中華民國計算語言學學會
通訊 ACLCLP
發行人：陳信宏
主　編：曾淑娟
執行編輯：黃　琪
The Association for Computational Linguistics and Chinese Language Processing
第十六卷第五期
E-Mail:aclclp@hp.iis.sinica.edu.tw 網 址：http://www.aclclp.org.tw
地 址：台北市研究院路二段128號中研院資訊所　電 話：(02)2788-3799 ext:1502
傳 眞：(02)2788-1638
劃 撥：19166251

本期要目

PACLIC-19

「第十九亞太地區語言、訊息暨計算國際研討會」謹訂於 94/12/1～94/12/3 假中研院活動中心國際會議廳舉行，日前已開放線上報名，報名截止日 **94/11/1** 日，暫訂議程請參閱第二～五頁。

第一屆 NTCIR 中文問答系統比賽

中研院資訊所許聞廉老師主持的「智慧型代理人實驗室」參加 NTCIR 跨語言問答系統比賽(CLQA)，於中文問答系統競賽獲得最佳成績。NTCIR 爲日本 NII 所主辦的國際會議，與會者多爲亞洲資訊處理相關學者，會議目的在促進相關領域技術交流。第一屆會議開始於 1999 年，每年都舉辦多項資訊處理比賽，今年的比賽項目包括跨語言文件檢索、專利文件檢索、網路文件檢索、日文問答、跨語言問答等五項競賽。其中日文問答比賽已經是第三屆，上屆（第二屆）第一名的團隊爲日本電信電話公司(NTT)，正確率爲 51.3%。今年，NTCIR 首次舉辦中文問答系統比賽。由於這是第一次嘗試提供此類比賽，因此，本次問題重點在人、地、物等較爲簡單明確之答案。智慧型代理人實驗室累積多年問答系統開發經驗，於本次競賽獲得 44.5%正確率，非常接近日文系統的五成正確率。相較於其他主要語言的問答系統（英文問答比賽舉辦多年，正確率已達70%以上，法、葡語也有 65%），中文問答系統的正確率仍然略低。其中一個原因爲，中文由於缺乏詞界的標示，「未知詞」詞界與屬性分析相當困難。該實驗室表示，此次由於初次參加，僅做了未知詞分析、問題分類、關鍵詞加權的部分，還有法則與模版的運用尚未加入，會持續改進該系統，提升答題的正確率。

第九屆理監事當選名單

理事長：陳信宏(交通大學電信工程系教授)
副理事：陳信希(台灣大學資訊工程學系教授)
理事：鄭伯順(中華電信研究所副所長)
陳光華(台灣大學圖書資訊系副教授)
吳宗憲(成功大學資訊工程系教授)
簡仁宗(成功大學資訊工程系副教授)
張景新(暨南國際大學資工系助理教授)
高照明(台灣大學外文系助理教授)
王新民(中研院資訊所副研究員)
簡立峰(中研院資訊所研究員兼副所長)
許聞廉 (中研院資訊所研究員)
候補理事：余明興 (中興大學資訊科學系副教授)
劉美君 (交通大學外文系教授)
監事主席：陳克健 (中研院資訊所研究員)
監事：黃居仁 (中研院語言所研究員)
張俊盛 (清華大學資訊系教授)
王小川 (清華大學電機系教授)
蘇克毅 (致遠科技有限公司總經理)
候補監事：王駿發 (成功大學電機工程系教授)
海外理事：
鄒嘉彥 (香港城市大學教授及語言資訊研究所中心主任)
李錦輝 (Prof. Georgia Institute of Technology, USA.)
舒維都 (Prof. Electrical Engineering and Computer Science at MIT)

附註：以上順序乃依據選票上之編號，不代表任何意義。

PACLIC 19

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

December 1-3, 2005
Centre for Academic Activities, Academia Sinica, Taipei
Organized by
Institute of Linguistics, Academia Sinica
Association for Computational Linguistics and Chinese Language Processing

The 19th Pacific Asia Conference on Language, Information and Computation (PACLIC 19) will be held in Academia Sinica on December 1-3, 2005. Following the long tradition of PACLIC conferences, PACLIC 19 emphasizes the integration of language processing – from linguistic understanding through information parsing to computational calculation and modelling. This year PACLIC 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). Research results of theoretical and empirical analysis of languages, automatic processing of linguistic information and their computational modelling will be presented at PACLIC 19. The PACLIC 19 online registration is now open. To be qualified for the discount of early registration, you are required to register before **November 1, 2005**. Please visit our website at http://paclic19.sinica.edu.tw. The programme of PACLIC19 is shown in following pages.

Tentative Programme

2005/12/1(Thursday)

Slot	Paper Title	Author and Affiliation
09:30-09:50	Opening Ceremony (Chair: Jhing-fa Wang)	
09:50-10:40	Keynote Speech: Steven Bird (University of Melbourne) Querying Linguistic Databases Chair: Chung- Hsien Wu	
10:40-11:00	**Coffee Break**	
Oral Presentation--Session chair:		
11:00-11:25	A Two-Level Morphology of Malagasy (21)	Mary Dalrymple, Oxford University Maria Liakata, Oxford University Lisa Mackie, Oxford University
11:25-11:50	A Framework for Data Management for the Online Volunteer Translators' Aid System QRLex (40)	Youcef Bey, Université Joseph Fourier Kyo Kageura, University of Tokyo Christian Boitet, Université Joseph Fourier
11:50-12:15	People in the State of the Union: Viewing Social Change through the Eyes of Presidents (45)	Kathleen Ahrens, National Taiwan University
12:15-14:00	**Lunch**	
Oral Presentation--Session chair:		
14:00-14:25	MARKET Metaphors: Comparison of Chinese, English and Malay (4)	Siaw-Fong Chung , National Taiwan University
14:25-14:50	Constructing Filler-Gap Dependencies in Chinese Possessor Relative Clauses (29)	Chien-Jer Charles Lin, University of Arizona Sandiway Fong, University of Arizona Thomas G. Bever, University of Arizona

14:50-15:15	In and Out: Senses and Meaning Extension of Mandarin Spatial Terms *nei* and *wai* (49)	Yiching Wu, National Tsing Hua University Cui-xia Weng, Institute of Linguistics, Academia Sinica Chu-Ren Huang, Institute of Linguistics, Academia Sinica
15:15-15:35	**Coffee Break**	
Poster Presentation (with 5 min oral presentation)-- Session chair:		
15:35-15:40	Multiply Quantified Internally Headed Relative Clause in Japanese: A Skolem Term Based Approach (10)	Rui Otake, Tohoku University Kei Yoshimoto, Tohoku University
15:40-15:45	A Study on Multiple Meanings of Frequency Adverbs in Japanese (41)	Tomoaki Ozawa, Tohoku University Hiroyuki Nishina, Saitama University Kei Yoshimoto, Tohoku University Shigeru Sato, Tohoku University
15:45-15:50	A Study on Implementation of Southern-Min Taiwanese Tone Sandhi System (54)	Iu^{n} Un-gian, National Taiwan University Lau Kiat-gak, National Taiwan University Li Sheng-an, National Taiwan University Kao Cheng-yan, National Taiwan University
15:50-15:55	Analysis of Machine Translation Systems' Errors in Tense, Aspect, and Modality (11)	Masaki Murata, National Institute of Information and Communications Technology Kiyotaka Uchimoto, National Institute of Information and Communications Technology Qing Ma, Ryukoku University Toshiyuki Kanamaru, Kyoto University Hitoshi Isahara, National Institute of Information and Communications Technology
15:55-16:00	Enhanced Rocchio's Method for Better Text Categorization with Fractional Similarity Measure (15)	K.Lakshmi, Anna University Saswati Mukherjee, Anna University
16:00-16:05	Predicate Composition and the Determination of Scope (31)	Kenji Yokota, Akita University
16:05-16:10	Language Identification for Person Names Based on Statistical Information (36)	Shiho Nobesawa, Tokyo University of Science Ikuo Tahara, Tokyo University of Science
16:10-16:15	A Constrained Finite-State Morphotactics for Korean (38)	Eunsok Ju, Yonsei University Chongwon Park, University of Minnesota Duluth Minhaeng Lee, Yonsei University Kiyong Lee, Korea University
16:15-16:20	Word Order in Mandarin Chinese and Grammatical Relations (44)	Antoine Tremblay, University of Alberta
16:20-16:25	Speech-Activated Text Retrieval System for Cellular Phones with Web Browsing Capability (46)	Takahiro Ikeda, NEC Corporation Shin-ya Ishikawa, NEC Corporation Kiyokazu Miki, NEC Corporation Fumihiro Adachi, NEC Corporation Ryosuke Isotani, NEC Corporation Kenji Satoh, NEC Corporation Akitoshi Okumura, NEC Corporation

