

本期要目

- | | |
|--|---------|
| 壹. Rocling-2005 CFP、議程 (暫訂) | 第二~三頁 |
| 貳. PACLIC-19 CFP | 第四頁 |
| 參. 語音訊號處理研討會- Language Learning Via Speech | 第五~六頁 |
| 肆. 專文-語音辨識技術在實用上的問題探討(許志興) | 第七~十四頁 |
| 伍. 專文-淺談數位內容之音訊標記(涂家章、簡世杰) | 第十五~十九頁 |
| 陸. 會費繳費通知單 | 第二十頁 |

九十五年會費開始繳交

九十四年度個人會員及學生會員會費即將於六月三十日到期，為保障各位會員之權益，敬請各位會員如期繳交會費，若您的個人資料有所異動，亦請惠予通知本會秘書處。繳費通知單及信用付費請參閱第 20 頁。

專文解答服務

為促進學術交流之目的，本會特開闢專文意見表達及問題解答之管道，凡對本期所刊登之專文，有任何意見或需進一步瞭解相關研究資料者，請將問題提供與本會秘書處 (aclclp@hp.iis.sinica.edu.tw)，本會會將讀者之意見及問題轉交作者，並於下期通訊刊登問題之解說。「詢問者」請確實留下姓名、所屬機構及身份。

ACLCLP/ISCA 雙會員會費折扣辦法

國際語音處理學會 (International Speech Communication Association, ISCA) 和本會的雙會員會費折扣辦法開始實施。同屬兩個學會的會員可申請 25% 的會費折扣，但終身會員者不能享有此一優惠 (因為 ISCA 沒有永久會員)。所有自動申請會員資格的程序 (線上註冊或繳費，在國際會議或研討會註冊時自動加繳會費等) 均不能辦理上述折扣；擬辦理上述折扣者請直接聯絡任何一個學會的秘書處：

aclclp@hp.iis.sinica.edu.tw (ACLCLP)
info@isca-speech.org (ISCA)

博碩士論文獎申請

博碩士論文將於七月一日開始申請，凡為國內大專院校博碩士班應屆畢業生，且從事計算語言學相關領域之研究者，皆可提出申請。

申請日期：七月一日~七月三十一日

名額：

1. 博士論文優等獎：一名，獎金二萬元，並頒給學生及指導教授獎狀各乙紙。
2. 博士論文佳作獎：一名，獎金一萬元，並頒給學生及指導教授獎狀各乙紙。
3. 碩士論文優等獎：一名，獎金一萬元，並頒給學生及指導教授獎狀各乙紙。
4. 碩士論文佳作獎：三名，獎金各伍仟元，並頒給學生及指導教授獎狀各乙紙。

其他申請相關之規定請參閱本會網站：

<http://www.aclclp.org.tw/doc/shipreg.htm>。

九十二年度博碩士論摘要彙編說明

本訊第十六卷第一期所刊出之「九十二年度博碩士論摘要」，讀者來函反應，論文漏列指導教授姓名。由於本資料收集之來源來自「國家圖書館博碩士論文資料網」及「各大專院校之圖書館」，因此，若資料有所誤或遺漏，乃原始資料之故，望讀者見諒。

獎助學生出席國際會議

補助會議：ACL-2005 Psoter/Demo Session

論文題目：Learning Source-Target Surface Patterns for Web-based Terminology Translation

獎助學生：吳鑑城 (清華大學資訊所博士班)

獎助金額：美金 600 元。



Conference on Computational Linguistics and Speech Processing

第十七屆自然語言與語音處理研討會

September 15-16, 2005, National Cheng Kung University, Tainan, ROC

<http://www.aclclp.org.tw/rocling2005.html>

CALL FOR PAPERS

Conference Chairs:

Jhing-Fa Wang

National Cheng Kung University

Program Committee:

Chung-Hsien Wu, Chair

National Cheng Kung University

Jen-Tzung Chien, Co-Chair

National Cheng Kung University

Wen-Hsiang Lu, Co-Chair

National Cheng Kung University

Shiaw-Shian Yu

CCL/TRI

Claire H. H. Chang

National Chengchi University

Jason S. Chang

National Tsing Hua University

Jing-Shin Chang

National Chi Nan University

Hsin-Hsi Chen

National Taiwan University

Keh-Jiann Chen

Academia Sinica

Kuang-Hua Chen

National Taiwan University

Sin-Horng Chen

National Chiao Tung University

Lee-Feng Chien

Academia Sinica

Zhao-Ming Gao

National Taiwan University

Wen-Lian Hsu

Academia Sinica

Chu-Ren Huang

Academia Sinica

Bor-Shenn Jeng

Chungwa Telecom Labs

Sur-Jin Ker

Soochow University

Lin-Shan Lee

National Taiwan University

Tyne Liang

National Chiao Tung University

Hsien-Chin Liou

National Tsing Hua University

Ren-Yuan Lyu

Chang Gung University

Chiu-yu Tseng

Academia Sinica

Shu-Chuan Tseng

Academia Sinica

Yuen-Hsien Tseng

Fu Jen Catholic University

Hsiao-Chuan Wang

National Tsing Hua University

H. Samuel Wang

National Tsing Hua University

Hsin-Min Wang

Academia Sinica

Yih-Ru Wang

National Chiao Tung University

Ming-Shing Yu

National Chung Hsing University

The 17th ROCLING Conference will be held at National Cheng Kung University, on September 15-16, 2005. Sponsored by Association for Computational Linguistics and Chinese Language Processing (ACLCLP), ROCLING is the most historied and major conference in the broad field of computational linguistics, speech processing, and related areas in Taiwan. ROCLING XVII will be hosted by Department of Computer Science and Information Engineering, National Cheng Kung University. The two-day conference will feature invited talks, papers and poster sessions and two workshops: Workshop on Intelligent Web Technologies and Student Workshop.

ROCLING XVII invites the submission of original and unpublished research papers on all areas of computational linguistics, natural language processing, and speech processing, including, but not limited to the following topic areas.

- | | |
|---|--|
| (a) cognitive/psychological linguistics | (k) query answering |
| (b) discourse/dialogue modeling | (l) semantics/pragmatics |
| (c) information extraction/text mining | (m) speech analysis/synthesis |
| (d) information retrieval | (n) speech recognition/understanding |
| (e) language understanding/generation | (o) spoken dialog systems |
| (f) lexicon/morphology | (p) spoken language processing |
| (g) machine translation/multilingual processing | (q) syntax/parsing |
| (h) name entity recognition | (r) text summarization |
| (i) NLP applications/tools/resources | (s) Web knowledge discovery |
| (j) phonetics/phonology | (t) word segmentation/part-of-speech tagging |

Paper Submission:

Prospective authors are invited to submit full papers of no more than 25 A4-sized pages in pdf or Microsoft Word format. Papers will be accepted only by electronic submission through the conference website. Prospective authors without Web access should contact the Program Committee Co-Chair (whlu@mail.ncku.edu.tw) before the submission deadline. The submitted papers should be written in either Chinese or English, and in single column, double-spaced format. The first page of the submitted paper should bear the items of paper title, author name, affiliation and email address. All these items should be properly centered on the top, with a short abstract of the paper following.

Best Paper Award:

The best paper will be selected and announced at ROCLING XVII.

