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(2/3)

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-009-046-
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：王逸如

計畫參與人員：魯柏暄、蕭希群、陳振模、許君豪

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 5 月 31 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(2/3)

計畫編號：NSC92－2213－E－009－046
執行期限：92 年 8 月 1 日至 93 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： 魯柏暄、蕭希群、陳振模、許君豪

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. 第二年中首先整理國內現有之兩套口語語料庫 PTSDN 及 MCDC，並利用閱讀語音(TCC-300)所建立之 HMM 辨認模型去做口語語音之切割, 接著利用上述切割位置產生口語語音 HMM 起始模型, 進而建立基本口語語音辨認器－其辨認模型除 411 音節還包含贅音、非語音信號之語言現象及填充模型，接著評估所建立國語口語語音辨認模型之效能及觀察造成錯誤之原因, 最後再加上常用音節連並詞組之聲學模式及語言模式以改善系統辨認率。

關鍵詞：口語語音辨認、贅音、非語音信號之語言現象、音節連並

Abstract

The speech recognizer of spontaneous Mandarin speech was established this year. Two Mandarin spontaneous databases – PTSDN and MCDC were used in the study. The speech data in both databases were first segmented by using the HMM model trained by read speech. Then, the 411 syllable, particles, paralinguistic phenomena and other non-411 syllable HMM models were trained according to the initial segmentation. And, the performance of the Mandarin spontaneous speech recognizers was evaluated and some error analyses were done. Finally, the language model and acoustic model of contraction syllable-pair was added to the recognizer in order to improve the performance of system.

Keywords: Spontaneous speech recognition, particle, paralinguistic phenomena, syllable contraction

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. 口語語音存在一些現象是 reading speech 中沒有的，如：(1)呼吸聲、笑聲等非語音信號之語言現象(paralinguistic phenomena)，(2) 贅音(filled pause, particle)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。

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. 所以不能完全仿造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)。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。

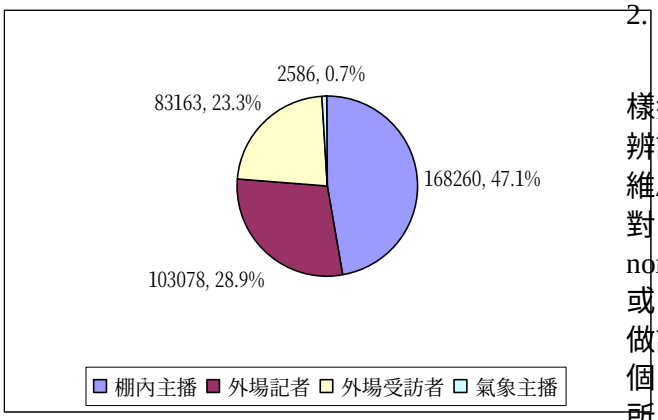
三、 研究方法

1. 國語口語語料庫之整理

國語口語語音語料目前在國內已可使用有兩個口語語料庫，一為公共電視廣播新聞(Public Television Service News Database, PTSND) [1]，它使用 LDC 的軟體 Transcriber 來做 transcription, transcription 資料是以 XML 格式標示，本計畫將第一年 40 小時與第二年 80 小時之語料加以整理, 先將 transcription 文字從 XML 格式之標示檔抽出及將音檔切割成對應的 turn，並利用 TTS 系統中的 parser 將 BIG5 資料轉為標音。因為節目中有許多是音樂、有背景音樂之語音信號、廣告等未做 transcription 之語料，將上述資料剔除後，僅剩 20 小時的可用語料，其中語料又依語者分為內場主播、氣象主播、外場記者與外場受訪

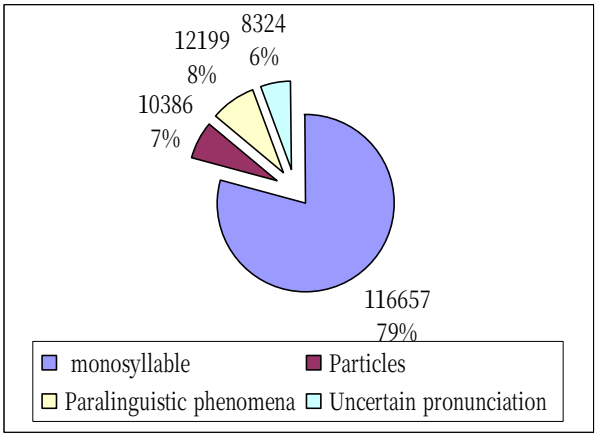
者，其上述環境分類其資料分佈如圖一所示(氣象主播之資料僅有小部分有 transcription 資料)。此口語語料庫之缺點是多為獨話(monologue)且主播之語料佔了一半。