16:25-16:30	*Mei ci* (每次) and *Mei yi ci* (每一次): Differences between Two Forms of Event Quantifier in Mandarin Chinese (50)	Huang Zanhui, Sun Yat-sen University Pan Haihua, City University of Hong Kong
18:00-20:00	**Banquet**	

2005/12/2 (Friday)

Slot	Paper Title	Author and Affiliation
10:00-10:40	**Invited Speech: Suk-Jin Chang (Seoul National University)** **Form-Meaning Interface in Constraint-based Unified Grammar: Linking Prosody and Pragmatics** **Chair: Chin-chuan Cheng**	
10:40-11:00	**Coffee Break**	
Oral Presentation-- Session chair:		
11:00-11:25	Empirical Verification of Meaning-Game-based Generalization of Centering Theory with Large Japanese Corpus (42)	Shun Shiramtsu, Kyoto University Kazunori Komatani, Kyoto University Takashi Miyata, Japan Science and Technology Agency Koiti Hasida, National Institute of Advanced Industrial Science and Technology Hiroshi G. Okuno, Kyoto University
11:25-11:50	Discourse Segment and Japanese Referring Expressions: Are These Bare NPs or Proper Names? (61)	Etsuko Yoshida, Mie University
11:50-12:15	Japanese Bare Nouns as Weak Indefinites (30)	Keiko Yoshida, Waseda University
12:15-14:00	**Lunch**	
Poster Presentation (with 5 min oral presentation)-- Session chair:		
14:00-14:05	Using Speech Recognition for an Automated Test of Spoken Japanese (63)	Masanori Suzuki, Ordinate Corporation Yasunari Harada, Waseda University
14:05-14:10	Understanding Poetic Effects in Advertising Discourse: A Relevance-Theoretic Perspective (7)	Vincent Taohsun Chang , National Chengchi University
14:10-14:15	Analysis of The Elements by HPSG (22)	Satoshi Tojo, Japan Advanced Institute of Science and Technology Ken Saito, Osaka Prefecture University
14:15-14:20	On the Web Communication Assist Aide based on the Bilingual Sign Language Dictionary (23)	Emiko Suzuki, Tsukuba Gakuin University Kyoko Kakihana, Tsukuba Gakuin University
14:20-14:25	Anaphoric Resolution of Zero Pronouns in Mandarin Discourses (58)	Pan Haihua, City University of Hong Kong Cui Yuzhen, City University of Hong Kong Hu Qinan, City University of Hong Kong
14:25-14:30	XNLRDF, an Open Source Natural Language Resource Description Framework (13)	Oliver Streiter, National University of Kaohsiung Mathias Stuflesser, Institute of Applied Linguistics, European Academy of Bolzano
14:30-14:35	Acquisition of Concentrated Modification Structure from Logical Formula (28)	Hiroshi Sakaki, Meisei University
14:35-14:40	Enhancing Usability of Information Extraction Results with Textual Data Profiling (37)	Jyi-Shane Liu, National Chengchi University Yung-Wei Cheng, National Chengchi University

14:40-14:45	An Approach to Improve the Smoothing Process Based on Non-uniform Redistribution (43)	Feng-Long Huang, National United University Ming-Shing Yu, National Chung-Hsing University
14:45-14:50	Repair Structures in Web-based Conversation: (8)	Ruowei Yang, Open University of Hong Kong
15:15-15:35	**Coffee Break**	
Oral Presentation--Session chair:		
15:35-16:00	Vowel Sound Disambiguation for Intelligible Korean Speech Synthesis (51)	Ho-Joon Lee, Computer Science Division, KAIST Jong C. Park, Computer Science Division, KAIST
16:00-16:25	A Structured SVM Semantic Parser Augmented by Semantic Tagging with Conditional Random Field (52)	Minh Le Nguyen, Japan Advanced Institute of Science and Technology Akira Shimazu, Japan Advanced Institute of Science and Technology Xuan Hieu Phan, Japan Advanced Institute of Science and Technology
16:25-16:50	From Text to Sign Language: Exploiting the Spatial and Motioning Dimension (24)	Ji-Won Choi, Computer Science Division, KAIST Hee-Jin Lee, Computer Science Division, KAIST Jong C. Park, Computer Science Division, KAIST

2005/12/3 (Saturday)

Slot	Paper Title	Author and Affiliation
10:00-10:40	**Invited Speech: Suzanne Stevenson (Toronto University)** **Automatically Determining the Semantics of Multiword Predicates** **Chair: Tingting He**	
10:40-11:00	**Coffee Break**	
Oral Presentation--Session chair:		
11:00-11:25	Learning Translation Rules from Bilingual English – Filipino Corpus (18)	Michelle Wendy Tan, De La Salle University Raymond Joseph Ang, De La Salle University Natasja Gail Bautista, De La Salle University Ya Rong Cai, De La Salle University Bianca Grace Tanlo, De La Salle University
11:25-11:50	Integration of Dependency Analysis with Semantic Analysis Referring to the Context (33)	Yuki Ikegaya, Shizuoka University Yasuhiro Noguchi, Shizuoka University Satoru Kogure, Shizuoka University Tatsuhiro Konishi, Shizuoka University Makoto Kondo, Shizuoka University
11:50-12:15	A Small Fan and a Small Handful of Fans: Exploring the Acquisition of Count-mass Distinction in Mandarin (27)	Becky, Hsuan-hua Huang, University of Los Angeles David Barner, Harvard University Peggy Li, Harvard University
12:15-12:30	**Closing Ceremony**	

第七屆漢語詞彙語意學研討會（CLSW-7）

徵稿通知

http://www.fl.nctu.edu.tw/CLSW-7/homepage.htm

「漢語詞彙語意學研討會」爲中央研究院鄭錦全院士、北京大學俞士汶教授與中研院語言學研究所研究員黃居仁等共同倡辦。自 2000 年發起，由兩岸四地輪流舉辦，先後在香港、北京、台北、新加坡及廈門舉行。 2006 年「第七屆漢語詞彙語意研討會」(CLSW-7)將在台灣新竹交通大學舉行，竭誠邀請賜稿。

相關訊息如下：

一、 會期： 2006 年 5 月 22-23 日

二、 地點：國立交通大學浩然國際會議廳

三、 特邀講員：**Christiane D. Fellbaum** （美國普林斯頓大學）

四、 主題研討：第七屆漢語詞彙語意學研討會的目的是匯集各相關領域的學者，探討漢語詞彙語意學各個層面的問題。本次研討會所涉及的主題涵蓋漢語詞彙語意學的理論、方法、計算及其應用。具體包括但<u>不限於</u>以下所列的研究領域：