Important Dates:

Preliminary paper submission due: July 8, 2005

Notification of acceptance: July 29, 2005

Final paper due: August 8, 2005

Sponsors:

Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

Department of Computer Science and Information Engineering, National Cheng Kung University

Department of Electrical Engineering, National Cheng Kung University

第十七屆自然語言與語音處理研討會 ROCLING-2005

專題演講：Statistical Language Modeling and Information Retrieval

主講人： Prof. Jian-Yun Nie

(http://www.iro.umontreal.ca/%7Eenie/index_Eng.html)

經歷：

Jian-Yun Nie is a professor in the Computer Science Department, University of Montreal, Canada. In 1999, he was awarded the "Best paper" award at the ACM-SIGIR conference for his work on cross-language IR using automatically mined parallel Web pages. Jian-Yun Nie also works on other related topics such as machine translation, Chinese processing, Web text mining, and so on.

Jian-Yun Nie's main research area is Information Retrieval, in which he has been active for almost 20 years. His research covers many of the IR problems, including theoretical modelling, IR in different languages, cross-language IR, and so on.

議程(暫訂)：

2005/09/15	
Time	Session
09:20-09:30	Opening Ceremony
09:30-10:30	Keynote Speech
10:30-11:00	Break
11:00-12:30	Speech Recognition/Enhancement
	Information Extraction
12:30-13:30	Lunch、中華民國計算語言學學會會員大會
13:30-15:00	NLP Applications
	Syntax/Semantics
15:00-15:30	Break
15:30-17:00	Panel Discussion (暫訂) 中文處理技術的產業機會與挑戰
18:00-20:00	Banquet

2005/09/16	
Time	Session
09:00-10:30	Translation & Multilingual System
	Speech Analysis/Synthesis
10:30-11:00	Break
11:00-12:30	Information Retrieval/Summarization
	Speech Recognition & Language Learning
12:30-13:30	Lunch
13:30-15:00	Question Answering
	Web Knowledge Discovery
15:00-	Closing

PACLIC 19

The 19th Pacific Asia Conference on Language, Information and Computation

December 1-3, 2005

Centre for Academic Activities, Academia Sinica, Taipei

Call for Papers

The 19th Pacific Asia Conference on Language, Information and Computation (PACLIC19) will be held in Academia Sinica on December 1-3, 2005. Following the long tradition of PACLIC conferences, PACLIC19 emphasizes the synergy of linguistic analysis and language processing – from theoretical frameworks to cognitive accounts, from information processing to language understanding, and from computational modelling to multi-lingual application. PACLIC19 will be hosted by the Institute of Linguistics, Academia Sinica, and Taiwan's academic association for computational linguistics, the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). The PACLIC19 website can be found at <http://pacific19.sinica.edu.tw>.

We invite submissions of extended abstracts of unpublished research on all aspects of theoretical and computational linguistics with special focus on language spoken in the Pacific Asia region. Possible topics include but are not limited to: (1) morphology, (2) phonology, (3) syntax, (4) semantics, (5) pragmatics, (6) discourse analysis, (7) typology, (8) corpus linguistics, (9) formal grammar theory, (10) natural language processing, (11) natural language systems and (12) related computer applications.

Abstracts should not be longer than six (6) A4 pages (about 1,500 words). They should use 11pt fonts and double line spacing throughout. Author information should appear in a separate page containing the following information: (1) title of the paper, (2) name(s) of the author(s), (3) affiliation(s), (4) postal address, (5) e-mail address for correspondence, (6) the preference for oral or poster presentation, and (7) the intention to be considered for young scholar award. Accepted papers will be published in the Conference Proceedings. The camera-ready final papers should be about 10 A4 pages with single line spacing.

The deadline for abstract submission is **July 15, 2005**. We accept only online electronic submissions. The online submission system will soon be available at the PACLIC 19 website. Hard copy submission will not be accepted.

Important dates of PACLIC 19:

Deadline for extended abstract submission	July 15, 2005
Acceptance notification to authors	Sep. 15, 2005
Deadline for camera-ready paper	Oct. 15, 2005
Early registration	Nov. 1, 2005
PACLIC 19 Conference Dates	Dec. 1-3, 2005

PACLIC19 Program Committee

Jhing-fa Wang (chair)	(National Cheng Kung University)
Tom Lai (co-chair)	(City University of Hong Kong)
Yuji Matsumoto (co-chair)	(Nara Institute of Science and Technology)
Yongkyoon No (co-chair)	(Chungnam National University)

For further information about paper submission, please contact the program committee chair Prof. Jhing-fa Wang at wangjf@csie.ncku.edu.tw.

2005 語音訊號處理研討會 - Language Learning Via Speech

會議時間：94 年 7 月 28 日(星期四)

會議地點：工業技術研究院 竹東中興院區 51 館 4F 國際會議廳

主辦單位：工業技術研究院電腦與通訊工業研究所、中華民國計算語言學學會

會議主旨：

語音訊號處理研討會是中華民國計算語言學學會一年一度的盛事，本次研討會由工研院電通所與學會共同主辦。會議以 Language Learning Via Speech 為主題，邀請國內知名學者專家發表專題演講，並就語音處理技術在數位/語言學習的新研究方向舉辦座談會。敬邀各位先進共襄盛舉，分享與交流此領域最新的研發成果，以提昇國內語音訊號產業技術水準。歡迎各界人士踴躍參加。

議程：

時間	講題	主講人
09:30-09:50	報 到	
09:50-10:00	開幕致詞	林寶樹所長(工研院電通所)
10:00-10:50	語音與語言技術於口/手語學習之研究	吳宗憲教授(成功大學)
10:50-11:10	Tea Break	
11:10-12:00	台語(閩南語)語音處理及資料庫建立技術	呂仁園教授(長庚大學)
12:00-13:30	午 餐	
13:30-14:20	語音技術與電腦輔助語言教學	張家麟博士(艾爾科技)
14:20-15:10	我唸得不夠好嗎？ —解讀語音辨識技術取代正音師腳色的迷思	尤菊芳教授(東海大學)
15:10-15:30	Tea Break	
15:30-16:30	座 談 會	李琳山教授主持 (台灣大學)

2005 語音訊號處理研討會
- Language Learning Via Speech
報名表

會議時間：2005 年 7 月 28 日（星期四）
會議地點：工業技術研究院 竹東中興院區 51 館 4F 國際會議廳
主辦單位：工業技術研究院電腦與通訊工業研究所、中華民國計算語言學學會

姓 名		單位	
電 話		E-mail	
聯絡地址			
收據抬頭			
報名費	一般人士： <input type="checkbox"/> 會員 400 元 <input type="checkbox"/> 非會員 600 元 學生： <input type="checkbox"/> 會員 200 元 <input type="checkbox"/> 非會員 300 元 (報名費包含講義、午餐及茶點)		

1. 非學生會員請附學生證影本
2. 報名截止日 **7/22** 日（現場報名加收 100 元）
3. 繳費方式：
 - 郵政劃撥：劃撥帳號 19166251、戶名：中華民國計算語言學學會
 - 信用卡：Card Type：(Visa / Master-card / JCB)

信用卡持有人：_____發卡銀行：_____

卡號：---有效期：_____

卡片後三碼：_____簽字：_____

4. 午餐請備素食
5. 報名表請傳真或郵寄至：

310 新竹縣竹東鎮中興路四段 195 號 51 館 709 室 工研院電通所
呂秀婷 小姐收 電話：03-5914646 傳真：03-5820098

語音辨識技術在實用上的問題探討

工業技術研究院

許志興

摘要

語音辨識技術近幾年來被廣泛的應用，然而真正成功的例子並不多，許多因素所造成，本文中我嚐試舉出幾個語音應用的例子，並且指出在該項應用中語音辨識技術所面臨的問題，包括實用上或是技術上的問題等，從我個人的經驗裡，把一些我曾經遇過的問題都提出來討論。

