行政院國家科學委員會專題研究計畫　期中進度報告

口語語音辨認與韻律模式之研究(1/3)

計畫類別：個別型計畫
計畫編號：NSC91-2219-E-009-048-
執行期間：91 年 08 月 01 日至 92 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：王逸如

計畫參與人員：江振宇、孫立諺、唐嘉俊

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 5 月 30 日

行政院國家科學委員會專題研究計畫成果報告

口語語音辨認與韻律模式之研究(1/3)

計畫編號：NSC91－2219－E－009－048

執行期限：91 年 8 月 1 日至 92 年 7 月 30 日

主持人：王逸如　國立交通大學電信工程系

計畫參與人員： 江振宇、孫立諺、唐嘉俊

一、中文摘要

本三年計畫針對國語口語語音辨認及韻律模式做探討，以其建立一套國語口語語音辨認系統。第一年在國內現在錄製中之國語口語語料庫尚未取得前，將先就朗讀語音(reading speech)語音辨認系統做改善，先改善 read speech 國語音節辨認器的 robustness，將針對說話速度及語者效應之移除作探討。並開始整理口語語料及基本性質探討。

關鍵詞：口語語音、隱藏式馬可夫模式、階層式信號偏移量移除法、語音信號變化率

Abstract

The speech recognition and prosodic modeling for spontaneous Mandarin speech will be studied in the three-year project. In the first-year progress, two speaker normalization methods, VTN and hierarchical SBR, were used to improve Mandarin speech recognition rate in reading speech database. Then, a new speaking rate measure, the rate of speech variation (ROSV), was defined and used to detect the speaking rate of training utterance. It was used to normalize the dynamic recognition features due to the variation of speaking rate in order to improve the syllable recognition rate. Finally, a basic Mandarin speech recognizer was established in this year.

Keywords: Spontaneous speech, HMM, hierarchical SBR, ROSV

二、 緣由與目的

口語語音辨認是現今國內外語音辨認研究學者正積極進行研究中的一個課題。口語

語音存在一些語言學上的現象是 reading speech 中沒有的，如：(1)呼吸聲、笑聲等非語音信號，(2) 贅音(filled pause)，(3) 不流利(influencies)的現象等問題，上述問題有些必須在上層語言處理來解決，但就聲學辨認之觀點也必須再深入研究，改進口語語音之聲學模式，以提高國語口語語音之音節辨認率，尤其是口語語音中之說話速度變異十分之大。

在國語口語語音中，有一些特性與歐美拼音語言中之特性不太相似，例如：英文之贅音常是一些有語法結構之字(如 you know)；而國語則沒有。所以不能歐美拼音語言之作法。除此，使用在語音辨認中除了辨認其音節內容外，還含有許多其他資訊，例如：語音信號中之韻律信息，尤其是在口語語音中，這些信息將可以輔助音節辨認、語言解碼(language decoding)、語意了解(understanding)。但這些韻律信息是一些隱含的資訊，如何蒐取這些資訊將音訊(acoustic)及語言模式(language model, LM)緊密結合則是一個研究中的課題。

三、研究方法

在第一年計畫中，主要使用 MAT-4500 (MAT-2000+2500)及 TCC-300 兩個朗讀語料庫用以製作計劃後兩年中所需之基本語音辨認器。

1. 基本 HMM 辨認系統

MAT 語料庫為經電話線路志之語料，其各項統計資料如表一所示。

表 1. MAT 語料庫統計表

類別	性別	人數	總人數	音節數		時間(小時)	
MAT 2000	男	1002	2229	115,940	260,133	13.62	31
	女	1227		144,193		17.38	
MAT 2500	男	1265	2570	163,364	324,712	17.86	36.2
	女	1305		161,348		18.33	

首先我們先對不同性別各自建立 100 韻母相關之聲母及 40 個韻母 HMM 模型，使用 38 維特徵參數(MFCC、ΔMFCC、ΔΔMFCC、ΔEng、ΔΔEng)混合式高斯機率觀測值，高斯分布個數則隨各狀態語料多寡而定，最多使用 32 個，狀態長度則使用 Gamma 分布，對男女 411 音節之各狀態分建立其狀態長度模型。並將 MAT 語料之 9/10 (508,146 音節)訓練語料，1/10(56,691 音節)作為測試語料。所得之辨認率為 57.77%(插入型錯誤:2.87%，刪除型錯誤:1.50%)。