圖 1. PTSND 語料庫統計圖(1,2 年語料)



計畫中使用的另一個口語語料語料庫為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)[2]，其內容共有 8 段對話(dialogue)已完成 transcription，其 transcription 資訊為奇特有檔案格式，在 transcription 資訊中已有漢語拼音資訊，且在語料中 read speech 中沒有的口語語音現象:如贅音(particle)，呼吸聲、嘆氣聲等語言現象(paralinguistic phenomena)，甚至有一些含糊不清的語音(uncertain pronunciation)等口語語音中之現象較 PTSND 多，其分佈如圖二所示，其中 411 音節所佔之比例僅 79%。但語料中之音節數較少，僅 14.7 萬字。MCDC 語料庫中之 transcription 中時間資訊都以 sub-turn(一個 turn 為一組對話)給定，所以我們亦將語音資料沏為 sub-turn。

圖 2. MCDC 語料庫統計圖



表一中則為 PTSDN 語料庫中各種語言現象之分佈，可以發現口語語音中特有之現象如 particles 之數量遠少 MCDC 語料庫，因為 PTSDN 中資料量最多之主播部份還是較接近 read speech。

表 1．PTSND 語料庫語音種類統計表

Model	Mono-syllable	breath	Particle	Inap _pron	English	MinNan	others
比例(%)	93.56	2.17	1.11	0.52	0.12	0.08	0.05

2. 基本口語語音 HMM 辨認模型之建立

在計畫中 PTSDN 或 MCDC 語料均相取樣頻率降至 16KHz，使用每秒 100 個音框，辨認參數是 12 維 MFCC、12 維ΔMFCC、12 維ΔΔMFCC、ΔENG、ΔΔENG，共 38 維；並對每一語句做 CMN (cepstral mean normalization)移除部分語者效應。在 PTSDN 或 MCDC 語料庫中，取十分之九 turn/sub-turn 做訓練語料，其餘十分之一作為測試語料。每個 turn/sub-turn 中所含之音節數多則上百個，所以我們在辨認模型的訓練時若使用 uniform segmentation 來做起始模型，可能會有非常大的誤差；將導致模型訓練時收斂至較差之結果。所以在計畫中我們使用 read speech (TCC-300 語料庫)所訓練的 411 音節 HMM 模型來做已知字串之切割，對 read speech 中沒有相對辨認模型的音節；如：particles、paralinguistic phenomena(呼吸聲、嘆氣聲、…)等音，我們採取下面步驟解決：

(1) 先以發音相近之 411 音節取代；
(2) 對呼吸聲等出現次數高之 paralinguistic phenomena，且無適當 411 音節可以取代者，以人工切割一部分來建初始模型；
(3) 對部分出現次數少的音；如：外來語、含糊不清的語音等先以所有非靜音之語料建立一個 filler 模型來描述這些語音資料；這個模式可以用來對應至我們無法建立正確聲學辨認模型之所有語音信號及非語音之口語現象，

如此我們可以將所有訓練語料，包含語句中含有口語語料特有現象之語句，做已知字串之切割(force alignment)，在獲得起始切音位置後，我們特別檢驗我們所採用的 filler 模型之切割位置是否正確，經檢查其切割位置確實可以正確對應至我們無法建立正確辨認模型之語音信號。

接著，我們使用 HTK [3]來做 HMM 語音辨認器之訓練，所建立之 model 除 411 音節外尚有 particles、paralinguistic phenomena(呼吸聲、嘆氣聲、…)及 filler model。PTSDN(僅使用第一年語料)或 MCDC 語料所使用的模型數如下表所示，其中各 HMM 模型每一狀態所使用的 mixture 數及之訓練資料多寡決定。

表 2．PTSND 語料庫使用之模型數

	411 syllable	Particles	Paralinguistic phenomena	Filler
Hmm Model 的數量	100 RD-initial 40 final	19	1(breath)	1
State no.	Initial:3 Final:5	3	3	3
mix. no.	Max. 16	Max. 16	Max. 16	32

表 3．MCDC 語料庫使用之模型數

	411 syllable	Particles	Paralinguistic phenomena	Uncertain pronunciation
Hmm Model 的數量	100 RD-initial 40 final	40	13	76
State no.	Initial:3 Final:5	3	3	3
mix. no.	Max. 32	Max. 32	Max. 32	Max. 32