一. 簡介

語音辨識技術應用在近幾年來越來越廣泛，然而在語音辨識技術並不是很成熟的情形下，它可能被應用的領域，仍相當受限，我所謂的不成熟，指的是這項技術仍不能像手寫辨識一樣至少讓90%以上的人都覺得滿意，而且很穩定；有些人認為很好，有些人認為很差，很兩極化，我們曾經邀請80個用戶來測試我們設計的語音入口網站系統(提供聲控查詢股票、路況、氣象、星座運勢及講笑話的服務)，並請他們寫下使用後的感想，歸納結果得到下列的評語：

1. 語音服務還不錯，覺得速度還蠻快的，提供很多生活化資訊，不需要考慮到網路連線問題，可以省去很多不必要的麻煩，整體感覺還蠻不錯的，系統很先進，蠻新鮮的，很猛且超神奇的，很好很方便，想知道什麼就查。
2. 有點小糟，不是很習慣，發音要很正確它才能辨識，假如發音不標準的話就沒辦法，例如台灣國語就沒辦法辨識。
3. 蠻爛的，覺得反應不靈敏，對使用者造成嚴重的困擾，使用後不會想要再次使用，辨音系統很差，它一直聽不懂我說什麼，像我說的很標準高雄，結果它居然反應說只能查詢台灣地區。而且笑話好難笑，聽不懂它在說什麼，浪費我的時間，有股很想打人的衝動，總之覺得它很白痴，有待加強，改進的空間很大。

同樣一個系統，用戶的感覺從很好到很壞的評語都有，我認為這也是目前現有語音產品同樣面臨的挑戰，一般人聽到語音辨識，直覺地認為，我說什麼，電腦都應該聽得懂我說的話，在預期心理與實際系統表現的落差呈現出來後，很多人便對語音辨識失去了信心，甚至抗拒去使用語音辨識技術的產品。像很多公司的自動總機系統寧可保守地採用傳統按鍵的轉接方式，而不買方便使用的聲控轉接系統，我認為這些問題可從兩方面來看：

1. 語音辨識技術核心的問題

以目前的語音辨識技術而言，各廠家或學術單位所做的語音辨識器，應該都已達一定的水準，差別應在強調對使用環境的強健性(Robustness)，使用環境包括：雜訊、通道、使用者口音、性別、年齡等，使用者的特性，可藉由訓練語料的收集來提昇辨識效能解決一部份的問題，然而雜訊及通道效應，仍是目前熱門研究的課題，尤其是在手機普及後，用戶可在各種不同的環境(雜訊)下使用各式各樣的手機(裝置通道)Access到語音辨識系統，這對語音辨識器又是另一項挑戰，以後如果VoIP流行後，除了要解決不同語音編解碼器(Voice Codec)對語音辨識產生的影響外，像是封包遺失(Packet Loss)或是不同麥克風及錄音裝置等的通道問題，都是陸續應該解決的問題。

2. 使用介面的問題

影響使用者對語音辨識系統第一印象好或壞的因素是使用介面，我所指的是廣泛的介面，不同應用有其特殊的介面需求，傳統的聽寫機，操作介面很簡單祇要把辨識出的文字傳送到使用者目前正在編輯的視窗內就符合需求，其它像PDA聲控撥號應用，辨識詞庫應該能跟PDA上的連絡人資料庫做整合，這樣用戶就不用再重覆鍵入要辨識的人名，又如應用在電腦電話語音系統時，系統應該提供插話功能，讓用戶可以隨時說出需要的服務，而不必等待系統的提示句(prompt)播放結束。好的介面設計有時甚至可以彌蓋掉辨識引擎的缺點。

在語音辨識系統可用但又不是那麼好用的技術水平下，如果要實際應用在語音產品上，許多問題便需加以注意及特殊設計，以下的章節，我將以幾種語音產品為例，嚐試說明其設計時所遇到的問題。

二. 實用範例

1. 以自動總機系統為例

自動總機系統允許用戶透過電話，不必用電話按鍵，直接用語音輸入經電腦辨識後依用戶的意圖，轉接給目的受話方[1]，從語音輸入到完成轉接的流程來看，可區分出下列幾個問題：

(1). 語音插話偵測問題

首先是語音輸入時是否允許用戶以插話(Voice Barge-In)方式輸入，以利嫻熟系統的用戶快速輸入語音，若允許插話，則在實現插話功能時，有下列幾個問題要考慮

- a. 電話語音卡需選擇可做雙向錄放音的介面卡，例如:Dialogic JCT系列語音卡，可安裝其CSP(Continuous Speech Processing)驅動程式，做回音消除(Echo Cancellation)及語音偵測(Voice Activity Detection, VAD)來偵測插話發生的時間點。
- b. 通常語音卡內附的回音消除器及VAD演算法並無法滿足實際使用環境的需求,因此如果覺得內附的回音消除或插話偵測效果不好，可再自行串接一級Echo Cancellator及利用較複雜的VAD演算法來達成較好的插話偵測效果，尤其當用戶的使用條件是在較吵雜的環境時，例如:以手機當做輸入，則VAD演算法必須有抗雜訊輸入的能力，且其判斷Voice/UnVoice的臨界值設定必須夠Robust，以免誤啓動Barge-In。
- c. 此外，對於較遠端的用戶，例長途電話，若系統回應句的能量太大時，Echo能量也會相對的增加，有時也會造成Barge-In誤啓動，因此適當的調整系統回應句的音量大小也是很重要的系統參數設定，通常在實際系統中會設定一個音量倍數，將所有輸出的回應句乘上該音量倍數再交給語音卡做輸出。

(2). 多語辨識問題

另一個在很重要的功能是語音辨識可否處理多語的問題，尤其是中英文夾雜的語音辨識，因為現今很多人除了中文名字外，都會再建另一個英文別名，如果分別使用中英兩組語音模型，在做Viterbi競爭最佳路徑時，兩組模型間的模型差異值 (bias)必須一併納入競爭的考量，另外也可考量是否將中文及英文的model做適當Level的Tying，以降低模型之間的分數差異。

(3). 拒辨能力的問題

拒辨能力允許系統對非屬於辨識人名的詞外集(Out of Vocabulary, OOV)的輸入做拒絕接受其辨識結果，一般會以一個語詞驗證器(Utterance Verification Unit)來處理這個問題，對同一個語音輸入的所有辨識候選詞，語詞驗證器都進行評分，並且以一個臨界值來進行拒絕或接受的判定，因此，便產生了假性拒絕(False Reject)：輸入的是在辨識詞集內，且被辨識出來，但因分數小於可接受的臨界值而被誤判為拒絕。假性接受(False Alarm)：輸

入的是在辨識詞集外，但被辨識出來且其分數超過臨界值而被接受。臨界值的設定決定了False Reject Rate及False Alarm Rate，在自動總機系統的應用裡，通常我們會設定一個較高的臨界值，當辨識的分數超過該高臨界值時，即逕行直接轉接，而當辨識分數介於高臨界值及另一低臨界值之間時，系統可播報這些分數適當的辨識候選詞，讓用戶進行確認的動作，以免誤轉接的情形發生。

(4). 交換機的問題

實作自動總機時，不同的交換機，具有不同的通道特性，通常數位式交換機，其通道雜訊較小，而傳統類比交換機通道雜訊較大，另外不同的交換機，其偵測發話或受話方掛斷訊號的方法也不同，通常，我們會以一自動偵測交換機訊號特性的程式，把交換機的各種訊號先錄下來進行分析(FFT Analysis)，並且把分析後的訊號特性設定於語音卡的訊號分析器中，以便其在分析電話線上的訊號時可以做出正確的判斷對方是否已掛斷電話，另一種做法是將分析程式寫在總機程式內，當要設定訊號參數時，直接撥打電話進線，然後在線上直接學習訊號參數。

