重要的研究題目，我們探討下面三個問題：(1) 在 HMM 訓練時，對已知切割之語音，如何估計 channel bias，以去除通道效應而獲得較為 compact 的 HMM models？(2)在辨認時無法事先知道語音之切割信息，因此一般使用 SBR (signal bias remover) 估計 channel bias，其效能如何？(3) 如果測試語料和訓練語料不 match，SBR 常不 work，如何改進？

通常我們假設電話通道之模型為

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{b}_t$$

其中 $\mathbf{y}_t$ 為 observed spectral feature vector (i.e., MFCC)，$\mathbf{x}_t$ 為不含 channel bias 之原始語音之 feature vector，$\mathbf{b}_t$ 為 channel bias。首先，我們討論由已知切割之語音做 channel bias 估計，若假設一個 HMM state (m,s)之語音訊號 $\mathbf{x}_t$ 為 normal distribution $N(\mathbf{x}; \mu_{m,s}, \Sigma^2_{m,s})$，則我們可將 $\mathbf{y}_t$ 減去 μ_{m_t,s_t} 而獲得 frame-base bias estimate $\hat{\mathbf{b}}_t$，再對整個 speaker/utterance 語音做 average 即得 channel bias estimate $\hat{\mathbf{b}}$，再將 $\mathbf{y}_t$ 減去 $\hat{\mathbf{b}}$，即得乾淨語音之 estimate $\hat{\mathbf{x}}_t$，可由其重新估計 HMM models。以 MAT-2000 做實驗，結果說明如下：(數字之說明以 MFCC 第一維參數為例)

(1) HMM state 的平均 variance 由 4,01 降為 2.24，F-ratio 由 0.86 增為 1.52；
(2) Speaker-based bias 之 variance 為 1.76；

SBR 則是要先訓練一個 codebook，再由其求 channel bias，32 codewords 之實驗結果為

(1) HMM state 的平均 variance 由 4,01 降為 3.01，F-ratio 由 0.86 增為 1.14；
(2) 和已知切割求 bias 之方法所得的 bias 之 correlation coefficient 為 0.94；如使用 MLVQ 則 correlation 增為 0.98；
(3) 含 2、4 音節之詞及長句之 SBR bias 和已知切割法之 bias 的 correlation 分別為 0.89、0.92 及 0.98；
(4) 對 silence 語音使用 SBR 求得之 bias 和已知切割法之 bias 的 correlation 為 0.57；
(5) 使用 MLVQ 對 55880 音節測試語料之音節辨認率為 63.8%。

最後，我們考慮 mismatch 問題，使用 MAT-2000 訓練 HMM models 及 SBR codebook，而對另一個工研院 ATC 的 3617 音節之 database 作測試，直接使用 SBR 之辨認率為 23.2%；使用 CMN 之辨認率為 56%；如先用本句之 CM 估計補償 mismatch，再用 SBR 去除 bias 之辨認率為 56.68%；如先去除兩個 databases 的 mean deviation，再用 SBR 去除 bias 之辨認率為 59.97%；如用前一句辨認切割信息與本句之 CM 共同估計 mismatch，再用 SBR 去除 bias 之辨認率為 58.42%。

(三) 連音聲學模型之建立

我們由 RNN 切割語音之研究結果發現，HMM 訓練過程有兩個缺陷，其一為音節邊界之切割常有偏差，此肇因於 HMM 訓練之 initial models 並非由正確之語音切割開始，使得一些聲音之 HMM models 和 silence model 有些混淆；另一為 HMM 對 inter-syllable coarticulation 嚴重之音段的 modeling 並不好。我們因此想利用 RNN 語音切割之結果來幫助 HMM models 的訓練。

首先我們觀察 RNN 語音切割對 coarticulation 嚴重之音節邊界常無法正確切割而產生一大段的 final segment，而對其他情形則能切割出正確的音節邊界。由此觀察，我們因此採取兩個下列行動：

(1) 使用 RNN 輸出加權來協助 HMM 訓練的語音切割步驟，以 RNN 的訓練結果為基準，根據 RNN 的輸出，我們對訓練 HMM model 所必須的對數觀測機率加入一加權值，此一加權是根據 RNN 的訓練結果所得來的。如此當 RNN 訓練出來的結果很確定為某一狀態時(即其 RNN 輸出很接近 1)，則我們在做 HMM training，會因加強了此一狀態的分數（觀測機率），而可增加此一狀態獲勝的機率。另外，當 RNN 訓練出來的結果為不確定狀態時(即 RNN 訓練結果屬於競爭狀態），則我們在做 HMM training 時，會因為每一狀態所增加的加權值相差不多，而變成單純由 HMM 訓練來決定 Viterbi search 路徑。
(2) 對 inter-syllable coarticulation 嚴重之音段建立額外的連音 final-initial HMM models，由 MAT-2000 語料，我們將較常出現嚴重 coarticulation 的 66 個 final-initial pairs 建立其 HMM models。辨認時使用 RNN 切割及連音 HMM models 協助找尋最佳之音節串。