其中我們為出現次數足夠之含糊不清/發音錯誤的語音(uncertain/inappropriate pronunciation)不使用原字音而另外建立模型是為了不污染原音節之模型且求得較佳之切割位置，出現次數不足之音則使用 1 及 3 個 state 的 filler 模式。

首先我們看 PTSDN 之辨認率，如表 3 所示。但 PTSDN 語料庫中之資料有內外場之分；內場語音品質較佳且語者數目有限，如果我們將內外場之辨認率分開統計，內場辨認率已可達 67.7%，與 TCC-300 之辨認率已相近。外場語料則語音品質較差，辨認率仍十分低。

表 3．PTSND(第一年語料)辨認結果

	Outside	Inside	內場 (outside)	外場 (outside)
字數	19367	16629	9369	4852
句數	220	219	109	57
Del	2.4%	2.6%	2.2%	4.0%
Sub	35.7%	30.7%	26.0%	52.5%
Ins	9.3%	8.0%	4.1%	19.4%
正確率	52.6%	58.7%	67.7%	24.2%

接著我們看 MCDC 之辨認率，如表 4 所示。其辨認率明顯 PTSDN 語料低許多。初步分析其原因：(1)MCDC 語料庫中口語語音現象:如贅音(particle)，呼吸聲、嘆氣聲等語言現象(paralinguistic phenomena)，含糊不清的語音(uncertain pronunciation)等口語語音中之現象較 PTSND 多，所以造成插入型錯誤之增加；(2) 在 MCDC 語料庫之標示中事實上提供了『音節連並』(contraction)之標示，若統計音節連並在語料庫中出現次數多達 20%，他們會造成嚴重的刪除型及取代型錯誤。而『音節連並』(contraction)又並非左右文相關辨認模式能解決，因為它們的音素結構都已改變了。

表 4．MCDC 辨認結果

	Consider all acoustic models	Consider 411 only
字數	17704	14396
Del	5.0%	6.0%
Sub	40.5%	37.0%
Ins	12.5%	13.0%
正確率	41.4%	41.35%

3. 口語語音辨認系統之改進及錯誤分析

對 PTSDN 語料庫因為其語者環境大略可分為 3 類:主播、外場記者及受訪者，且加上第二年語料後資料足夠建立不同環境之辨認器，所以我們就建立 3 組辨認器以克服環境差異因素。其辨認結果如表 5 所示。主播之辨認率以高達 75%，因為它僅能視為 multi-speaker read speech 情況。受訪者則因較接近口語語料且音質較差所以辨認率僅 41%，與 MCDC 語料之辨認率接近。

表 5．PTSND 語料庫使用之模型數(使用 1,2 年語料)

	411 syllable	Particles	Paralinguistic phenomena	Filler
Hmm Model 的數量	100 RD-initial*3 40 final*3	35	1(breath)	3 (English,Min Nan,others)
State no.	Initial:3 Final:5	3	3	3
mix. no.	Max. 32	Max. 32	Max. 32	Max. 32

表 6．PTSDN 辨認結果(3 組辨認器)

資料分類	anchor	reporter	interviewee
字數	14694	9279	10377
字數/Turn	78.45	44.19	56.70
Del	3.3%	2.7%	7.4%
Sub	21.5%	29.1%	46.7%
Ins	0.8%	0.8%	4.4%
正確率	74.80%	67.28%	41.51%

接著我們再加上語言模式(Language Model)，所使用的詞典(lexicon)是由光華雜誌(所有文字資料庫可參考[4]之網頁)、語言學會之平衡語料庫[4]中詞頻最高之 58954 個詞加上補齊國語語音中之 411 音節未在詞典出現者後構成。並由上述 3 個文字資料庫求得 3-gram 語言模式。加上 bigram 語言模式後之音節辨認率如表 6 所示。

表 7．PTSDN 加語言模式後之辨認結果

Model	Anchor	Reporter	Interviewee
字數	14694	9250	3430
句數	187	209	48
字數/sub-turn	78.58	44.26	71.46
Del	3.3%	2.7%	12.4%
Sub	10.8%	15.2%	32.6%
Ins	0.4%	0.5%	4.5%
正確率	85.50%	81.54%	50.47%

但上述語言模式中並未包含口語語料中

特有的贅音(particle)，呼吸聲、嘆氣聲等語言現象，初步我們使用語言模式中 OOV 之 N-gram 來取代贅音(particle)，呼吸聲、嘆氣聲等語言現象之 N-gram，若再加上 NTCIR[4]文字庫則訓練語言模式之語料共七千多萬詞，初步辨識結果受訪者之音節辨識率可再提升 10%；進一步結果仍在執行中。上述結果可發現由文字資料所獲得之語言模式與口語語料之語言模式還是有相當程度的差異。