(5). 多線處理

通常自動總機系統不會祇提供單線服務，4至16甚至更多通道數目的線路是必須的，因此在設計系統時必須考慮是否使用多執行緒(Multi-Thread)的寫法，又若提供的通道數目更高時，語音辨識核心對系統記憶體的需求也不能太高，因此，可將同一時間可啓用的語音辨識器數目降低，而要求做語音辨識的通道則依序排隊，等待有語音辨識器可用時再進行語音辨識，當然，這會延長使用者的等待時間，所以適時安排使用者進線的時機，例如：利用響鈴的次數來控制，也可提昇使用者對系統整體效能的觀感。

(6). 其它

在自動總機系統的應用中，其它像是同名同姓的處理方式(同音詞)或是利用關鍵詞粹取(Keyword-Spotting)技術來處理綴詞問題(如:請轉xxxx先生)等，都是語音辨識器必須提供的功能。

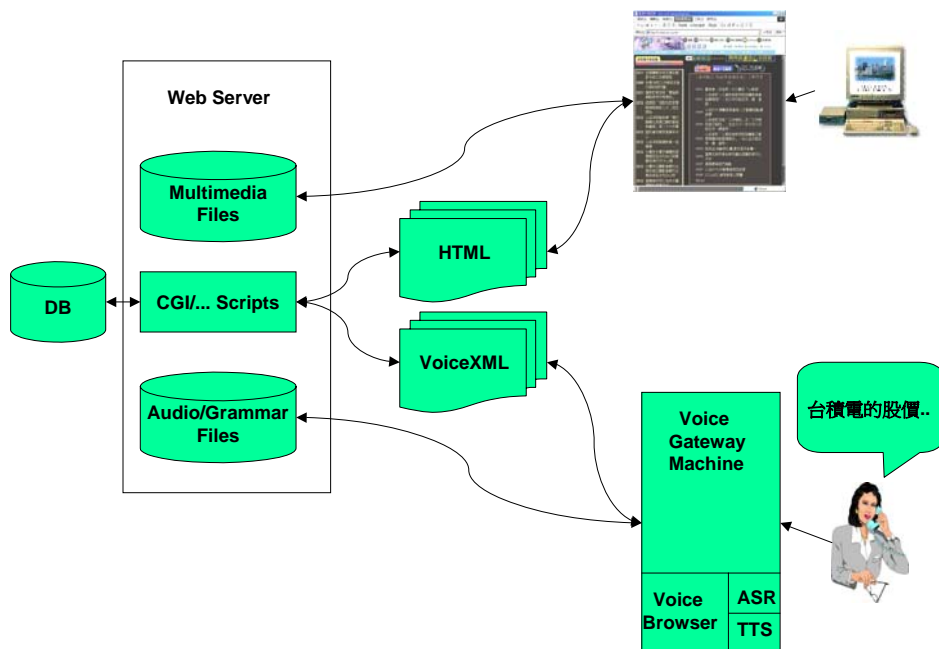
2. 以語音入口網站系統為例

語音入口網站系統(Voice Portal System)是近幾年很流行的語音應用，該系統提供用戶透過電話操作即可取得網際網路上的資訊，其中用到的核心技術包括了語音辨識及語音合成，而爲了不同語音入口網站相互間的溝通性，W3C組織制定一種稱爲VoiceXML(Voice Extended Markup Language)的標準化語言[2]，提供用戶用來發展各式各樣的語音應用系統。由於它的設計相當簡易且高階，因此幾乎是一般會寫HTML網頁的使用者即可利用它來設計符合自己需要的語音應用系統，不需學習任何的語音API函式庫，除了易學易用的特性外，VoiceXML還包括了下列特性：

- (1). 它是一種描述人和電腦之間對話流程的語言，並且以語音合成技術或播放音檔做爲輸出，以語音辨識或電話按鍵(DTMF)做爲輸入；
- (2). 主要的輸入輸出裝置是電話；
- (3). 符合目前 Internet 技術的標準：以電話替換 PC，以 VoiceXML 替換 HTML，以語音瀏覽器替換 Web Browser；
- (4). 一種標準化且具跨平台性的語言；
- (5). 一種高階化(High Level)的語言；
- (6). 針對特定領域(domain-specific)應用所設計的語言；

由於VoiceXML具有這些特性，因此在許多應用場合裡，使用VoiceXML便成了很好的選擇，例如：

- 資訊擷取：如氣象、股價查詢等；
 - 電子商務：如銀行交易、客戶帳戶、訂貨進度查詢等；
 - 電話服務：如自動總機等；
 - 個人化服務：如個人化代理人(Personal Agent)及個人化語音入口網站(Personal Portal)等。
- 我們可以把 VoiceXML 和 HTML 來做一比較，請看下圖，



HTML使用者以GUI介面利用Web Browser從一個Web Server上取得要瀏覽的資訊，包括HTML文件及聲音或影像等多媒體資料，如果要查詢資料庫則透過一些script language如CGI程式，而後將查詢結果以HTML文件型式傳回給Web Browser。VoiceXML的使用者則以電話為介面，透過撥入一台裝有VoiceXML語音瀏覽器的Voice Gateway機器，從Web上取得VoiceXML文件及文法或音檔等資料，如果需要查詢資料庫，則一樣是由CGI程式執行並將結果轉成VoiceXML文件回傳給語音瀏覽器繼續做Browsing的動作。

VoiceXML 1.0版在2000年3月被提出，同年5月即被W3C接受成為業界標準（2.0版於2001年10月發表，是目前最新的版本）。在這之前許多公司有各自的語言以發展他們自己的語音應用系統，例如：AT&T及Lucent有PML(Phone Markup Language)、IBM有Speech Markup Language、Motorola有VoxML，事實上VoiceXML就是由這些公司共同推動及制定的。

有了VoiceXML的標準，便需要可以解譯VoiceXML文件的瀏覽器，也就是我們所謂的語音瀏覽器。根據VoiceXML標準，一個語音瀏覽器至少必須具備下列模組，

- **VoiceXML 語言解譯器**

VoiceXML 語言解譯器將所取回的 VoiceXML 網頁剖析並執行。剖析網頁可用一般的 XML Parser，因為 VoiceXML 本身即是符合 XML 規格的語言；執行動作則包括執行 JavaScript 及 VoiceXML 語法中有關語音及電話功能的元素(element)，底下其餘模組即是用來支援這些元素所要執行的功能。

- **電話界面服務模組**

電話界面服務模組負責執行有關電話介面的工作，例如：等待使用者打電話進來(Wait for Call)、播放一段聲音(Audio Playback)、錄製使用者的聲音(Recording)或是接受按鍵(DTMF)輸入等，並且可執行電話轉接及偵測電話是否掛斷等工作。

- **語音辨識服務模組**

語音辨識服務模組是語音辨識核心，負責將使用者的聲音輸入辨識成文字輸出。由於 VoiceXML 規範的語音對話(Voice Dialog)是一特定領域內容，因此語音辨識的內容僅限於 VoiceXML 文件中所設定的語音文法(Speech Grammar)的範圍內。

- **文字轉語音服務模組**

文字轉語音服務模組負責執行將文字轉換成語音的工作，為了合成更自然的語音，VoiceXML 也制定一種標準的 Markup Language 用來標示想要的聲音特性，例如調整音調高低、音量大小及速度快慢等。