- 詞彙語意
- 詞彙與形式語意
- 詞彙與功能語法
- 詞彙與訊息處理
- 詞彙與認知
- 詞彙與句構
- 詞彙的神經心理基礎
- 詞彙與言談語用
- 詞彙化與語法化
- 詞彙網的建構

任何關於中文詞彙語意之研究皆歡迎賜稿。

五、 徵稿程序：分爲三階段（一）摘要徵求，（二）摘要審查，（三）全文集稿，截止時間如下：
（一）摘要徵求截止日： 2005 年 12 月 31 日
（二）通知審查結果： 2006 年 2 月 5 日
（三）全文寄送截止日： 2006 年 4 月 20 日

六、 摘要內容須包括標題、論文主旨、方法、取材、預期成果等。

七、 摘要格式：篇幅以A4 紙二頁爲限。
・ 請寄送<u>兩</u>份PDF檔案，一份具名、一份不具名。
・ 具名 PDF 檔案請附上作者姓名、服務或肄業單位、及 email 地址
・ 請一律以電子郵件附加PDF檔方式惠寄：clsw.nctu@gmail.com
・ 投稿人應於寄件 3 天內收到確認回函，方表示寄件成功。

八、 會議網址：http://www.fl.nctu.edu.tw/CLSW-7/homepage.htm

九、 聯絡方式：
國立交通大學外國文學與語言學研究所
劉小姐 (886)-3-573-1661

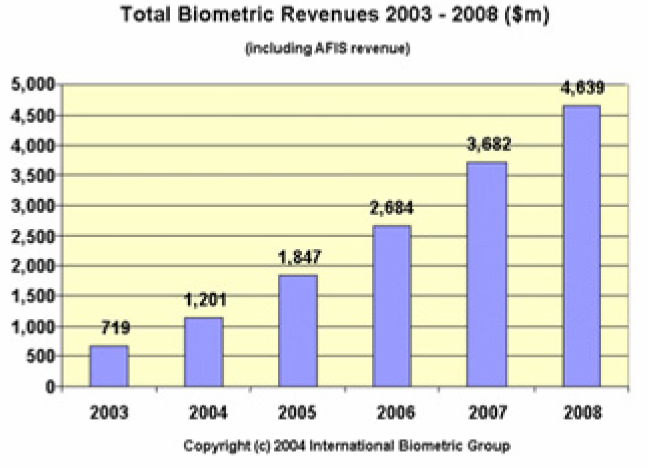
語音生物測定技術之簡介

蔡偉和、王新民
中央研究院資訊科學研究所

美國麻省理工學院（Massachusetts Institute of Technology, MIT）將「生物測定學」（Biometrics）或「生物認證」（Biometric Person Authentication）視爲最具改變世界的十大潛能科技之一。何謂生物認證？簡而言之，是以個人的生理特徵或行爲特徵來進行使用者身分識別與驗證。自動確認使用者身分無疑是達成安全防護的重要關鍵。近幾年由於電腦硬體、數位信號處理及圖像識別等技術的進步，促成了生物認證研究的蓬勃發展。包括指紋、語音、人臉、視網膜、虹膜、掌形、靜脈血管分佈、簽名及手勢等多種生物特徵都已被試圖利用爲身分確認之依據。根據生物特徵進行身分確認的好處甚多，其中較爲公認的優點包括：(1)僞造及破解困難、(2)使用方便、(3)適用性廣泛。舉凡需要確認使用者身分的場合都是這類技術可應用的範圍，例如門禁管制、電子商務及資料存取等。

回顧傳統的身分確認技術，一般可分爲兩類：「基於知識」(Knowledge-based) 與「基於憑證」(Token-based)。基於知識的方法是利用個人所知道的資訊來作爲身分確認的依據，例如使用密碼或 PIN (Personal Identification Number)即屬於此類的方法。基於憑證的方法則是利用個人所擁有的物品來作爲身分確認的依據，例如使用護照或身分證等。目前，仰賴密碼與證照的保全方式正面臨相當大的考驗，主要原因是個人帳號愈來愈多，密碼易被遺忘，加上現階段所使用的證照容易被盜拷複製，因此其實用性與安全性漸趨不足。相對地，利用生物特徵進行身分確認正可彌補這些不足。

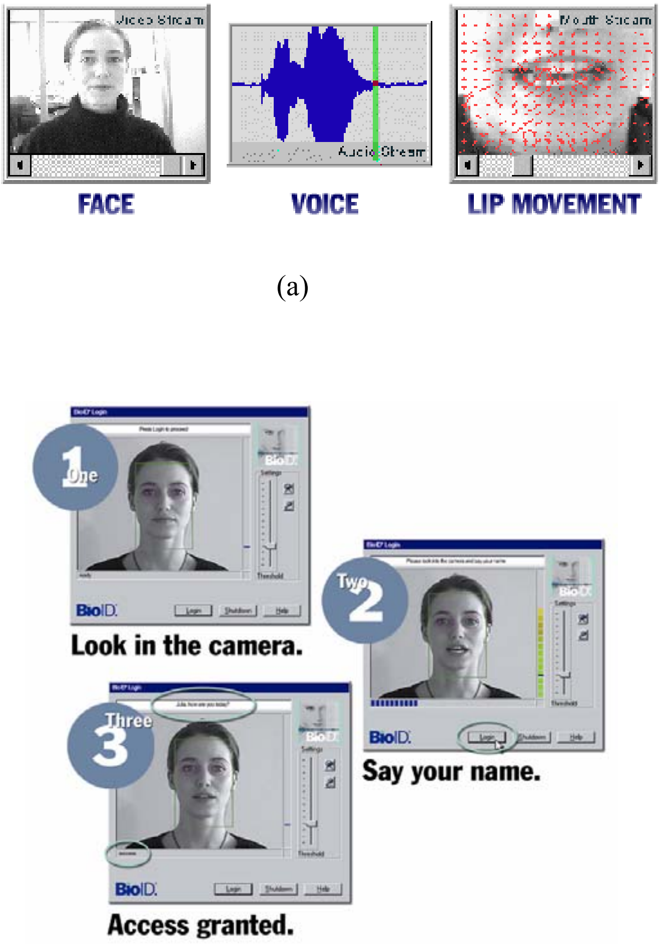
目前生物辨識技術已經被定位爲維護電子商務及網際網路安全的解決方案之一。圖一爲節錄自 Biometric Market Report 2003-2008 [1]之生物測定技術市場預估。西元 2008 年生物認證的市場將會達到 46 億美金，這些成長歸因於電腦與網路存取以及電子商務的蓬勃發展。目前世界上各先進國家都積極建構電子化政府，加上行動通信日益普及，這些因素都使得資訊保安之需求日益殷切，更確定了生物辨識技術未來的發展潛力。



圖一、生物辨識技術收入統計圖

爲了達到高安全性，高親和力與低佈建成本的要求，目前生物測定技術普遍的發展方向爲整合多模式生物特徵。整合方式可大略分爲三類：(1)可選擇性。系統使用多種生物特徵之身分確認