2. 語者效應之移除

接著我們將各種語者及通道效應去除的方法，如：SBR (32 codewords)、CMN 等加入基本辨認系統中，在 SBR 方法是先建立一組 VQ 碼書，然後使用 MMSE 方法來找各音框之參數偏移量，我們提出一個改進方法將 SBR 中之 MMSE 要求改為使用 ML (Maximum likelihood)，於是我們建立一組高斯分佈，再使用 ML 方法來找各音框之參數偏移量，我們在下面稱之為 ML SBR。

表二. 以語者及通道效應去除的方法之較能比較。

	插入率	刪除率	辨認率
基本系統	2.87%	1.50%	57.77%
SBR	2.40%	1.36%	60.76%
CMN	2.41%	1.28%	61.02%
None-SBR	2.27%	1.45%	59.38%
ML_SBR	2.44%	1.31%	61.13%

由上面結果可以發現一些有趣的事情，(1) CMN 方法較 SBR 簡單但對辨認率之改善不會比 SBR 差；但(2)僅在測試時加上 SBR 移除語者及通道效應(表二之第四行)也可大幅改善辨認結果，CMN 不可僅在測試步驟做，因為它會對所有語料扣除一個方向一致的偏移量。

在檢查所建立之 HMM 模式後發現狀態長度則使用 Gamma 分布會有變異數(variance)過小的情形，會影響辨認結果，於是對 Gamma 分布變異數設限，辨認率可提升至 61.41% (插入型錯誤:2.38%，刪除型錯誤:1.28%) 。

接著，我們使用計畫中所提出之使用 hierarchical codeword 之 SBR 方式，也就是使用多次的 SBR，但每次的 codeword 數是遞增的，這樣在 codeword 數較小時能避免 SBR 會將輸入參數誤判 codeword 造成偏移量錯誤，在 codeword 數多時因已移除部份偏移量，偏移量機率減小且精確度會提高。我們做了一個實驗，使用 3 次 SBR，codeword 數分別為 1、8、32，結果辨認率為 61.98%(插入型錯誤:2.23%，刪除型錯誤:1.28%)，辨認率提升了 0.57%。

3. 使用 VTN 移除語者效應

在 VTN 移除語者效應之研究，我們使用的是另一個語料庫 – TCC-300，因為它是一個麥克風輸入、16KHz sampling rate 的語料庫，我們在第二年起所使用的國語口語語音研究中所能取得之語料庫均為麥克風輸入。

表 2. TCC 語料庫統計表

類別	性別	人數	總人數	音節數		時間(小時)	
訓練	男	137	274	150,344	300,856	12.08	24
	女	137		150,512		11.96	

測試	男	15	29	16,297	324,712	1.28	2.4
	女	14		15,114		1.17	

我們首先使用經 SBR 做去除語者效應之 HMM 辨認器做參考系統，其辨認率為68%(插入型錯誤:0.75%，刪除型錯誤:1.34%)。

我們使用標準 VTN 方法[2]做語者效應去除，使用之 warping factor 為 0.88-1.12(13組)。發現其辨認率為 65%(僅較不作補償高 1%)，且其運算量增加了 13 倍，其辨認率改善幅度並不如原論文中所得到之那麼理想。

接著我們使用在語者辨認中十分成功的GMM (Gaussian Mixture Model)方法來做warping factor的預判別，若能快速判別warping factor即可將VTN方法之複雜度降低則VTN方法可與SBR共同使用來移除語者效應，但我們發現warping factor 的預判別僅達60%，使用預判別的warping factor後，音節辨認器之辨認率較不補償僅上升0.1%，所以往後將不使用VTN語者正規劃方法。