在 MCDC 語料庫辨認器之效能探討方面，首先我們檢視增加贅音(particle)，呼吸聲、嘆氣聲等語言現象等 HMM 辨認模型做辨認對辨認結果之影響。我們希望加入這些因學辨認模式後能夠降低因為這些現象所造成之音節插入型錯誤，但是也不希望將 411 音節辨認為這些音。於是我們對訓練語料作了測試，由表 7 可看出贅音與 411 音節中混淆最為嚴重，事實上一個音節是否為贅音單由聲音信號是無法判別的，這也引發另一個問題，如何訂定口語語音音認效能之好壞。在表 7 中有兩項資料並未列出，那就是插入及刪除型錯誤，口語語音辨認中插入及刪除型錯誤明顯較 read speech 為高，其主要原因在前面已提過。在辨認結果中由於 paralinguistic phenomena, particle, uncertain 模型造成之插入及刪除型錯誤可由表 7 算出分別為 3 及 5%，比起它們在語料庫所佔之比例 21%低了許多，可以看出將非 411 音節信號建立辨認模型之好處。

表 8. 加入非 411 音節模型後，411 音節與贅音、語言現象等模型間相互錯誤(between-class)表

辨認結果 / 答案	Para-linguistic	Particle	Uncertain	411	總數
Para-linguistic	71.75%	3.15%	2.4%	7.6%	10542
Particle	3.2%	70.9%	2.25%	16.05%	8114
Uncertain	1.9%	2.8%	55.6%	32.8%	3857
411	1.4%	1.5%	2.15%	85.7%	116175
Correct rate	62.95%	65.35%	51.25%	60.6%	

在 MCDC 語料庫辨認器之改善方面，首先我們先檢視『音節連並』(contraction)現象對口語語音辨識之影響。在 MCDC 語料庫中語言學家們已標示了發生『音節連並』(contraction)現象的地方；如果我們單就文字資料來統計，出現前 100 名的連並詞組以佔所有音節連並次數的三分之一，並且多為常見詞及 function word，表 7 中列出 MCDC 12 萬字中一些最常見之音節連並詞組。這些詞組在口語語音中，其音素結構均已改變，在語音辨認時絕對不是由音節模型連接或使用左右文相關辨認模型可以解決的。在計畫中，我們首先對這些常見音節連並詞組另外建立 word 模型，而非使用音節模型連接。在我們建立 90 個常見音節連並詞組之 word 模型後，可將 MCDC 語料庫之音節辨認率可提高 7%。

表 9. MCDC 中常見音節連並詞組(最常見的 18 個)

詞組	出現次數	詞組	出現次數
就是	918	現在	367
我們	796	他們	365
然後	697	什麼	362
覺得	622	其實	347
因為	605	一個	332
沒有	448	這樣	327
可是	438	比較	312
所以	419	不是	305
對對	396	那邊	275

MCDC 語料庫加入語言模型之辨認則正在執行中，含糊不清/發音錯誤的語音(uncertain/inappropriate pronunciation)將可因語言模型之加入而獲得改善。

4. 結論

本年度計畫中已整理國內現有之兩套口語語料庫(包含更正了許多標記錯誤)，並建立口語語音辨認器、評估其效能並作了初步錯誤分析。但也發現有幾個有趣的課題，也將是下一年度計畫的重點：(1) 口語語音之辨認單元之選定；(2) 口語語音之語言模式，尤其是加入贅音(particle)，呼吸聲、嘆氣聲等語言現象後之語言模式；事實在我們還應在語言模式中加入韻律(prosody)資訊，這都將是下一年度計畫中欲探討的課題。

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結果，有些部分因需要較多的計算機時間以獲得較精確之結果都在進行中。

五、參考文獻

[1] Hsin-Min Wang, Shi-Sian Cheng and Yong-Cheng Chen, "The SoVideo Mandarin Chinese News Retrieval System", *Int. Journal of Speech Technology*, Vol. 7, pp 189-202, 2004.

[2] 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。

[3] HTK (HMM Tool Kit) manual, Cambridge

University Engineering Department, http://htk.eng.cam.ac.uk

[4] 中華民國計算語言學學會，http://rocling.iis.sinica.edu.tw/ROCLING/corpus98/index_cf.htm