方式。使用者可以選擇其中的一種生物特徵來接受其身分檢測。例如一個身分確認系統可能提供人臉，語音和指紋辨識的認證功能，並允許使用者經由上述任何一種功能進行身分確認。系統的主要優點是使用方便，缺點是使用者必須針對每一種生物特徵，事先登錄相關的資料，系統則必須同時擁有擷取各種生物特徵所需的儀器設備。(2)非同步性。使用者在認證的過程中，必須依照一定的順序通過多個生物特徵認證關卡。例如使用者必須先通過指紋辨識確認身分後才能進行人臉辨識系統進行第二階段的認證。這類系統的主要優點在於提高身分確認系統的整體安全性，破解這種系統必須要分批破解多道生物特徵認證關卡，難度較高。主要缺點在於使用者必須學習各種不同的認證程序，而認證時間增加也造成了使用者的不便。(3)同步性。系統在一個認證的過程中同時執行多個生物特徵身分確認模組。例如使用者在認證的過程中同時包含了人臉和語音的辨識程序，如此一來，更加提高了破解系統的困難度，因為冒充者必須同時假冒多種生物特徵，難度更高。因為認證是同時進行的，所以使用者不需要依序通過不同的認證程序，所以節省了認證時間。圖二為一個整合人臉，語音與唇形動作的身分確認系統範例[2]。使用者看著攝影機說出自己的姓名（宣告自己的身分），系統則利用人臉，語音與唇形辨識進行身分確認。



(a)

(b)

圖二、BioID 認證系統範例

在生物辨識技術中，語者辨識（Speaker Recognition）是利用人類最自然的表達方式（即語音）作為辨識身分的依據。依使用目的不同，語者辨識一般可分為二大類：語者識別（Speaker Identification）與語者驗證（Speaker Verification）。語者識別必須事先對一組使用者收集各自的語音資料，然後抽取資料中的語者特徵參數或建立語者模型。當使用者對系統輸入測試語音時，系

統經由某種聲音相似度計算或模型比對，決定出說話者是誰。所以語者識別是從眾人中辨識出使用者的身分，亦即問「我是誰？」（Who am I？）。而語者驗證是用來鑑定處理對象所宣稱之身分的真實性，即判斷「我就是我所宣稱的那個人」事件之真偽（Am I who I claim I am？），也因此判斷錯誤的情況包括兩種，一是錯誤接受(False Accept)，即誤認爲冒充者是其所宣稱的人，另一爲錯誤拒絕(False Reject)，即將真實用戶誤認爲冒充者。另一方面，由於語者驗證的對象包括非用戶語者(冒充者)，其語音資料大多無法事先取得。

相對於利用其他生物特徵進行身分辨識的系統，語者辨識技術具有方便使用的優點，特別是可藉由電話或麥克風等透過電話線和網路進行遠端身分辨識。舉凡銀行業務等電子商務、個人網路資料的存取與其他資訊服務的安全管制皆可利用語者辨識技術來達成。軟體巨擘微軟創辦人比爾蓋茲在談到未來科技發展時曾說過：「語音科技不但是 Windows 的未來，更是整個電腦界的未來」。他強調，儘管個人電腦和網際網路已經徹底改變人類生活，但這些僅是剛開始而已，其實軟體研發仍在非常早期的階段，未來主要的技術突破將落在語音、書寫等辨識技術上。創造了摩爾定律的晶片巨擘英特爾的共同創辦人摩爾（Gordon Moore），不久前也曾在接受媒體訪談時預測「語音辨識將是可以大幅改變未來科技發展的關鍵領域」。語音科技與生物認證同樣深具發展潛力，現階段加速發展語者辨識(基於語音生物特徵的身分確認)技術是必要的。

國內外相關研究概況

在生物認證研究領域中，多模式生物特徵之身分確認技術在這最近幾年特別受到重視。The International Association for Pattern Recognition （IAPR）自 1997 年起每兩年舉辦一次 International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA)。這個研討會的主題是整合各種生物特徵之多種模式身分確認技術，並舉辦身分確認評比競賽。

在語者辨識研究方面，美國 NIST (National Institute of Standards and Technology)自 1996 年起，每年舉辦語者辨識評比（Speaker Recognition Evaluation, SRE）[3]，所有參賽單位均使用 NIST 提供的標準對話電話語料（Conversational Telephone Speech）。評比項目包含單一語者偵測（Single-Speaker Detection）、雙語者偵測（Two-Speaker Detection）、語者分段（Speaker Segmentation）與語者追蹤（Speaker Tracking）等，其中單一語者偵測之目的爲判斷一段語音是否爲假設語者（Hypothesized Speaker）的語音，相當於語者驗證(Speaker Verification)；雙語者偵測則則是從一段兩個人的對話中判別假設語者是否在其中；語者分段是藉由找出一段語音中各語者的聲音段落，進而將這些聲音段落依據語者分群；另外語者追蹤則是將一段語音中屬於某一假設語者的段落一一標示出來。近幾年在相關期刊包括 IEEE Trans. on SAP、Speech Communication、Computer Speech and Language、Digital Signal Processing 等與研討會 ICASSP、Eurospeech、ICSLP 等論文中大多已採用 NIST 語者辨識評比語料進行實驗，因此 NIST SRE 已是國際公認的語者辨識評比基準(Benchmark)。

美國約翰霍普金斯大學的語言及語音處理研究中心（The Center for Language and Speech Processing, The Johns Hopkins University）自 1995 年起每年舉辦一暑期研習會（Summer Workshop），挑選 4 個專題研究計畫，每個計畫由約 10 位來自各大學或研究機構的學者專家及學生負責進行，經過密集的腦力激盪及實驗，目的是爲相關研究領域開拓新方向或建立基準。語者辨識（SuperSID: Exploiting High-Level Information for High-Performance Speaker Recognition）是 2002 年獲選的四個計畫之一[4] ，而 2000 年的研習會亦曾針對Audio-Visual Speech Recognition [5]專題進行探討。該計畫的重點是利用唇形與語音的同步資訊進行語音辨識研究。另外有關Audio-Visual Speech Processing較重要的學術研討會則是International Speech Communication Association (ISCA)主辦的International Conferences on Auditory-Visual Speech Processing (AVSP)，1998 年首度舉辦，2005 年已邁入第五屆。

國際上從事語者辨識研究的知名單位包括MIT Lincoln Laboratory、The Human Language Technologies Group at IBM、The Speech Technology and Research (STAR) Laboratory at SRI、The Speech Group at Microsoft Research Asia等，而國內包括台大、清大、交大、成功、中山等大學及工研院、中科院、中華電信研究所等研究機構都曾經有過零星的研究，不過長期投入的單位並不多見。