實驗結果顯示步驟 (1)可微幅調整 HMM 訓

練之語音切割，而獲得較正確的音節邊界；步驟 (2)在目前則僅可稍微增進辨認率，此方法須做進一步改進。

(四) 韻律片語邊界之偵測

我們將 TCC-300 的 30%語料，以人工找出 major/minor breaks 來研究 prosodic modeling，研究題目包括：(1)已知音節切割做 break detection，(2)由 RNN 切割做 break detection，(3)對兩 breaks 間之 prosodic phrase 做 modeling。(1)將使用在聲學辨認之後處理，協助做 lingiuistic decoding；(2)將使用在聲學辨認之前處理，(3)則是希望建立好的 prosodic model 協助做 speech dialogue。目前已進行(1)，由相鄰兩音節之語音抽取 pause duration、energy、pitch、final duration 等 features，detection rate 約 97%。其餘兩項工作正在進行

(五) 音節長度模型之進一步研究

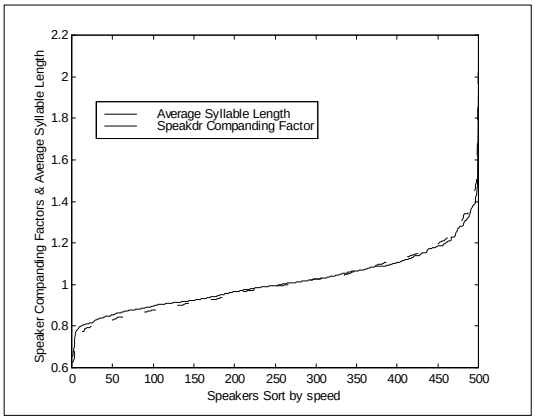
我們對上年度提出的 multiplicative 音節長度模型加以改進，研究項目包括：Tone 3 的 refinement、對 MAT 電話語料的 duration modeling、additive duration model 的比較、伸縮係數(Companding factor, CF)的分析等。

首先，我們考慮 Tone 3 的三個 patterns，包括：falling-rising (full tone)、middle-rising (sandhi tone)及 low-falling (half tone)，分別標示為 Tones 3、6 及 7，得到的 CF 如下表所示

Tone	1	2	3	4	5	6	7
CFs	1.01	1.03	1.04	1.02	0.85	0.92	0.82

由表中可看出，full-tone 最長，變嗓成二聲時長度縮短且較一般的二聲短，half-tone 最短，此結果與語言學上的 knowledge 相符。

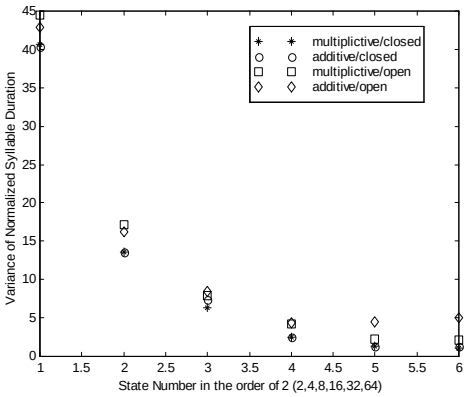
接著，我們考慮對 MAT 電話語料的 duration modeling，我們由 MAT-2000 中任意選出 500 人的句子語料，使用 100 聲母及 39 韻母 HMM models 切割，對音節長度進行 modeling，音節長度之 variance 由 66.78 frame^2 降為 2.54 frame^2，這顯示此音節長度模式 performance 很好。所估計出的 speaker CF 和 speaking rate estimated based on speaker-based average syllable duration 比較如下圖所示，由圖中可看出它們很相符。



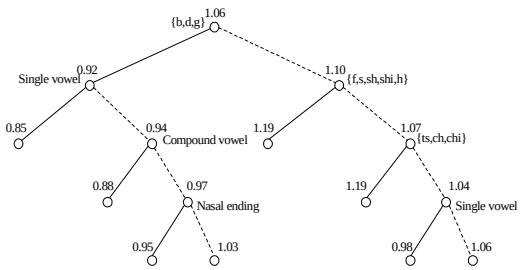
我們接著比較 multiplicative model 和 additive model，後者使用下式表示

$$Z_n = X_n + \mu_{t_n} + \mu_{y_n} + \mu_{j_n} + \mu_{l_n} + \mu_{s_n}$$

其中 Z_n 和 X_n 分別為 observed 及 normalized syllable durations，下圖顯示 normalized syllable durations 的 variance 的實驗結果，圖中顯示此兩個 models 的 performance 相當，並且我們經由統計分析發現它們所標示的 prosodic states 有很高的一致性，因此它們一樣好。