但在此我們做了一個小實驗，因為第二年我們要用 TCC-300 訓練出的 HMM 模式去做國語口語語料之切割(已知字串)，在國語口語語料上未準備就緒前，我們使用了一個自行錄製 20Khz 取樣頻率之語料庫(文字內容為 TreeBank 資料庫之文章)，並以人工切割了 1200 字，比較兩者字首字尾之切割位置發現其誤差之均方值為 0.2 音框。

4. 語者說話速度之調適

在一段連續語音中，語者說話速度的快慢會直接影響到音節的長度和音框與音框間的變化關係。

我們認為語者的說話速度主要會影響到特徵參數中的差量化 MFCC(ΔMFCC、ΔΔMFCC)部分，而對於 MFCC 的影響較大的則是連音(coarticulation)效應或其它語言效應。因此，我們主要即針對特徵參數中的差量化 MFCC，原來 MFCC 差量是 MFCC 參數對一組正交基底展開求得，

$$\left(1, \frac{i}{\sum_{i=-3}^{3} i^2}, \frac{i^2-4}{\sum_{i=-3}^{3}(i^2-4)^2}\right)$$

其中 $i \in [-3, 3]$，而此正交基底是由 $(1, i, i^2)$ 正規化所求得。

現在，為估計語音信號中說話速度的效應，我們假設在語音信號中存在一個 time scale factor γ，所以我們將基底置換為 $(1, \gamma i, (\gamma i)^2)$，經正規化後，正交基底則變成

$$\left(1, \frac{\gamma i}{\sum_{i=-3}^{3}(\gamma i)^2}, \frac{(\gamma i)^2 - 4\gamma^2}{\sum_{i=-3}^{3}((\gamma i)^2 - 4\gamma^2)^2}\right)$$

若 $\Delta\tilde{x}_t, \Delta\Delta\tilde{x}_t$ 是原本的 MFCC 差量，而 $\Delta x_t, \Delta\Delta x_t$ 則是考慮說話速度時的正規化後之 MFCC 差量，則 $\Delta x_t = \gamma\Delta\tilde{x}_t$，　$\Delta\Delta x_t = \gamma^2\Delta\Delta\tilde{x}_t$ 。

我們可將 time scale fator 換置為速度轉換參數 k, $k = 1/\gamma$ 。我們將 k 稱之為語音變化速度(Rate Of Speech Variation, ROSV)。

基於上述假設，若要補償因 ROSV 變化造成的影響，由於我們在求取語者說話速度時僅考慮 Δx_t 的變化量而不考慮其變化的方向，因此我們可對 Δx_t 取 norm，來求得其變化的大小。而觀察特徵參數的特性，我們會發現，特徵參數會隨著維度的增加而遞減，因此為了不使第一維的差量特徵參數主宰我們所估計的語者說話速度，我們定義一個量測量

$$\Delta\hat{x}_t(d) = \frac{\|\Delta x_t(d)\|}{\sqrt{E(\Delta x_t^{\,2}(d))}} \qquad (1)$$

此量測量與 ROSV 之相關性十分強。接著，我們考慮如何由上面定義的 $\Delta\hat{x}_t$ 來估計 ROSV。首先，我們假設 HMM 每一狀態中去除說話速度後之 $\gamma\Delta\hat{x}_t$ 參數成一高斯分佈，

$$p(\gamma\Delta\hat{x}_t \mid s_t, m_t) = \sum p_i(\gamma\Delta\hat{x}_t \mid S_t, m_i)$$

$$= \sum_i c_{i,s_t,m_t} N(\mu_{i,S_t,m_t}, U_{i,S_t,m_t})$$

其中 c_{i,S_t,m_t} 表示在 m_t 模型、S_t 狀態中第 i 個混和的加權值，而 coraiance matrix U_{i,S_t,m_t} 為 diagonal。