主流技術概述

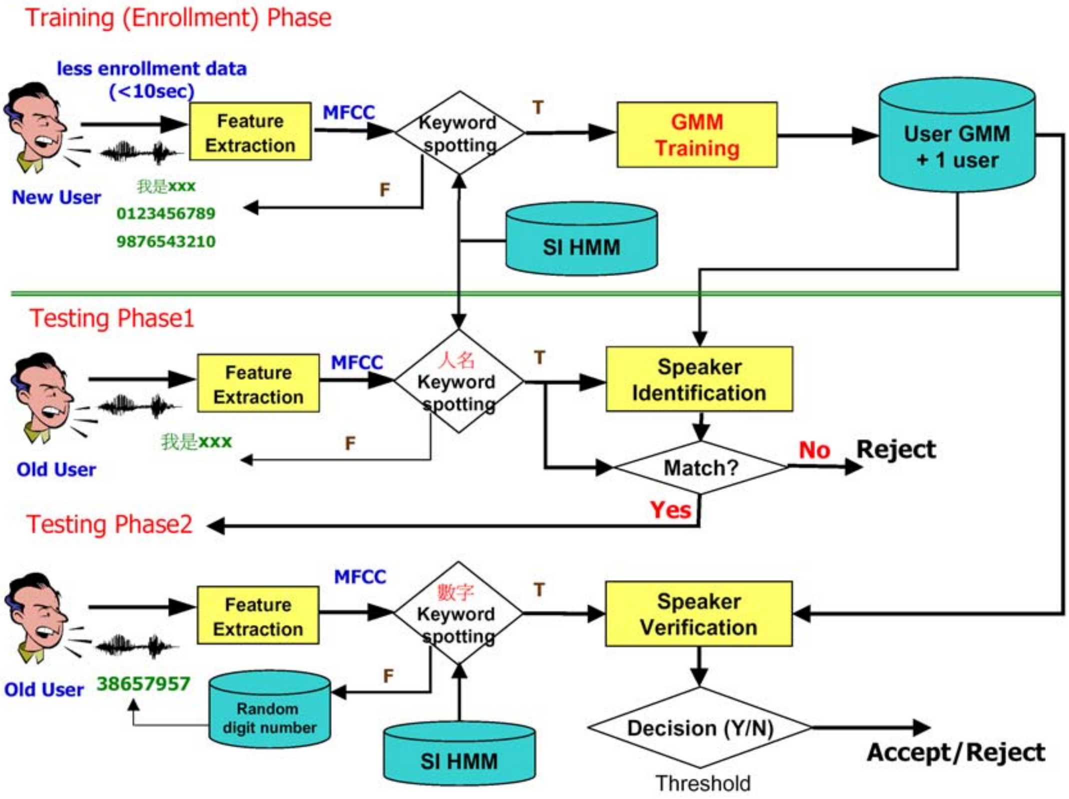
語者辨識之操作模式可分爲訓練與測試兩階段。訓練階段的工作主要是擷取各用戶之語音資料的語者特徵，一般將語者特徵表示爲參數模型，而測試階段則進行使用者輸入語音與資料庫中參數模型間的相似性比對。另一方面，語者辨識的方法可區分爲特定文本（Text-Dependent）與非特定文本（Text-Independent）二大類。特定文本方法限制使用者必須輸入與訓練語料相同的關鍵詞語或文句，而非特定文本方法則無此限制。由於特定文本之語者辨識已知測試語音內容(Linguistic Messages)，因此系統可事先建立各語者所屬的音素(Phoneme)模型，以控制某些與語者特徵無關的聲音變異。目前語者辨識的主流方法屬於統計模型分類法，對於特定文本之語者辨識最常使用的統計模型爲隱藏式馬可夫模型(Hidden Markov Model, HMM) [6]，而非特定文本之語者辨識則以高斯混合模型(Gaussian Mixture Model, GMM) [7]最爲普遍。這些統計模型主要用以描述語音頻譜特徵的靜態與動態分佈情形，且頻譜特徵大多表示爲倒頻譜係數，與語音辨識所使用者相同。

然而，由於一般基於統計模型的語者辨識技術使用與語音辨識相同的聲學特徵參數，其辨識效能往往因爲訓練語料不夠充足而嚴重衰退。爲了克服此問題，目前較常用的語者辨識方法是一種基於模型調適的架構。最具代表性者爲 Reynolds 等人[8]所提出的 GMM-UBM 方法。此方法利用具有大量非用戶語者的語料庫來訓練一通用背景模型(Universal Background Model, UBM)，該模型代表一般非特定語者的聲音特性。接著根據用戶語者各自的少量語料，利用最大後機率估算法(Maximum A Posteriori Estimation)，將通用背景模型調適成爲個別語者的特定模型(Speaker Dependent Model)。另一種調適方式是 Thyes 等人[9]所提出稱爲 Eigenvoice 的方法。其原理同樣是利用具有大量非用戶語者的語料庫來補捉一般非特定語者的聲音特性，但不同於 GMM-UBM 方法的是 Eigenvoice 將此一般非特定語者的聲音特性表示爲具有正規基底(Orthonormal Bases)的特性空間(Eigenspace)。主要步驟是先產生訓練語料中非用戶語者的個別模型，然後將每一模型中的參數串成一超向量(Super-vector)。接著利用主成分分析法(Principal Component Analysis, PCA)求出構成所有超向量之特性空間的基底。於是每位語者都可以表示爲特性空間上的一個投影點或座標，其中求取座標的方式僅需少量的語者登錄語料即可完成，而此座標也可以重建出語者的模型，稱爲EigenGMM。

本實驗室研究現況

本實驗室目前已完成一結合語音辨識及語者辨識之身分確認雛形系統。如圖三所示，此系統進行兩個階段的身分確認步驟：第一階段要求使用者說出個人身分(User Identity)，若通過後則進入第二階段：要求使用者說出由系統隨機產生的明碼(目前爲數字串)。這個作法的好處是冒充者若利用盜錄之使用者語音入侵系統，即使通過第一階段也將無法順利通過第二階段。由於已限定使用者說話內容，語音辨識部分採用關鍵詞擷取(Keyword Spotting)技術取代大字彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)技術，使系統辨識速度加快且辨識較準。由於第一階段的辨識結果可知道使用者所宣稱之身分，所以第二階段的語者辨識部分可採用語者驗證(Speaker Verification)技術取代需要把所有目標語者(Target Speakers)都計算一次相似度的語者識別(Speaker Identification)技術。當系統存在的目標語者數目很多時，語者驗證的速度將遠較語者識別快。目前新使用者登錄系統時只要輸入”我是 xxx(姓名)；0123456789；9876543210”即可。我們使

用一個基於貝氏資訊準則(Bayesian Information Criterion, BIC) [10]結合 Expectation Maximization (EM)的演算法[11]來訓練出語者的高斯混合模型，此法的優點在於可以根據訓練語料多寡及特性而動態決定出 GMM 的最佳混合數(Mixture Densities)。



圖三、結合語音辨識及語者辨識之身分確認雛形系統架構

另一方面，我們也嘗試運用不同於傳統操作模式的一種語者模型訓練策略，目標為同時達成減輕使用者操作負擔與提昇系統性能。考慮一般的語者辨識系統大多要求新用戶提供若干登錄語音資料作為建立該用戶模型參數的依據。然而為了減輕新用戶操作的負擔，系統的設計理念多朝向儘可能地降低用戶所須提供的登錄語料量，因此研究方法也著重於發展僅需極少量訓練語料條件下的語者辨識技術。但就系統可靠度而言，使用非常少量語料所訓練獲得的語者模型畢竟不如大量語料所建立的語者模型。因此，如何在系統可靠與操作簡便間取得平衡是一不易掌握的難題。為了解決此問題，我們的基本構想是不斷地利用增量學習(Incremental Learning)技術去調適初始僅用少量語料所建立的語者模型，使得用戶語者模型越來越好。同時，我們試圖在新用戶不知不覺的情形下取得其語音資料，用以調適其語者模型。主要原理是考慮在許多應用中，一個語者辨識系統除了在使用者進行登錄程序時可取得其語音資料外，其他時刻包括使用中與非使用中皆有機會收錄其語音資料。例如一門禁系統，用戶可能在進行身分確認操作前後正在使用手機談話，或與身旁友人交談，此時系統便可收集這些語音加以利用。另外，正與該用戶交談的身旁友人可能隨後將進行登錄程序而成為新用戶，事先收錄其語音資料將便可簡化該新用戶原本需要進行的惱人的錄音步驟。

欲達成上述操作模式，有賴於識別收錄語料的屬性或身分。首先，我們允許系統隨時處於偵測聲音的狀態，一旦所偵測出的聲音被判斷為語音訊號，則收錄至語料庫中。於是，語料庫將逐漸包含各式各樣語者(用戶與非用戶)的語音。由於在錄音的過程中，可能存在多位語者同時或非同時說話的情形，為了使收錄的語料可以善加利用，首要工作是將連續的語料長串進行切割，使之

