行政院國家科學委員會專題研究計畫　期中進度報告

中文語音聲學模式及韻律模式之進一步研究(1/3)

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-009-128-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：國立交通大學電信工程學系

計畫主持人：陳信宏

計畫參與人員：郭威志、唐嘉俊、孫立諺

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 93 年 6 月 1 日

行政院國家科學委員會補助專題研究計畫 □ 成 果 報 告
∨期中進度報告

中文語音聲學模式及韻律模式之進一步研究(1/3)

Further Studies on Acoustic Modeling and Prosodic Modeling for Mandarin Speech

計畫類別：∨ 個別型計畫　　□ 整合型計畫
計畫編號：NSC－92－2213－E－009－128
執行期間：　92 年　8 月　1 日至　93 年　7 月　31 日

計畫主持人：陳信宏
共同主持人：
計畫參與人員： 郭威志、唐嘉俊、孫立諺

成果報告類型(依經費核定清單規定繳交)：∨精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
∨出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立交通大學電信工程學系

中　華　民　國　93　年　5　月　31　日

一、中文摘要

本計畫探討中文語音訊號之頻譜辨認參數及韻律參數之模式,在頻譜辨認參數模式方面，我們深入討論參數域之線性轉換，以及它在語者調適上的效果；在韻律參數模式方面，我們進一步討論基週軌跡之統計模式，以及它在聲調辨認上的應用。

關鍵詞：參數轉換、語者調適、基週軌跡模型、聲調辨認

Abstract

This report presents our studies on acoustic modeling for Mandarin speech recognition and on prosodic modeling. In acoustic modeling, we exploit the effect of feature-domain linear transformation on speech recognition and on speaker adaptation based speech recognition. In prosodic modeling, we improve the statistical pitch contour model proposed previously by considering the coarticulation effect in detail. An application to tone recognition is also discussed.

Keywords: feature transformation, speaker adaptation, pitch modeling, tone recognition

二、研究目的

語音辨認與合成技術在近年來有長足進步，但在推出一些試用系統後，使用者接受的程度仍低，顯示要達到實用階段仍有待學術上進一步的研究，本計畫擬以三年的時間探討兩個基本的問題，一為辨認參數之聲學模式，擬研究語音訊號的數學模式，考慮語者及環境的影響，以電話語料庫 (MAT-2000, MAT-2400) 及麥克風語料庫 (TCC-300)為研究對象；另一是韻律模式，擬研究中文韻律變化之數學模式，以及韻律和語法之間的關係，以自行錄製的單一女性語者語料為對象，text為中研院提供的tree-bank語料庫。

三、研究方法及結果

3.1 辨認參數之聲學模式

我們延續過去使用 affine transform 來考慮通道/語者之效應，使用下面之語音辨認參數模型

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

其中 $\mathbf{y}$ 和 $\mathbf{x}$ 分別為正規化和原始的語音頻譜參數向量，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：

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

其中 $\mu_{s_t m_t}$ 為 frame t 所屬 HMM model m_t, state s_t 之頻譜參數平均向量。藉由 minimize Q_k 可求出最佳的 A_k 及 $\mathbf{b}_k$ for speaker k。

使用此模型我們進行語者調適之語音辨認，圖 1 為使用 speaker-independent model 幫助語音切割以求取語者調適參數 A_k 及 $\mathbf{b}_k$ 之流程圖，圖 2 為語者調適之語音辨認測試方圍。使用十分之一的 MAT4500 約 450 之語料，以長句部分進行測試，實驗結果列於圖 3，由表中可看出當調適語料增加時，音節之辨認率隨之增加，在使用 6 句調適語音(約 26.7 seconds)時，辨認率超過使用 cepstrum mean normalization (CMN)方法，另一參考之基準是由每一測試語者之所有語料先求取調適參數 A_k 及 $\mathbf{b}_k$ ，再做辨認測試，它的辨認結果 64.6%可視為 upper bound。