- **HTTP 服務模組**

HTTP 服務模組執行和目前 HTML browser 類似的工作，例如：Submit 一個 Request 到 HTTP Server，執行 CGI 程式以連結資料庫及交換資料。

- **VoiceXML 範例**

VoiceXML 網頁就如同 HTML 網頁一樣，可以存放在任何的 Web Server 上，其語法符合 XML 語言的規格，但在 VoiceXML 中定義了許多特有的元素(Element)來使語音功能能輕易的寫在 VoiceXML 文件中，以下是一個簡單的 VoiceXML 文件範例：

```
<?xml version="1.0"?>
<vxml version="2.0" encoding="iso-8859-1">
  <form>
    <field name="breakfast">
      <prompt>你要點什麼早餐</prompt>
      <grammar>土司 | 熱狗 | 漢堡 </grammar>
      <filled>
        <submit next=http://www.food.com/breakfast.cgi
          namelist="breakfast" />
      </filled>
    </field>
  </form>
</vxml>
```

上述範例是一個簡單的語音點餐應用，以下是使用者與電腦間可能的對話：

C：你要點什麼早餐

U：熱狗

其實你可以想像和電腦間的對話就像在填表格一樣，一個表格有許多欄位(field)，每個欄位可以用聲音(voice)或按鍵(DTMF)輸入，而且可以有個別的文法來限制使用者的輸入範圍。

除了表格形式的對話外，VoiceXML 還提供了較簡單的<menu>形式的對話，以下是一個使用<menu>對話的範例：

```
<?xml version="1.0" encoding="ISO8859-1" ?>
<vxml version="1.0">
  <menu>
    <prompt> 請選擇你要的服務項目<enumerate/>
    </prompt>
    <choice next="sports.vxml"> 運動 </choice>
    <choice next="weather.vxml"> 天氣 </choice>
    <choice next="news.vxml"> 新聞 </choice>
  </menu>
```

</vxml>

上例中當使用者的輸入符合某一選項(<choice>)時，語音瀏覽器即載入該選項對應的VoiceXML網頁並執行之。撰寫VoiceXML網頁時可根據使用者能輸入範圍的複雜度來決定使用<form>或<menu>建構VoiceXML網頁。

此外，爲了讓對話更具彈性，在VoiceXML中允許電腦主導式(Computer Directed)及互動主導式(Mixed-directed)兩種對話。電腦主導式的對話流程依VoiceXML網頁的設計流程固定進行；而互動主導式的對話流程可因使用者的輸入而改變要使用的對話流程，而且允許同一句話填滿兩個以上的欄位，例如：

<例>

C：請問你要查詢什麼天氣

U：降雨量

C：請問你要查什麼地方的降雨量

U：台北市

<例>

C：請問你要查詢什麼天氣

U：台北市

C：請問你要查台北市的什麼天氣

U：降雨量

你會發現電腦的對話流程會根據使用者的輸入而改變，此即所謂的互動式對話。

基於上述VoiceXML的特性，若要設計一個可瀏覽VoiceXML的語音瀏覽器(Voice Browser)，語音辨識核心至少必須具備下列能力：

- a. VoiceXML的Grammar Parser
- b. 可動態建立及載入辨識詞庫
- c. 具有辨識多個關鍵詞(Multiple Keyword)的能力

而VoiceXML解譯器則有開放的源碼OpenVXI可供下載免費使用[3].

3. 嵌入式語音辨識技術

2004是智慧型手機（Smart Phone）蓬勃發展的一年，從早期以記事功能爲主的PDA發展到目前各式功能都有的智慧型手機，這類產品體積越做越小功能卻越做越強，幾乎各種多媒體應用都被希望能放入這項產品中，使得此類消費產品更具市場吸引力；語音辨識便是其中一項，在小型化的產品上提供最直接及方便的輸入方式，例如：聲控自動撥號。然而在有限資源的嵌入式硬體及軟體OS平台上，CPU執行速度慢，記憶體小，許多以往在PC平台上開發的技術及演算法都必須做大幅修正才能放進這類平台內，例如：做浮點轉定點化(Fixed-Point)以提昇辨識速度，減少記憶體使用量以符合硬體規格，同時還需確保辨識效能不因此而變差。因此在從PC Porting到嵌入式裝置時有下列問題必須考量：

(1). 嵌入式OS平台的問題:

目前常見的OS平台有下列幾種

- Symbian: 常使用於手機中
- WinCE: 常使用於PDA上或手機
- Embedded Linux: 使用於手機中

不同的作業系統平台有不同的程序來撰寫語音辨識核心，以WinCE來說，常用的發展軟體是Embedded-Visual C++或Embedd-Visual Basic，而Embedded Linux大部份以ANSI C來發展即可，Symbian由於是RealTime OS撰寫方式大部份是以Object-Oriented的觀念進行，因此要

Porting的程序較為複雜。

(2). 語音模型定點化

不像PC，在嵌入式系統中，為了降低成本，CPU並不支援浮點數運數，如果還是以浮點數進行語音辨識，執行速度將會非常慢，因此，浮點轉定點化是必須的步驟，首先是模型的定點化，如果以HMM模型為例，必須將所有模型的Mean, Variance, Determine等參數做定點化，定點化的方法可將不同的參數先進行統計，然後取適當的Q值(2^Q 倍)，儘量讓Overflow及Underflow的參數個數降到最低，通常會將Mean, Variance, Determine分別取一適當Q值，然後將模型中的所有參數都乘上 2^Q 倍成爲一定點整數值。

(3). 求取特徵參數及比對搜尋演算法的定點化

另一必須的定點化步驟是求取特徵參數及比對演算法，進行時必須針對每一個函式或運算式的結果進行統計，儘量讓量化後的值不Overflow或Underflow，並且確保最後得到的參數的Q值，和定點化模型的Mean的Q值是一致的，如果不一致的話，可在進行前先將參數做移位(SHIFT)使其和模型的Q值一致。

(4). 啓動VFR否

爲加快辨識速度，可以使用VFR(Variable Frame Rate)來減少輸入的語音特徵參數數量，然而，大部份的靜音或母音的特徵參數經由VFR的臨界值刪除後，可大量的被減少，然而必須注意，VFR臨界值的設定必須確保，辨識率不要下降太多，常用的VFR臨界值來源有Delta-Cepstrum的Norm Sum等。

(5). 使用Frame同步的演算法

Frame同步語音辨識可在錄音同時就進行Viterbi Search，因此可減少辨識時間，然而常用語音參數CMN(Cepstrum Mean Normalization)是必須等整段語音輸入後才可求得Mean值，因此，CMN可以用Delayed-CMN延遲一小段時間後，如0.5秒，再開始求取Cepstrum Mean，然後接下來的所有參數再以Accumulated Cepstrum Mean的方式減去該Mean值或是簡單的以前一句的Cepstrum Mean值來做爲本句的Cepstrum Mean減除的依據。

(6). 在吵雜的環境下VAD及語音辨識核心

手機或PDA等裝置由於可攜性高，使用的環境多樣化，綜言之，在吵雜環境下使用語音辨識是常發生的，因此，除了語音辨識核心必須具備抗噪能力外，在語音端點(End-Point)偵測時也必須具抗噪能力，若在更吵雜的環境下，則可建議使用Push To Talk的方式進行，以免收到非語音段的背景音使辨識效果降低。