成爲若干僅含單一語者之語音區段(Speaker Homogeneous Segments)。另外，系統需判別含有多位語者同時說話的語音區段(Overlapping Speech Segments)並予以剔除。經由此一處理後，語料庫將存有若干單一語者但未知身分的語音區段。接著，系統判斷每一語音區段所屬語者是否爲現有用戶或是其他未知語者。若屬於某一現有用戶之語音，則此資料將用於調適該現有用戶之模型參數，以彌補進行登錄時僅收錄少量語料的不足。若不屬於任何現有用戶之語音，則這些語音區段將被集合起來並進行分群。分群的目標是希望所有屬於相同語者的語音區段均被集中至同一群中，而任何屬於不相同語者的語音區段則被分開至不同群。依此方式進行，語料庫將存放若干單一語者但未知身分的語音區段群(Speaker Homogeneous Clusters)。由於這些未知身分的語者將來可能進行登錄而成爲合法用戶，因此，系統先利用每一群內的語音區段產生一組代表該群的模型參數，待任何一位新用戶欲進行登錄程序時，系統可決定由該新用戶依其語音「認領」一組模型參數或重新爲其建立一組新的模型參數。如此一來，系統可持續更新原有使用者的模型參數，同時快速建立新使用者的模型參數，不但提昇身分確認與辨識的效能，更能提供使用者簡便而人性化的操作模式。圖四所示即爲上述的整個流程架構。

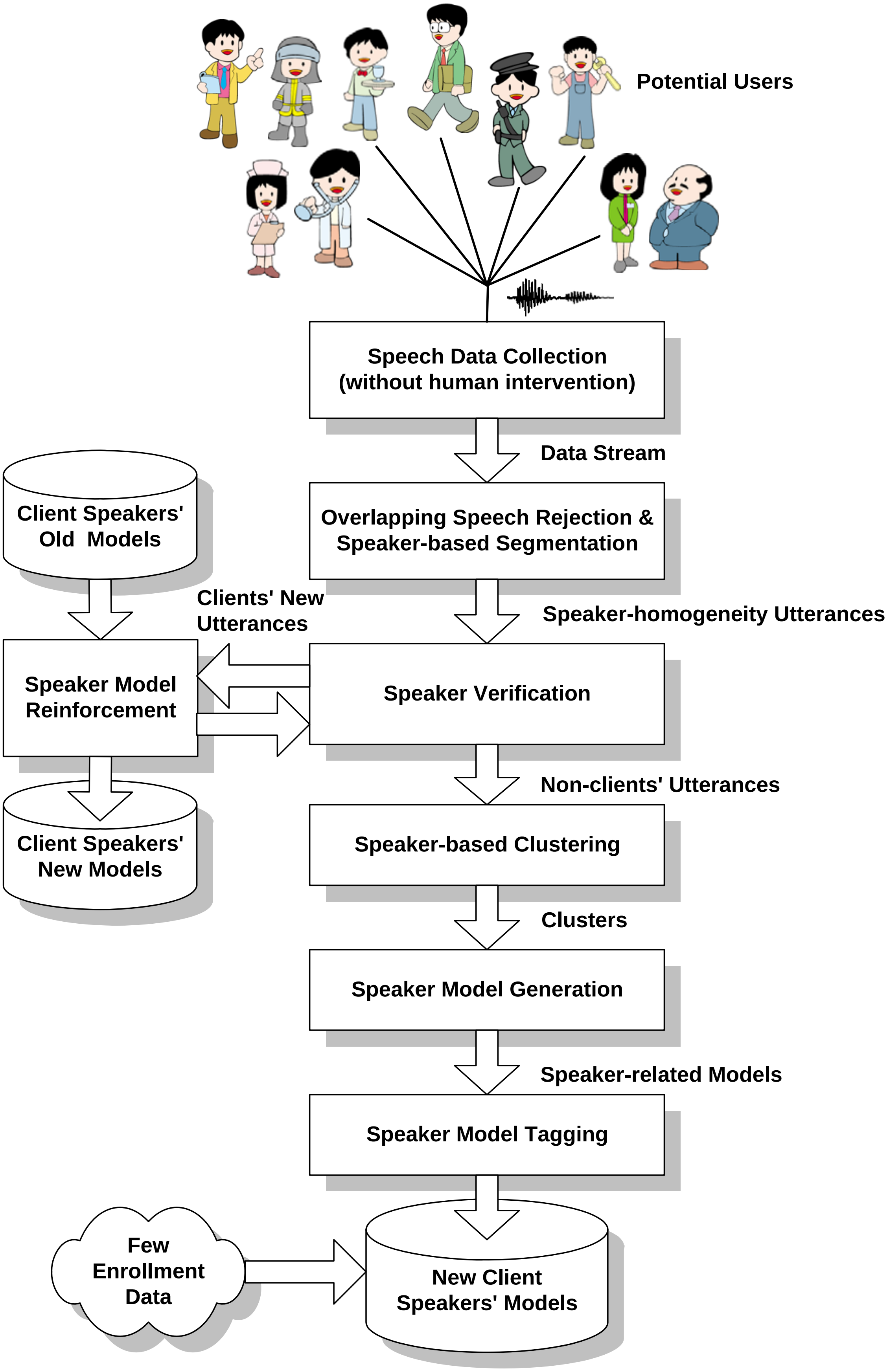
具體而言，我們已發展下列核心技術：

1. 自動語音偵測：需求爲有效區隔人聲與非人聲的辨識技術，以避免非語音資料影響系統參數模型的訓練與調適。我們應用語音訊號編碼常見的語音活動檢測（Voice Activity Detection, VAD）方法，包括利用語音的能量、頻譜等統計特性[12]以及模糊理論的決策機制[13]，搭配語音辨識的進一步驗證來決定訊號的屬性。

2. 語者分段(Speaker Segmentation)：在目前的研究中，語者分段的方法大致可分爲三類：(1)以模型爲基礎的分段(Model-based Segmentation)[14]：這類方法需事先訓練各種聲音的高斯混合模型(GMM)，例如音樂的模型、環境聲音的模型、人聲的模型等，然後將欲處理的音訊串流切分成固定長度的音段(如 3 秒)，再將這些音段與事先訓練好的聲音模型做比對，發生聲音模型變換的位置即是分段點。(2)以距離爲基礎的分段(Metric-based Segmentation) [15-18]：此類方法需先定義兩個語音區段的距離，然後利用一個滑動，固定長度或可變長度的分析窗框(Analysis Window)得到一距離曲線(Distance Curve)。在距離曲線中，數值大於所設定的閥值(Threshold)的尖端點(Peak)即可能是語者切換點。(3)前兩類方法的融合(Fusion of Model-based and Metric-based)。三類方法中，以距離爲基礎的分段法一般有較佳的效能，也較容易實作。有興趣了解我們在此問題上所提出之改良方法的讀者可參考論文[19][20]。

3. 重疊語音偵測(Overlapping Speech Detection)：在一般交談中，多位語者同時說話的情形相當普遍。然而，由於同時說話將造成不同來源之語音訊號彼此重疊，使得語者屬性難以界定，因此這類的語音資料應予以剔除。根據觀察，重疊語音與非重疊語音彼此在頻譜能量分佈上有著明顯的差異，目前我們使用一 GMM 分類器來區別重疊語音與非重疊語音。

4. 語者分群：給定一個語音區段的集合，語者分群的目的是要將屬於相同語者的語音區段集中至同一群中，而屬於不同語者的語音區段則被分開至不同群。我們將不屬於任何現有用戶之語音區段集合起來並進行分群，利用每一群內的語音區段產生一代表該群的模型。如此，當任何一位新用戶欲進行登錄程序時，系統可決定由該新用戶依其語音「認領」一模型或重新爲其建立新模型。我們在 2004 年語音訊號處理研討會中曾針對語者分群問題進行介紹，有興趣的讀者可至[21]下載講稿。