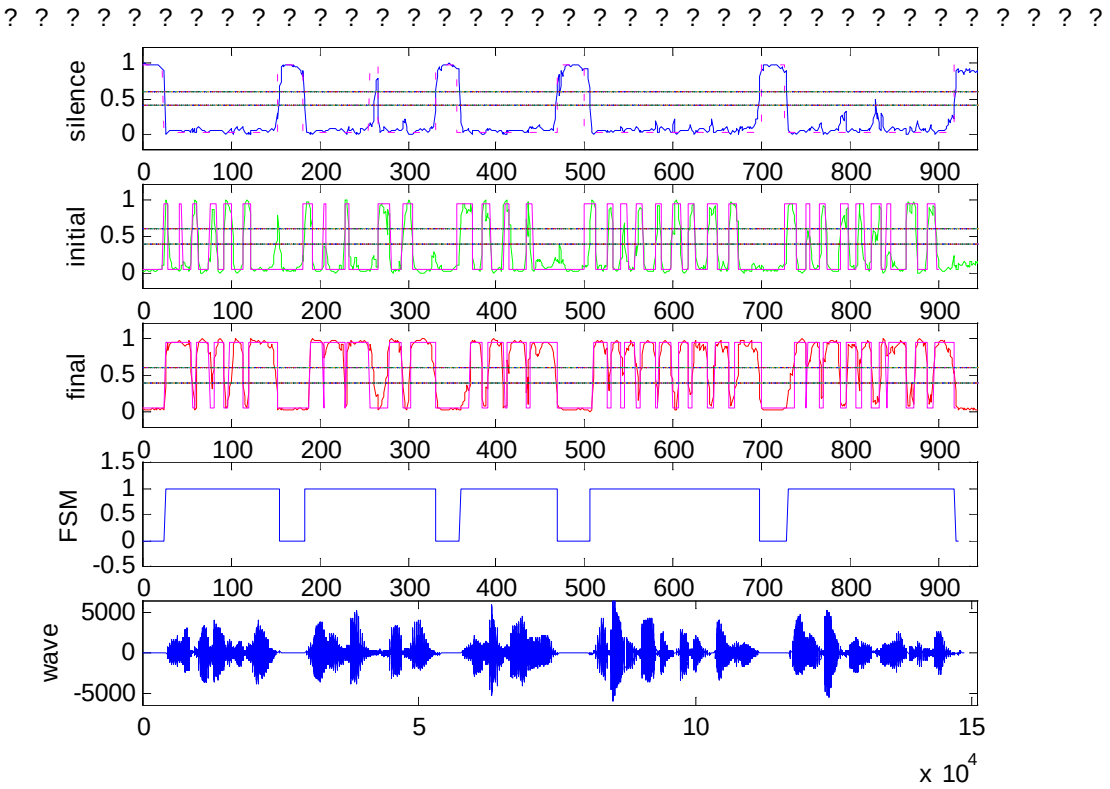


最後我們分析 model 的基本音節伸縮係數，以了解音節的音素組成成分對音節長度的影響，我們使用 decision tree method 來做分析，建立一個 hierarchical tree (見下圖)，由圖中顯示含 stop 聲母及單母音韻母之音節較短，含 fricative 及 affricate 聲母之音節較長。

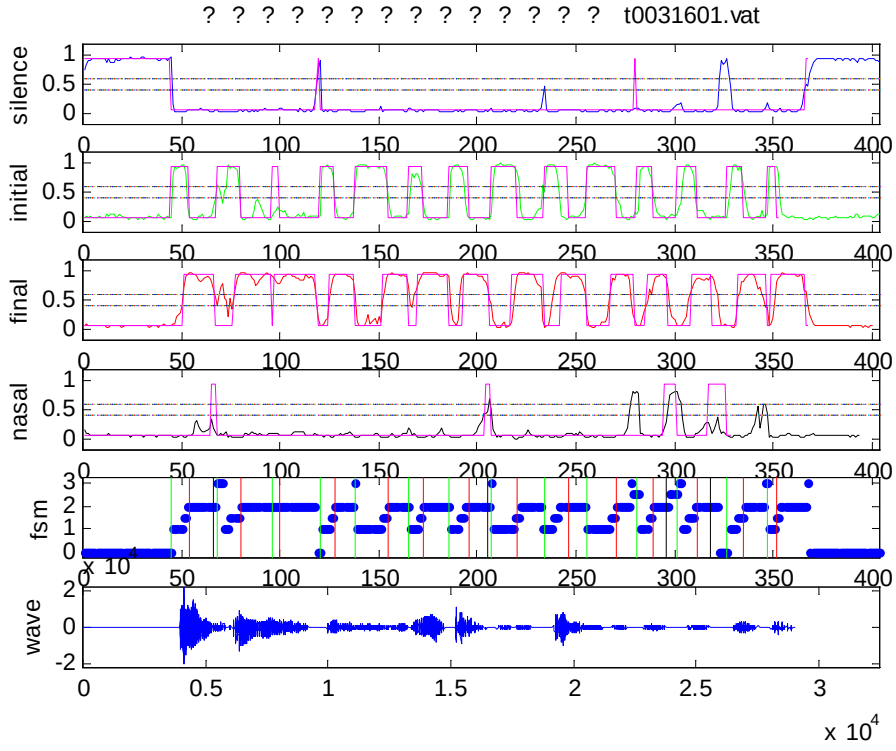


四、計畫成果自評：

計畫進行順利，與預定時程相符。



圖一. 使用 FSM 之音段粗分類例句：“以一天的時間藉演講、座談、心聲交流、活動引導等各種方式來引導婦女朋友成長”



圖二. 使用 FSM 之音段細分類例句：“但偶爾的失落感是在所難免的了”