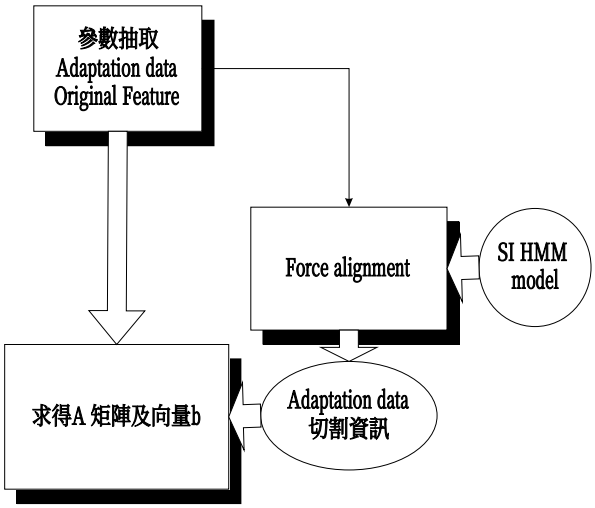

圖 1：從調適語料求取 A 及 b 之流程圖

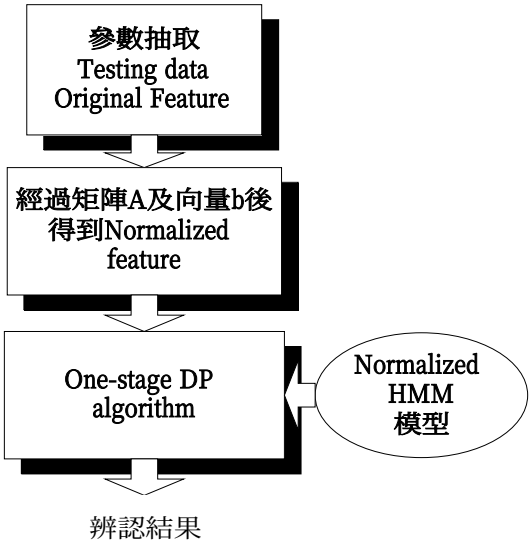

圖 2：語者調適之語音辨認測試

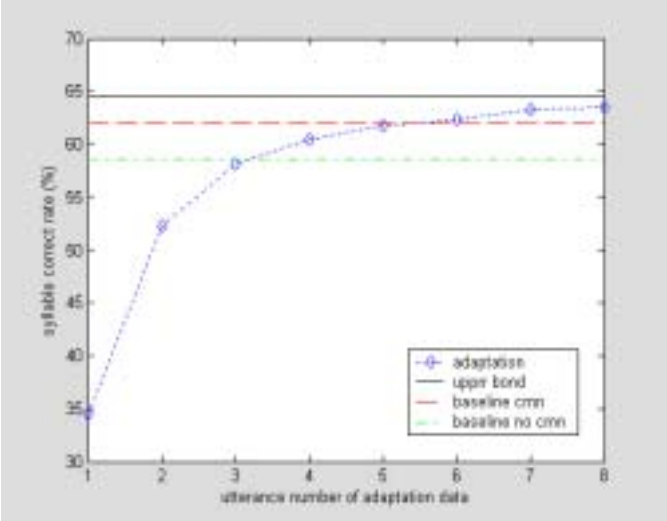

圖 3：語者調適之實驗結果

上述是基於 MSE criterion 來求取 A 及 $\mathbf{b}$，我們可以改用 ML criterion 來做，定義客觀的目標函數如下

$$Q = \sum_t (\mathbf{A}\mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_{S_t m_t})^{\mathrm{T}} \mathbf{R}_{S_t m_t}^{-1} (\mathbf{A}\mathbf{x}_t + \mathbf{b} - \boldsymbol{\mu}_{S_t m_t})$$

其中 $\mathbf{R}_{s_t m_t}^{-1}$ 為 frame t 所屬 HMM model m_t, state s_t 之頻譜參數共變異矩陣，圖 4 顯示語者調適之辨認結果，由表中可看出此方法之語者調適效果較差，但它有較高的 upper bound 辨認率 67.2%，這顯示它需要較多的調適語料來精確估計 A 及 $\mathbf{b}$，才能將辨認參數轉換以 match HMM model。