四、具備自動遞增學習及快速建立新使用者模型等功能之語者身分確認系統架構圖

5. 模型屬性標定(Speaker Model Tagging)與信任度量測(Confidence Measure)：當進行語者分群後，每一群內的語音區段將可假設爲相同語者所有，因此，系統可爲每一語音區段群建立一語者參考模型，這些參考模型將等候新用戶前來認領。認領的方式是由新用戶對系統任意說一句話，而系統將根據這句話找到與其最匹配的參考模型，此即爲模型屬性標定。回顧先前所述之典型的語者確認系統，其運作模式是由已知屬性的模型來判斷未知屬性的測試語句，這裡定義的模型屬性標定，其運作模式則是由已知屬性的語句來判斷未知屬性的參考模型。雖然運作模式相反，但語者確認與模型屬性標定基本上同爲假說測定(Hypothesis Testing)之問題。另外，考慮系統現有之待認領的模型可能皆不屬於或皆不適合一新用戶，或是有可能發生多位用戶認領同一參考模型的情形，因此在認領過程中必須加入信任度量測，使參考模型必須足夠適合某一新用戶時才予以指定爲該新用戶專用，否則該新用戶仍須進行正規的錄音登錄過程。最直接的信任度量測是由用戶自行決定，即聆聽並判斷所選模型所對應的部分語音區段是否爲其本人所說的話，但此方式必然造成使用者的負擔，因此由系統決定參考模型相對於使用者的信任度值有其必要性。我們的做法是：假設$\lambda_1^s,\lambda_2^s,\ldots,\ \lambda_N^s$爲系統中現有 N 個用戶的參考模型，$\lambda_1^c,\lambda_2^c,\ldots,\ \lambda_K^c$爲未知語料分割 K 群後所訓練獲得的參考模型，$\mathbf{X}$ 爲新用戶所提供的少量語料，則參考模型λ_i^c有資格接受該新用戶所認領的條件爲

$$i = \arg\max_{1\le k\le K} p(\lambda_k^c \mid \mathbf{X}),$$

並且

$$\log p(\lambda_i^c \mid \mathbf{X}) - \frac{1}{N+K-1}\left(\sum_{k\neq i}\log p(\lambda_k^c \mid \mathbf{X}) + \sum_{n}\log p(\lambda_n^s \mid \mathbf{X})\right) > \eta$$

其中η爲一閥值。在實際使用上可考慮先由系統認可後，再由用戶試用並決定是否認領。目前我們已驗證了此一語者辨識架構的可行性，並正著手進行上述技術的最佳化與整合。

結語

從“Minority Report”到“FACE OFF”，科幻電影中隨處可見生物測定學的蹤影。我們也經常可看到電影中歹徒以挖出他人眼睛或砍下手指等不擇手段方式冒充用戶入侵一安全系統的劇情。相較於視網膜、虹膜與人臉識別等生物測定技術，語者辨識可利用非限定文本模式避免冒充者以盜錄用戶語音的方式入侵系統，因此能達到較可靠且自然的安全防護。然而，目前以語者辨識爲主的安全系統幾乎仍不存在，主要原因歸咎於現階段的準確度仍無法達到令人滿意的地步。一般，視網膜與虹膜識別技術的錯誤率(接受或拒絕率)可達 0.1%以內，指紋與手掌掃描識別技術的錯誤率接也可低於 1%，然而非限定文本之語者辨識的錯誤率則普遍高於 10%。顯然，語音生物測定技術距離實用階段還很遙遠，而這也暗示著這方面的研究還有很大的空間。

參考文獻

[1] International Biometric Group http://www.biometricgroup.com/

[2] BioID http://www.bioid.com/

[3] NIST Speaker Recognition Evaluation http://www.nist.gov/speech/tests/spk/index.htm

[4] SuperSID Website http://www.clsp.jhu.edu/ws2002/groups/supersid/

[5] Audio-Visual Speech Recognition http://www.clsp.jhu.edu/ws2000/groups/av_speech/

[6] J. P. Campbell, “Speaker Recognition: A Tutorial,” *Proc. of the IEEE*, 85(9), pp.1437-1462, 1997.

[7] D. A. Reynolds, “Speaker Identification and Verification using Gaussian Mixture Speaker Models,” *Speech Communication*, 17, pp.91-108, 1995.

[8] D. A. Reynolds, T.F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, 10, pp.19-41, 2000.

[9] O. Thyes, R. Kuhn, P. Nguyen, and J. C. Junqau, “Speaker Identification and Verification Using Eigenvoices,” in *Proc. ICSLP*, 2000.

[10] G. Schwarz,“Estimating the Dimension of a Model,” *The Annals of Statistics*, 6:461–464, 1978.

[11] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, 39: 1-38, 1977.

[12] A. Pasanen,“Voice Activity Detection in Noise Robust Speech Recognition,” Master Thesis, Tampere University, 2002.

[13] F. Beritelli, S. Casale, and A. Cavallaro,“A Robust Voice Activity Detector for Wireless Communications Using Soft Computing,” *IEEE Transactions on Selected Areas in Communication*, 16(9), pp. 1818-1829, 1998.

[14] R. Bakis et al, “Transcription of broadcast news shows with the IBM large vocabulary speech recognition system,” in *Proc. Speech Recognition Workshop*, 1997.

[15] S. Chen, P.Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[16] P. Sivakumaran, “On the use of the Bayesian Information Criterion in multiple speaker detection,” in *Proc. EUROSPEECH* 2001.

[17] M. Cettolo and M. Vescovi, “Efficient audio segmentation algorithms based on the BIC,” in *Proc. ICASSP*, 2003.

[18] M. Siegler, U. Jain, B. Raj and R. Stern, “Automatic Segmentation, Clasification and clustering of broadcast News Audio,” in *Proc. Speech Recognition Workshop*, 1997.

[19] S. S. Cheng and H. M. Wang, “A Sequential Metric-based Audio Segmentation Method via The Bayesian Information Criterion,” in *Proc. Eurospeech*, 2003.

[20] S. S. Cheng and H. M. Wang, “METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation,” in *Proc. ICSLP*, 2004.

[21] http://sovideo.iis.sinica.edu.tw/SLG/speechworkshop2004/sp04-WHTsai.pdf

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。

A New Framework for Automatic Pronunciation Assessment Based on Distinctive Features

Chih-Chung Kuo, Cher-yao Yang, Ke-Shiu Chen, and Miao-Ru Hsu
Advanced Technology Center, Computer & Communications Research Labs,
Industrial Technology Research Institute Hsinchu, Taiwan
cck@itri.org.tw

Abstract

This paper presents a novel framework for automatic pronunciation assessment based on distinctive feature analysis. The major idea is to analyze learner's speech segment to verify whether it conforms to the correct combination of distinctive features. A *Distinctive Feature* (DF) is a primitive phonetic feature that distinguishes minimal difference of two phones [1]. The overall framework is organized as three layers: DF assessment (DFA), phone assessment, and continuous speech pronunciation assessment. Various methods can be designed to build DFA modules by extracting suitable acoustic features for each specific DF and classifying the features into score of opposite values. In contrast with conventional method that is based on speech recognition or verification of phonetic units (e.g. phonemes or syllables), the DFA is language independent and therefore universal. The performance of rudimentary experiments has shown the framework a feasible and compelling new approach.

1. Introduction

The ability to communicate in second language is an important goal for language learners. Students working on fluency need extensive speaking opportunities to develop this skill. But students have little motivation to speak out because of their lacking of confidence due to the poor pronunciation. The intent of pronunciation assessment systems is to provide learners with diagnosis of problems and improve conversation skill.