4. 語言學習

語音辨識技術在語言學習中也是近幾年來很流行的應用，其功能包括：語音跟讀，發音評量等，

(1). 語音跟讀的問題

語音跟讀指的是使用者唸出指定的文字，而跟讀游標跟著一起移動到使用者唸出的文字內容，這包括了動態文字對齊(Dynamic Alignment)技術及動態載入語音辨識詞庫的技術，當使用者唸出文字時，必須每隔一段時間去比對使用者到目前爲止唸出的最有可能的文字內容，這意謂著語音辨識器必須具有同步(Frame-Synchronized)語音辨識的能力，當使用者停頓時或是一段固定的時間後，便產生一最有可能的辨識結果，並且載入下一句或下一段要唸的文字詞庫內容。

(2). 發音評量

發音評量是針對使用者唸的內容與已知的文字內容做比對，並且以Utterance Verification技術針對每一個單字，甚至每一個Phoneme或Syllable進行評分的動作[4]，評分時，由於

對象的不同，可用不同的模型，以給定適當的分數，例如：

模型選擇的因素有下列幾種：

- a. 年齡:如小孩或成人
- b. 性別:男性或女性
- c. 口音:本土腔調或外國腔調模型

(3). 麥克風問題

在實作這類的系統時，有一點很重要的是麥克風的選擇及錄音裝置調整等問題，通常我們會選擇具指向性的麥克風以使雜訊不易進入輸入的語音中，而錄音裝置若具有AGC Control，通常會建議將AGC關閉以免雜訊也跟著放大潰入輸入的語音訊號中，若錄音裝置有DC偏移值，也必須在調整錄音裝置時量測出該DC值，以供線上求取語音特徵參數時先扣除。

三. 結論

在本文中，我舉出幾個目前流行的語音辨識技術應用的實例，不同的應用要克服的問題，也大都點了出來，個人認為，語音辨識技術要真正達到實用化，除了不斷的提高語音辨識率外，在不同的環境下進行不同壓力測試也是必須的，例如：以自動總機為例，可用電腦自動撥號給系統，並以播放音檔而模擬人聲輸入的方式，進行系統的測試工作，由於這些工作可以很大量且長時無間斷的測試，系統上線後的穩定性便可以從這些壓力測試得到保證。

語音辨識技術發展這麼久，實際成功應用的例子並不算多，環境因素是首要克服的問題，包括使用者的口音，性別，年齡等都是，如果這些問題能夠解決，那麼語音辨識技術將是人機界面最自然的使用方式，現在很多論文都一直強調Robustness的問題也就在此。

參考文獻

- [1] 游山銳、簡世杰、許志興、陳科旭、涂家章、林政賢、張森嘉，「中英文混雜關鍵詞萃取技術」，電通月刊第 108 期。
- [2] VoiceXML Forum, “Voice Extensible Markup Language VoiceXML 1.0 Spec.”, www.voicexml.org, March, 2000.
- [3] SpeechWorks, Inc. “Release Notes for VoiceXML Interpreter (VXI)- Release 1.4,” <http://www.speech.cs.cmu.edu/openvxi>.
- [4] Sukkar R. A., Setlur A. R., Lee C.-H., Jacob J., “Verifying and Correcting Recognition String Hypotheses Using Discriminative Utterance Verification”, *Speech Communication* 22 (1997), pp. 333-342.

附註：對於本文若有任何意見或相關資訊擬進一步瞭解者，請將意見及問題轉至學會秘書處，問題之回覆將於下期通訊刊登。

淺談數位內容之音訊標記

工業技術研究院
涂家章、簡世杰

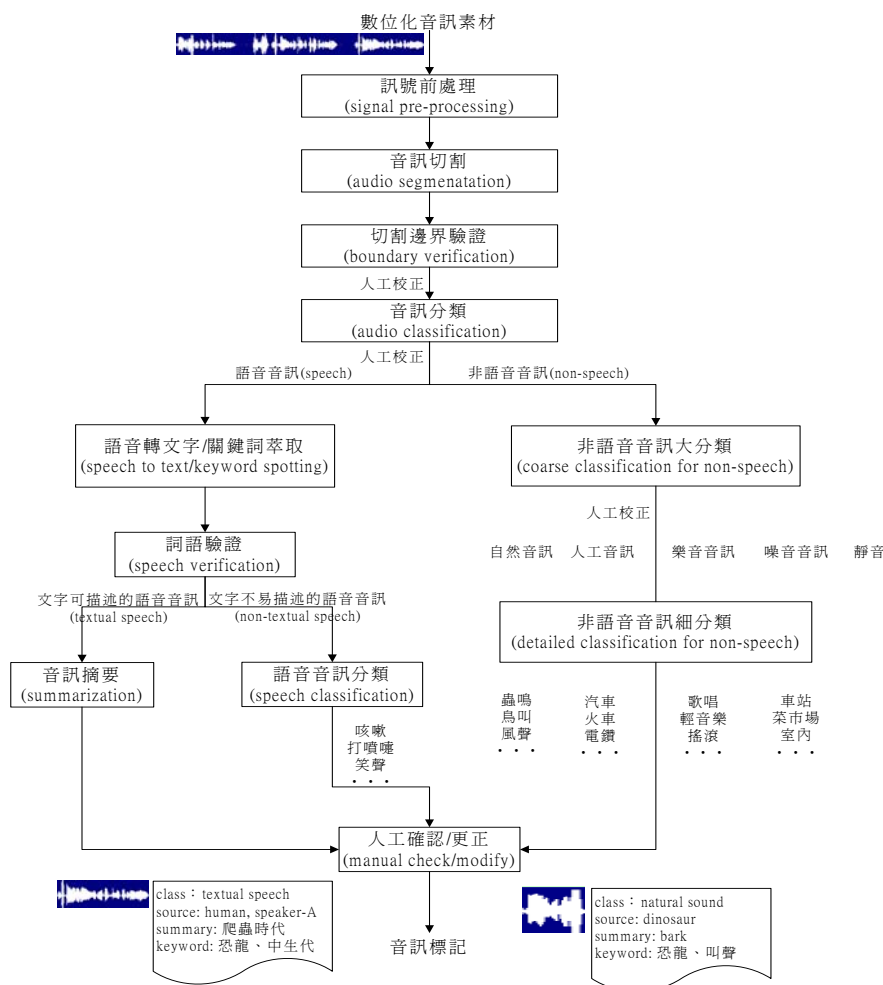
一. 前言

數位內容是數位學習成功的關鍵，如果沒有很好的數位內容，純粹將傳統的教學帶拷貝到電腦裡，以滑鼠按鍵取代翻書，是很難引發學習者的興趣，也很難讓數位學習帶來真正的學習成效。因此，如何活化數位內容以觸發學習興趣，並增加學習效果，將是推動數位學習所必須關注的焦點；譬如，將過去枯燥無味的教學帶分割成多個片段單元，並與相關的影像、動畫及文字搭配，以呈現出一個生動活潑且具有關聯性/啟發性意義的數位學習內容，才能夠對學習者的興趣及學習效果上有所助益。然而，在活化數位內容上，目前所遭遇的普遍困難點是數位內容的再用性低、分享共用性差、製作耗時費力及製作成本高等，而這也是目前數位內容產業的推動瓶頸所在。因此，如何提昇數位內容的可使用性、可再用性、可互通性、耐久性進而縮短開發時程、降低開發成本，以順利推展數位內容產業，不但已列入國家型科技方案裡，同時，數位內容也早已成爲目前行政院所推動的「兩兆雙星計畫」裡的明星之一。而同樣的，爲了因應上述的問題及思考其解決對策，目前國際上也有幾項標準正在積極推行，其中，SCORM (Sharable Content Object Reference Model) 也就是根據這個理念所制定的一項標準。這裡面包括素材 (Raw data)、可分享共用的物件 (Sharable Content Object, SCO) 以及這些分享共用物件的聚合體 (Aggregation；或稱之爲節目、課程)。而其中的SCO則是由素材所拆解成可再利用的 Meta-data 及其相對應的執行程式組合而成，也是這項標準所希望達到分享共用的精神所在。然而不管是要對過去龐大的媒體資料進行再利用，或者在這個 SCORM 架構裡將 Raw data 轉化爲 Meta-data，可以預期的，資料的整理工作將會是一個相當大的負擔，在人力和時間上的耗費將是相當可觀的。也因此可以預期，一個自動化或者半自動化的數位內容整理技術將會在未來數位內容產業裡扮演的一個相當重要的關鍵性角色。