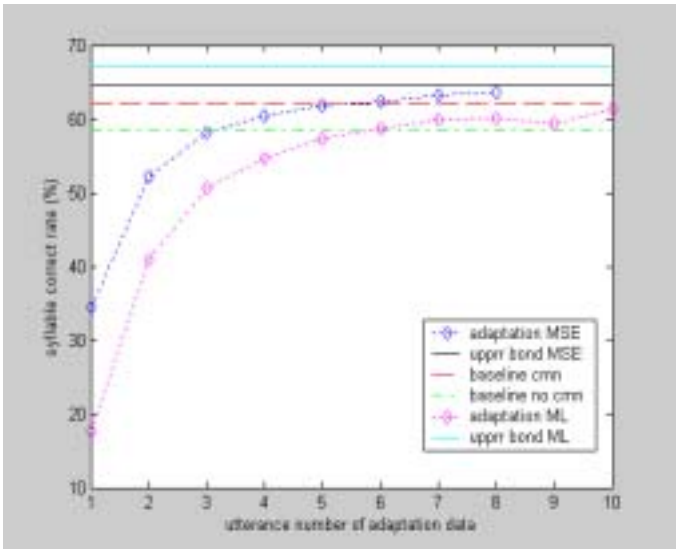

圖 4：使用 ML criterion 之語者調適實驗結果

我們更進一步考慮將語音信號的韻母(final)和聲母(initial)部分區分開，分別求韻母與聲母類型資料的矩陣 $A_{k,(v,u)}$ 及向量 $b_{k,(v,u)}$，而靜音(silence)部分採用下面三種方式考

慮：

1. 我們將所有訓練語料的靜音頻譜參數向量分群成 16 群，再針對每一個靜音頻譜參數向量是屬於哪一群，去對所屬那一群的平均值與共變異矩陣來求得矩陣 A_s 及向量 b_s。辨識時，每位語者靜音部份的轉換 A_s 及 b_s 與訓練與料用的轉換 A_s 及 b_s 是同一組。
2. 我們用較相近靜音部分的聲母轉換 $A_{k,(u)}$ 、 $b_{k,(u)}$ 來正規化靜音的頻譜參數。
3. 靜音部分不處理。

方法一所辨識出來的結果是 deletion error 數量大量增加，也就是大部份原本是 syllable 的音段被辨識成靜音。方法三所辨識出來的結果則是 insertion error 數量大量增加，也就是大部份原本是靜音的音段被辨識成 411 syllables，尤其是被辨識成帶ㄕ與ㄙ聲母的音節。方法二則獲得較好的辨認率，達 69.2%，如果我們事先知道語音的切割資訊，則辨認率可再提高至 77.6%，當然這違反測試的常規，只能拿來做為比較的參考。

由上述實驗結果顯示，將語音依性質分類，每一類採用不同的轉換參數，將可獲得較佳語音正規化效果，而提升辨認率。另外，silence 部分需要仔細處理，才能去除它帶來對音節辨認的干擾。我們將在後續的研究中加以考慮。

3.2 中文韻律模式

我們延續過去對中文的音節基週軌跡使用統計模型來描述它的變化的研究，首先對它加以深入分析，然後提出改進。此模式首先使用 frame-based speaker normalization

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

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

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

其中 tc_n 為考慮前後 tones 之組合、q_n 為韻律狀態。

此音節基週軌跡模式以一個 5 人的 TL-database 來評估，對訓練及測試語料的 RMSE (root mean square error)分別為 0.362 及 0.373 ms/frame，其中包含 0.17 and 0.19 ms/frame 的 orthogonal expansion errors。我們使用主觀測試(subjective test)來評估 reconstructed 基週軌跡，AB test 的結果顯示 41.25%的句子是使用原始基週軌跡的合成

語音較好，25%的句子是使用模式產生的基週軌跡的合成語音較好，而 33.75%的句子是沒有差別；另外，MOS test 的結果顯示使用原始基週軌跡和模式基週軌跡的合成語音之 MOS 值分別為 3.94 及 3.68，此結果顯示此基週軌跡模式效果良好。