利用 ML 的方法來做估測一個音框之 γ_t ，

$$\gamma_t = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad (2)$$

其中

$$a = \sum_t \sum_i p(i \mid \Delta x_t, s_t, m_t, \gamma)\left(\frac{\Delta\hat{x}_t^2}{\sigma_{i,S_t,m_t}^2}\right)$$

$$b = \sum_t \sum_i p(i \mid \Delta x_t, s_t, m_t, \gamma)\left(\frac{\mu_{t,S_t,m_t}\Delta\hat{x}_t}{\sigma_{i,S_t,m_t}^2}\right) \quad p(i \mid \Delta\hat{x}_t, s_t, m_t) = \frac{p_i(\gamma\Delta x_t \mid s_t, m_t)}{p(\gamma\Delta x_t \mid s_t, m_t)}$$

$$c = -\sum_t \sum_i p(i \mid \Delta x_t, s_t, m_t, \gamma)$$

而句子之 ROSV $k = 1/\gamma$, $\gamma = \sum_{t=1}^{N} \gamma_t / N$ 。

1. 根據(1)式我們將訓練語料的特徵參數做語者說話速度的正規化，初始的語者說話速度預設為 1。

2. 求 Δx_t 的變化量，並對 $\|\Delta x_t\|$ 做正規化。

3. 根據原模型對訓練語料庫所訓練出的語句切割位置，我們可訓練出用來估計語者說話速度的模型。

根據步驟 3 的模型和(2)式，我們可找到新的語者說話速度。如圖 1. 所示。

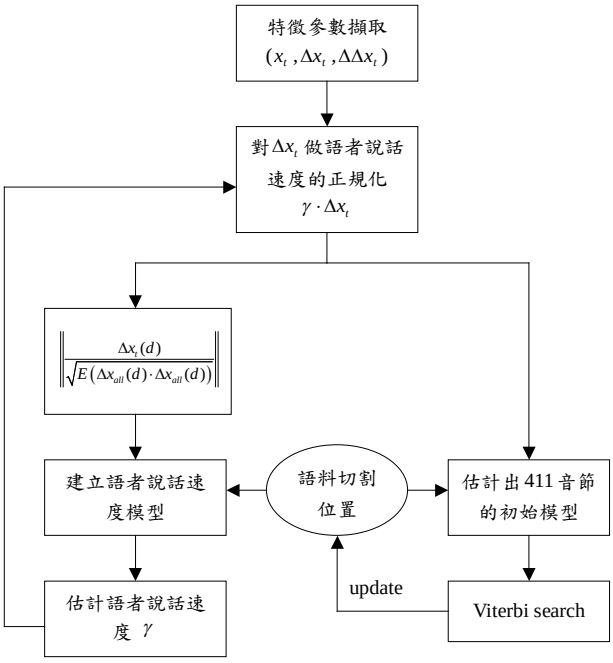

圖 1. ROSV 估計流程圖。

在此實驗中我們採用 TEST-500 做為測試語料(4,731 音節)。使用上述方法求得之 ROSV 值分佈如圖 2. 所示。再使用估計之 ROSV 做輸入參數差量化正規劃並做狀態長度分佈之正規劃，重新訓練 HMM 模型，並重新辨認其辨認率如表 3. 表示。在表 3. 中我們將辨認率分為 ROSV>1 及<1 兩部份，圖 3. 中則更可看出對 ROSV 作補償前後，辨認率與 ROSV 之關係。在 ROSV 較大時辨認率反而下降，我們認為 ROSV 較大時影響辨認的原因會是連音效應，它必須靠建立連音甚至詞組的辨認模型才能解決。而 ROSV 非常小時之辨認率較平均值低甚多，在國外早就提出 connected speech 與 continuous speech 應該分別建立辨認器。由我們的實驗結果發現，語者說話速度之調適這個課題還是需要做更深入的研究。我們會在第二年計畫中對口語語料繼續探討說話速度補償之問題。