而在數位內容的處理中，音訊標記是一項頗重要之技術，如果能夠準確的標記出音訊的種類與內涵，將來要製作數位學習教材時，就能夠快速的檢索出相關音訊，例如要製作恐龍相關的教材時，只要輸入恐龍的關鍵字，就能夠檢索出與恐龍相關的音訊資料。因此，要能夠將各種類型的音訊資料針對其特質及內涵進行標註以製作成學習元件，就必須仰賴音訊標記技術，本文將以語音處理的角度來探討音訊標記。我們以過去在語音處理方面的經驗並且參考現有的一些技術，將音訊標記進行的程序以圖一表示，以下，將根據這個程序依序探討其所牽涉到的技術和該項技術的技術狀況與所面臨的問題，並且探討一些可能之處理策略。

二. 訊號前處理

數位化音訊素材的格式可能非常多樣，爲降低系統處理的複雜度，音訊格式的一致性處理是一項必要的前處理工作。這裡的前處理工作主要是將不同格式的音訊資料轉成統一的分析格式，並且求出其特徵參數供後續處理使用。這項工作的問題點就在於所提供的數位化音訊素材格式是否低於分析格式的解析度上，譬如使用的分析格式爲 16 bits/sample，而音訊素材格式爲 8 bits/sample，雖然可以將音訊素材以等比例方式放大，但其解析度仍然低於分析格式，因此可能會得到令人無法滿意的處理結果。所以在處理前，輸入的數位化音訊素材的音訊解析度高於分析格式的解析度是必須事先加以規定的。



圖一、音訊標示程序

三. 音訊切割

音訊切割又可稱為音訊邊界偵測 (boundary detection)，其重點在偵測不同性質音訊的邊界轉換點，以將具有不同性質的音訊片段分離出來，譬如不同語者之間的轉換、語音音訊與非語音音訊和樂音音訊的邊界偵測等等。目前無論是國內或國外在音訊切割技術上的研究都以提高邊界偵測的準確率為研究重點，但仍無法完全取代人工作業[1] [2]。因此，這項技術從實際的應用面來看要達到全自動是有困難的，僅能以半自動的輔助方式來進行，也就是必須配合人工校正的型式來進行處理。由於邊界錯誤偵測 (False Alarm) 可由後續人工校正加以處理，邊界偵測失誤 (Miss Detection) 則是無法以後續的作業加以彌補的，因此邊界偵測失誤率 (Miss Detection Rate, MDR) 的降低是這項處理的技術重點；決定邊界偵測失誤率和邊界錯誤偵測率 (False Alarm Rate, FAR) 的操作點和藉著不同分析條件的使用來進低邊界偵測失誤率。

四. 切割邊界驗證

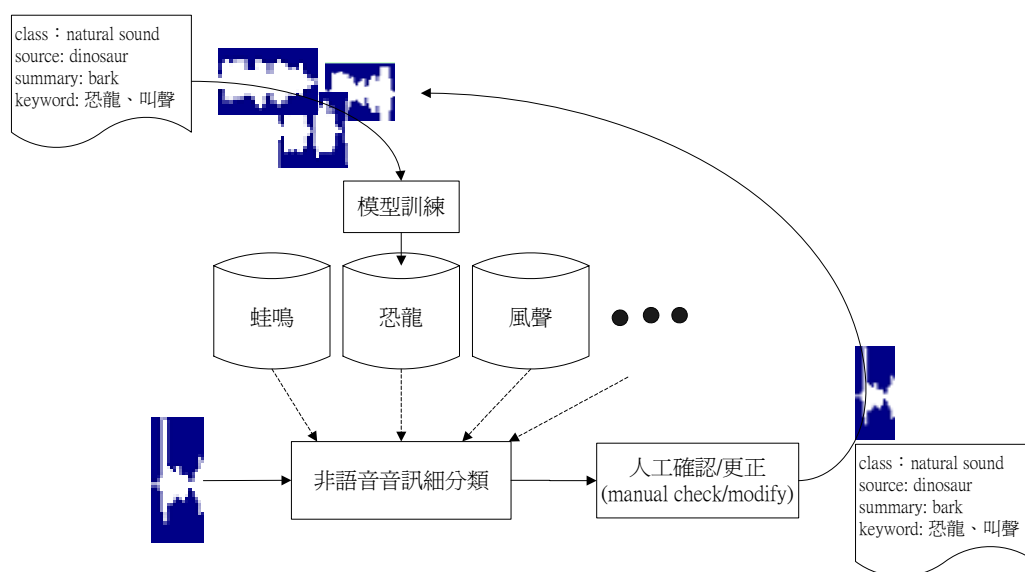
由於在音訊切割部分，我們以降低MDR方式來盡量避免邊界資訊的遺漏，然而相對的FAR也就會因此而提高，人工校正動作也因此而變的相當繁瑣。這裡切割邊界驗證的重點就是在為每一音訊邊界提供一個信心值來輔助人工判斷，以降低人工作業負擔。雖然這項技術目前沒有與之直接相關的文獻出現，不過以過去在語音驗證 (speech verification) 及語者驗證 (speaker verification) 技術上的基礎，發展邊界驗證技術是極為可能的。未來當這項技術成熟時，也可能與音訊切割部分合併，以提供更為精準的音訊切割結果。不過以現階段可能達到的技術水準來看，這部分的輸出結果仍必須仰賴部分的人工作業來進行最後音訊切割結果的檢驗與校正。

五. 音訊分類、非語音音訊大分類及非語音音訊細分類

在人耳可鑑別的音訊中，我們將它分為語音音訊與非語音音訊兩部分，而這也是目前從事語音研究裡所常使用的分類方式。而在非語音音訊部分，我們依照音訊的特質又再以大分類及細分類等程序將音訊依其發聲源和屬性加以細分，以進行非語音音訊部分的標記。

由於語音音訊含有人類可認知的語言資訊在，也是目前在人機介面技術中較被關注的焦點並也已發展出許多較為成熟的解析技術，因此這部分採用的程序與非語音音訊部分會有明顯的技術區隔。相較於語音音訊，非語音音訊部分顯然的就沒有被充分的瞭解也沒有特別為這個部分發展出較為具體的解析工具與技術，以目前的技術來看，通常只使用一般通用的分類方法來對具有不同特質的音訊加以分類，譬如類神經網路（Neural Network，NN）、高斯混合模型（Gaussian Mixture Model，GMM）或是支持向量機（Support Vector Machine，SVM）等等。但這些分類方法通常都必須要有相當的音訊樣本才能訓練出分類模型，並且不同類別的音訊之間的特性必須要有明顯的區隔性才能得到其分類的準確性。因此，在面對自然界裡不勝枚舉的音訊，以及在音訊特質無法被完全瞭解的情況下，要準確的將各種音訊分離出來幾乎是難以達到的，僅能以目前有限的音訊資料為基礎，以漸進的方式逐步進行各類音訊資料的收集和建立其分類模型來完成。譬如以圖二為例，在非語音音訊的細分類裡，我們希望能夠將屬於自然音訊的部分加以細分；判斷輸入的音訊是蛙鳴、風聲或是恐龍的叫聲等等。當然，在實際的狀況下我們不可能真的收集到恐龍的叫聲，它只是人類根據史料的記載和過去所出土的恐龍化石的生理結構進行猜測所製作出來的虛擬音訊，而在有限的時間下，我們也不可能為各種類別的音訊進行廣泛的收集，因此從這樣的環境限制下，就處理的音訊素材來進行樣本的收集，並在收集足夠的樣本後來建立其分類模型才是較為可行的模式。所以對於非語音音訊的部分，初期應該將先以人工方式來進行分類，並以逐步建立其音訊資料類別和其分類模型的方式來完成。