我們再以 100 人的 TCC300 麥克風語音測試，此時採用自動的語音切割及基週偵測，其 RMSE 為 0.384 ms/frame，其中包含 0.172 ms/frame 的 orthogonal expansion errors。此結果顯示此基週軌跡模式對一般人也都有效。

我們也分析 prosodic state 所代表的意義，圖 5 顯示 prosodic state 所代表的音節基週均值序列，和原始值比較，它含較少因聲調(主要為 Tones 3 及 5)而產生的 zigzag 變化，可明顯看出逐漸下降的 prosodic phrase patterns。圖 6 畫出 prosodic state 所代表的音節基週均值序列之 autocorrelation function，和原始值之 autocorrelation function 比較，它的值較高，顯示它確為較平滑的序列，其最低值為 6 個音節，顯然和平均的 prosodic phrase 長度 6.14 音節相吻合。

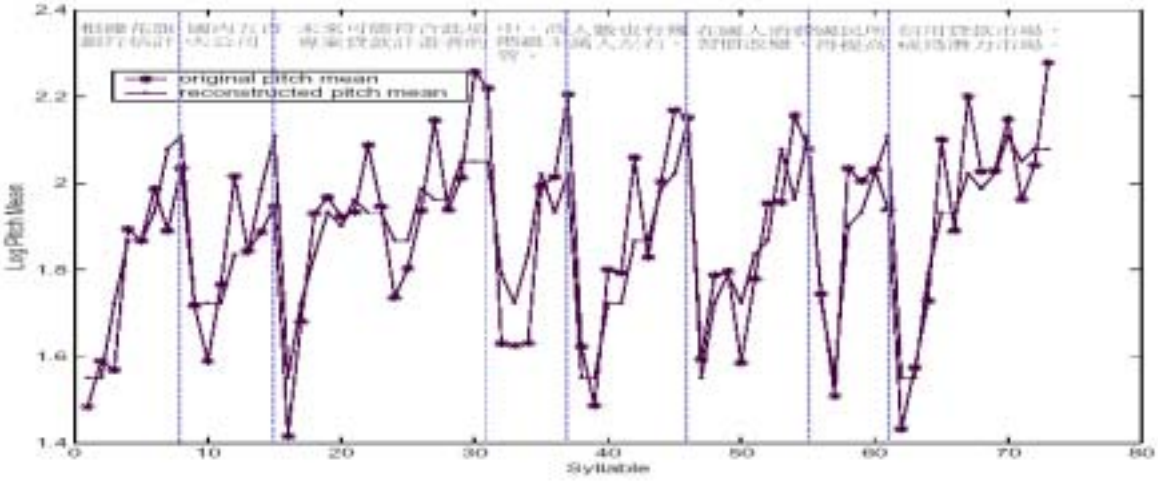

圖 5：prosodic state 所代表的音節基週均值序列和原始值之比較

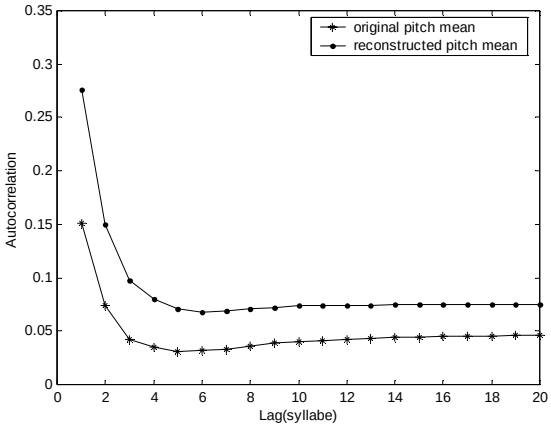

圖 6：prosodic state 所代表的音節基週均值序列和原始值之 autocorrelation function 比較

爲了進一步改進此基週軌跡模式，我們檢驗所有的參數後發現，此模式對相連接的音節基週軌跡的 modeling 效果較差，其原因在於人類為了節省力氣，在發此類音時，沒依照聲調該有的基週位準發出，而改以平平的聲調發出，因此造成此模式的過度聲調