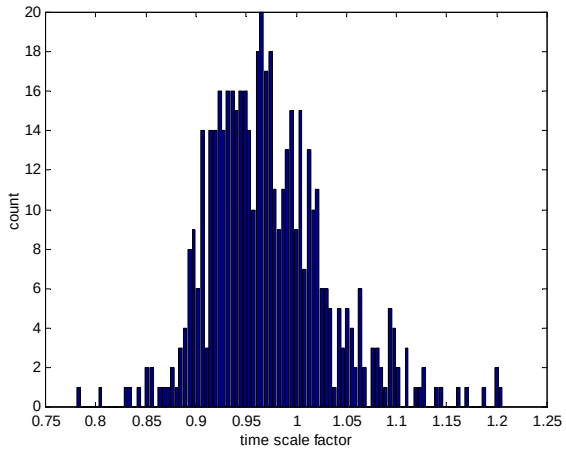

圖 2. 測試語料之 ROSV 分佈。

表 3. 經說話速度調適前後之辨認率。

ROSV		插入率	刪除率	辨認率
<1	調適後	1.67%	2.40%	59.54%
	原始前	1.88%	2.40%	58.81%
>1	調適後	1.75%	1.35%	67.05%
	原始前	1.19%	1.54%	67.36%

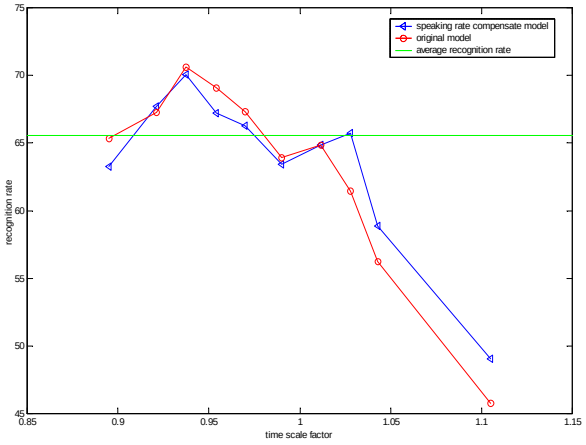

圖 3.辨認率與 ROSV 之關係圖。(橫軸為γ=1/ROSV)

5. 國語口語語料庫之整理

國語口語語音語料目前國內有兩個語料庫正在進行蒐集中，一是中研院王新民博士所負責蒐集之廣播新聞節目，已有 prerelease 版，其 transcription 資料是以 XML 格式標示，目前已將 transcription 從標示檔抽出及將音檔切割成對應的 term，因為無標音資訊現正利用 TTS 系統中的 parser 標示標音中，預計七月前可將標音及人工檢查工作完成，這個語料庫之缺點是多為獨話(monologue)。另外，中研院曾淑娟博士也錄製一個『現代

漢語對話語料庫』(Mandarin Conversational Dialogue Corpus、MCDC)[5]，其內容均為對話(dialogue)，transcription 內容中已有標音，在此語料庫發行後，應會使用此語料庫繼續第二年之計畫。

今年做了口語語音初步的探討發現口語語音中贅音出現的機率遠較英文高，約有9-10%，而像呼吸聲、嘆氣聲等語言現象也有 3-5%，甚至有一些含糊不清的語音也有3-5%。上述口語語音中之現象，約佔口語語音中 20%[5]。這些現象將會使國與口語語音之辨認率大幅下降，現也正在觀察這些現象中。

6. 結論

本年度計畫中所建立之辨認器，雖然辨認率改進幅度不大，但其強健性已可使用於口語語音辨認之上，在第二年度處理口語語音時預期會遇到的問題將會更多，且提前取得國語口語語料，所以已提前做了一些口語語料的前處理工作。

四、計畫成果自評

在計畫書中所列舉之項目均已執行完畢，所建立之辨認器將可在第二年計畫中使用。並已提前做國語口語語料庫之整理工作。

五、參考文獻

[1] 呂儲仰，“國語連續音節辨認系統之改善與分析”，國立交通大學碩士論文，民國九十年六月。

[2] Li Lee, and Richard Rose, “ A frequency warping approach to speaker normalization “, pp. 49-60, IEEE Trans. on AU, Vol. 6, No. 1, Jan. 1998.

[3] D. A. Reynolds, “ Speaker Identification and Verification Using Gaussian Mixture Speaker Models, “, Speech Communication, Vol. 17, pp. 91-108, Aug. 1995.

[4] M. G. Rahim, and B. H. Juang, “ Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition “, IEEE Trans. on SA, vol. 4, pp. 19-30, Jan, 1996.

[5] 曾淑娟、劉怡芬，“現代漢語對話語料庫標注系統說明” ，中央研究院中文詞知識庫小組，技術報告02-01。