對於語音音訊與非語音音訊兩部分的分類，目前大部分的研究也僅著重在如何區分語音音訊與靜音、噪音甚至樂音音訊之間的判斷，對於自然音訊和人工音訊則甚少被提及，因此，這個部分雖然有一些可資參考技術，但其分類效果仍是有待商榷的，因此仍免不了人工校正的動作。而對於後續的非語音音訊大分類及非語音音訊細分類等程序，由於前述在音訊樣本的缺乏下，僅能以前述漸進式的方式來進行。至於這些部分所使用的分類技術，除了方法上可採用前述的幾種分類方法之外，其所面臨的挑戰就在於找出能夠有效的分出不同音訊類別的特徵參數上。



圖二、漸進式音訊類別模型建立

六. 語音轉文字與關鍵詞萃取

音訊素材在判斷出是屬於語音音訊之後，接下來通常就希望能得到該音訊與其所對應的文字內容，以期得到描述該音訊摘要和關鍵詞。目前對於語音轉文字的部分已有許多被認為是成熟的聽寫機商品，然而其中卻有許多的技術盲點存在，譬如必須要限定語音的類別是新聞類或科技類等等，以選擇其適當的語言模型（language model），進而縮小其資料搜尋的範圍；必須要有語音調適機制（adaptative training）以調整其聲學模型（acoustic model）才能達到“可用”的程度。當然在音訊素材處理時，我們或許可以預先知道素材的範圍，進而限定其使用的語言模型，然而當素材範圍超出這些語言模型可描述的範圍時，我們可能又必須再為該範圍收集足夠的語言資料來訓練語言模型，但這對整個音訊標記工作而言，無疑的是相當大的負擔。而對於語音調適的部分，它必須在事前收集足以代表該語者的語音資料才能進行語者調適（speaker adaptation）動作，但對於音訊素材已固定的情形下，這點是難以辦到的。因此，對於語音轉文字部分的另外一個選擇是使用關鍵詞萃取技術。

關鍵詞萃取技術旨在抽取音訊素材裡與處理範圍一致的關鍵詞出來，因此它只要有關鍵詞的定義即可達到。而在範圍限定下，這項技術所能達到的準確率是相當高的，譬如工研院的自動總機系統使用關鍵詞萃取技術，關鍵詞為七千多位左右的員工姓名，以語者無關（speaker independent）的聲學模型進行辨識，辨識率已可達到90%以上的水準[3]。此外，透過關鍵詞萃取，我們也可直接取得音訊標記所要標註的關鍵詞欄位，也不用再為這個部分額外進行處理。因此，這個部分以關鍵詞萃取技術並配合處理範圍的關鍵詞定義，應當是相當可行的。

另外，語音音訊有時可能會有相對應的文字資料可提供參照，在這種情況下，我們是可以得到語音音訊與文字資料的對應關係的（使用Viterbi algorithm即可），然而，對應的文字資料有時也可能會與語音音訊有不一致的情形，這時候這些文字資料就不一定可以完全參考，也可能會產生負面的干擾，將這些文字資料拆解成關鍵詞再套用到關鍵詞萃取上可能是較為可行的方式。

七. 詞語驗證

由於語音轉文字或是關鍵詞萃取技術都不可能得到百分之百的正確率，因此詞語的驗證是相當重要的一環。這裡除了在確認文字與語音的一致性程度之外，也將藉著這個確認機制將文字不易描述的語音音訊部分（如：咳嗽、打噴嚏及笑聲等等）分離出來。目前這部分已有許多可參考的技術可運用，如使用驗證模型（verification model）與反模型（anti-model）的相似率（likelihood ratio）來進行驗證[4]或使用類神經網路的分類機制進行判斷[5]等，在技術上都是具有相當高的技術成熟度。

八. 音訊摘要

音訊摘要是整個音訊標記程序裡最困難的一項技術。它除了要配合許多事前的資料收集和分析，在應用領域上也必須事先加以限制，並且這項技術目前所能達到的效果也是相當有限的[6,7]，因此是一項無法在短期可以完成的技術項目。因此，以現階段的技術而言，這個部分多半只能仰賴人工作業，並以前述的關鍵詞萃取後所得到的關鍵詞出現頻率來加以輔助。

九. 語音音訊分類

這部分主要是針對咳嗽、打噴嚏、笑聲等等文字不易描述的語音音訊以其相對應的事件或行為來加以描述，而要為這些事件進行判別，其狀況是與前述的音訊分類和非語音音訊分類相同的；必須有相當足夠的音訊樣本來建立其參考模型，因此所採取的可能策略與音訊分類和非語音音訊分類是相同的。

十. 人工確認與更正

在完成音訊標記之前，除了現階段某些技術不成熟必須以人工方式介入之外，不論標記的正確與否，確定標記的正確性仍是一項必要且重要的步驟。因此，我們將它列為最後完成音訊標記所必經的最後一道程序。

十一. 結論

綜合以上的探討，音訊標記牽涉到的技術不少，處處充滿挑戰，以目前的語音處理技術而言，要將音訊標記做到百分之百是不容易的，不過要達到半自動的程度是有機會的，一旦能達到半自動化的程度，對於數位內容的處理將有相當的助益，也會加速數位學習產業之推動，因此，我們將來會對前述所提的相關技術進行加強與突破，希望能達成半自動化的音訊標記，將來更希望能達到全自動的目標！

計畫相關資訊

本文係工研院電通所執行經濟部九十四年度前瞻研究專案 A341XS1010 計劃成果之一

參考文獻

- [1] S. S. Chen and P. S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," Proc. Of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.
- [2] S.-S. Cheng and H.-M. Wang, "A Sequential Metric-based Audio Segmentation Method via The Bayesian Information Criterion," Eurospeech'03, 2003.
- [3] 謝偉強、簡世杰、許志興、張森嘉，"工研院 104 自動總機系統的改進過程"，電腦與通訊月刊第 96 期。
- [4] Sukkar, R.A. and C.-H. Lee. 1996. Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition. *IEEE Trans.On Speech and Audio Proc.*, 4(6): 420-429.
- [5] S.-C. Chien, T.-H. Hwang, and S.-C. Chang, "MLP-based Utterance Verification for In-Car Speech Recognition," O'COCOSDA2003, 2003.
- [6] T. Kikuchi, S. Furui, and C. Hori, "Automatic Speech Summarization based on Sentence Extraction and Compaction," ICASSP'03, 2003.
- [7] T. Hori, C. Hori, and Y. Minami, "Speech Summarization using Weighted Finite-State Transducers," Eurospeech'03, 2003.

附註：對於本文若有任何意見或相關資訊擬進一步瞭解者，請將意見及問題轉至學會秘書處，問題之回覆將於下期通訊刊登。