補償，反而使 prosodic state 所代表的音節基週均值產生較大的 jump。

爲了補償上述基週軌跡相連的效應，我們將音節間連音效應 (coarticulation effect) 考慮進基週軌跡模型。首先，將相鄰音節之連音分成 4 類：

(1) Type 1：前後音節的 pitch contours smoothly 相連
(2) Type 2：pitch contours 不相連、沒有 pause、energy dip 高過 background silence energy level 一個 threshold (say 10 dB)
(3) Type 3：pause duration 不為零但小於一個 threshold (say 80 ms)，或是不合 Type 2 的 energy condition
(4) Type 4: pause duration 大於一個大的 threshold。

其中 Type 1 是基頻軌跡相連，表示前後音節之連音嚴重；Type 2 和 Type 3 是一般的音節間的關係，有些連音但不嚴重；Type 4 是前後音節間有一長停頓，彼此不相關。

依此連音分類，我們以單一語者的 treebank database 且先考慮 pitch mean model，將其調整為：

$$Z_n = Y_n = X_n + \beta_{t\_set}\Big|_{C_n} + \beta_{p_n}$$

其中 companding factor $\beta_{t\_set}\Big|_{C_n}$ 考慮四種前後音節間連音狀態組合 C_n：

$$C_n = \begin{cases} Type1 - Type1 \\ Type1 - !Type1 \\ !Type1 - Type1 \\ !Type1 - !Type1 \end{cases}$$

而使用多個不同的值

$$\beta_{t\_set}: \begin{cases} preTone - curTone - folTone & :125 \\ preTone - curTone & :25 \\ curTone - folTone & :25 \\ preTone \quad curTone \quad folTone & :10+5+10 \end{cases}$$

我們將在本年度完成此部分的 study，希望藉著仔細考慮音節連音效應而改進基頻軌跡模式。

另外，我們也將此基週軌跡模式用於 tone labeling for Taiwanese speech，此問題說明如下：我們在收集台語語音後需要標示每一音節之聲調，由於台語語音有複雜的 tone sandhi rules，語音的聲調和 text 標示的 lexical tone 會有所不同，如果使用人來標示，需要由專家來做，將耗費極大的人力，且長時間工作或多人一起工作都會有標示不一致的問題，因此需要對自動化的標示方法加以研究。直接的做法是對音節的基週軌跡進行分類，例如使用向量量化做分群，對每一群標示其最可能的 tone，但因音節的基週軌跡會受前後音節以及 intonation pattern 的影響，直接分類的做法項果將較差。因此我們使用前述的音節基週軌跡統計模式，嘗試將 tone 的 variation 考慮在此 model 中，同時去除前後 tones 以及 prosodic state 的影響，在做完 training 後，直接標示最可能的 tone。

我們使用一個含 23633 音節之單一男性語料進行實驗，表 1 列出使用 model 的方法和兩個 VQ 法的比較，其中 VQ-4 和 VQ-3 分別使用基週軌跡 4 個 orthogonal expansion

參數和 3 個代表 shape 的參數，由表中可看出 modeling method 將音節標示正確率由 52.4%提升至 61.9%，此結果顯示 modeling method 是有效的。

表 1： The correct rates of the three tone labeling methods of VQ-4, VQ-3, and MODEL. (unit: %)

Tone (sandhi tones)	1 (7)	2 (1)	3 (2,1)	4 (2,1,8)	5 (7,3,7)	7 (3,7)	8 (3,7,4)	Ave.
VQ-4	61.9	82.9	55.4	40.9	28.1	34.0	33.9	50.9
VQ-3	58.7	84.8	44.1	28.7	43.7	47.2	35.8	52.4
MODEL	72.4	89.3	51.7	55.7	50.6	51.1	41.9	61.9

四、成果討論

我們在計畫中對中文語音訊號的頻譜辨認參數及韻律參數之模式進行深入探討，未來將可應用於改進中文語音辨認系統之辨認率及 TTS 系統之自然度。

五、計畫成果自評

計畫執行符合進度，成果已陸續寫成論文投稿，執行成效尚稱良好。