The computer-assisted pronunciation assessment (PA) can be divided into two categories: text-dependent PA (TDPA) and text-independent PA (TIPA). Most systems exploit prior speech recognition technology to evaluate the pronunciation quality [2][3][4]. TDPA constrains the text for reading to pre-recorded sentences. The learner's speech input is compared to the pre-recorded speech for scoring. The scoring method usually adopts template-based speech recognition technique like Dynamic Time Warping (DTW). On the other hand, the TIPA usually adopts speech recognition method of statistical approach like Hidden Markov Model (HMM). Witt [2] has further grouped the existing algorithms into two main classes: a-posteriori-based or classification-based confidence scores. No matter what approach is, the TIPA is language dependent because the statistic speech recognizer requires acoustic modeling of phonetic units.

Another crucial problem for the conventional automatic PA systems is that they provide very limited information except a pronunciation score or grade. The score itself give neither any diagnosis for the cause of pronunciation problems nor instruction for correction. Also, the scoring is general and fixed; teachers or students cannot emphasize on specific phonetic focus by adjusting the scoring mechanism. For pronunciation course designer, this is also a serious limitation.

Therefore, we proposed a novel approach based on the phonetic distinctive features (DF) of speech for pronunciation assessment. Each feature is an opposition between two relative values; for example, vocalic (or syllabic) sounds have a *relatively* clear formant structure in comparison with nonvocalic sounds [1]. Each speech phone may be described as a set of features. To evaluate the quality of a learner's pronunciation in phone level, we can thus check if one's pronunciation presents the correct value combination of distinctive features.

Using distinctive features for pronunciation assessments provides several important advantages:

- Text-independent PA
- DF assessment is language independent.
- Assessment result can offer information for problem diagnosis and correction instruction
- Users can control the scoring mechanism according to the learning or teaching focus.
- Phonological rules for continuous speech can be easily incorporated into the PA system.

These advantages will be explained and clear in the following sections.

2. Distinctive Feature

Linguists have long been using features or components to describe speech either explicitly or implicitly. Especially, it has been recognized that any language has a limited number of phonological contrasts or oppositions, called distinctive features. In 1952, Jakobson, Fant and Halle first expounded this approach [5]. Jakobson and Halle (1956) have proposed a set of only 12 features [6]. Chomsky and Halle (1968) have further emphasized on a universal set of distinctive features (27 features) [7]. They said of their feature system: “The total set of features is identical with the set of phonetic properties that can in principle be controlled in speech; they represent the phonetic capabilities of man and, we would assume, are therefore the same for all languages” (1968, pp.294-5) [7].

Table 1 shows part of the distinctive feature table of English phonemes used in our experiment. Although the concept of distinctive feature is universal, not all of the distinctive features are useful for a specific language. English, for example, requires about 16 features to distinguish all phonemes. Most of the features are polar opposites (symbolized by ‘+’ and ‘−’) for a specific phoneme. Some features, however, are not useful for some phonemes (to be distinguished from other phonemes), in which case the symbol ‘0’ is used.

Phoneme DF	e	ɛ	æ	o	ə	…	b	d	ʧ	ð	…
sonorant	+	+	+	+	+	…	−	−	−	−	…
sylllabic	+	+	+	+	+	...	−	−	−	−	...
consonantal	−	−	−	−	−	…	+	+	+	+	…
coronal	0	0	0	0	0	…	−	+	+	+	…
anterior	0	0	0	0	0	…	+	+	−	+	…
high	−	−	−	−	−	…	−	−	+	−	…
low	−	−	+	−	−	…	−	−	−	−	…
back	−	−	−	+	+	…	−	−	−	−	…
round	−	−	−	+	−	…	−	−	−	−	…
nasal	0	0	0	0	0	…	−	−	−	−	…
lateral	0	0	0	0	0	…	−	−	−	−	…
continunant	0	0	0	0	0	…	−	−	−	+	…
delayed release	0	0	0	0	0	…	−	−	+	+	…
tense	+	−	−	+	−	…	0	0	0	0	…
voiced	0	0	0	0	0	…	+	+	−	+	…
strident	0	0	0	0	0	…	−	−	+	−	…

Table 1: Distinctive feature table of English phonemes (partial). (‘+’ means the phoneme bears such feature, ‘−’ means not, and ‘0’ means redundant or irrelevant.)

3. Description of the Framework

The overall pronunciation assessment system is bottom-up organized as three layers: distinctive feature assessment, phone assessment, and continuous speech pronunciation assessment.

3.1. Layer 1—Distinctive Feature Assessment

The architecture of a distinctive feature assessor (DFA) is shown in Figure 1. Speech waveform is input into the DFA and goes through feature extraction module, binary classifier, and finally score mapper for the result. The output of a DFA (DF score) is a variable with value, without loss of generality, ranging from −1 to 1. One extreme value, 1, means the speech sound consists of the specified distinct feature with full confidence, -1 means extremely not. The DF score could also be defined as other value range such as [0, 1] or [0, 100].

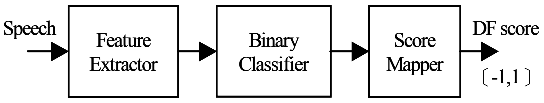


Figure 1: Distinctive Feature Assessor.

3.1.1 Feature Extraction

The DF could be described or interpreted either in articulatory or in perception point of view. However, for automatic detection and verification of DFs, only acoustic sense of them is useful. Therefore, we must define or find out appropriate acoustic features for each DF.

Different DF can be detected and identified by different acoustic features. Therefore, the most relevant acoustic features could be extracted and integrated to represent the characteristics of any a specific DF. Some acoustic features are more general that could be used for many DFs. The popular acoustic features used for conventional speech recognition, Mel-frequency cepstral coefficients (MFCC), are one apparent example. On the other hand, some features are more specific and can be used particularly to determine some DFs. For example, auto-correlation coefficients and formant estimation may help to detect such DFs like voiced, sonorant, consonantal, and syllabic. Some other possible examples of acoustic features include voice onset time, energy (low-pass, high-pass, band-pass), zero crossing rate, pitch, duration, etc.

Currently, we adopt the DFs defined by the linguists. In the future, we could re-define the set of DFs from the signal point of view so that the feature extractor can be more straightforward and effective.

3.1.2 DF Binary Classifier

DF classifier is the core of DFA. First of all, speech corpora for training are collected and classified according to the distinctive feature table. Then we can use the classified speech data to train a binary classifier for each distinctive feature. Many methods can be used to build the classifier, such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Neural Network (NN), Support-Vector Machine (SVM), etc. We can design and deploy different classifiers for different DFs so as to minimize the classification error and maximize the scoring effectiveness.

3.1.3 Score Mapper

Variant feature extractors and/or variant binary classifiers could be used for differnet DFs. Thus, a score mapper is designed to normalize the classifier scores to a common interval of values (for example, [-1, 1] in our experiments). This is to linearize and standardize the output for each DFA so that different designs of feature extractor and classifier can produce output of the same format with the same sense. This will assure the smooth integration of all DFAs in the next layer.

3.2. Layer 2—Phone Assessment

Multiple DFAs are integrated to construct a phone level assessment module as shown in Figure 2. The assessment controller can automatically determine a proper DF weighting for each DFA upon the input phoneme for assessment. It's naturally to refer to the DF table as in the Table 1. The users can also explicitly specify the distinctive features they wish to practice for pronunciation by setting the DF weighting factor (note that value 0 representing specific meaning of disabling the DFA). The output of each DFA can also be chosen between soft decision (that is a continuous value in the interval [-1, 1]) or hard decision (that is binary value –1 and 1). Finally, the integrated grader can be controlled to output various types of ranking result for the phoneme pronunciation assessment. It could be a N-levels or N-points ranking result (N>1). It could also be a vector of rankings for several groupings of DFs to express some learning goals.

.

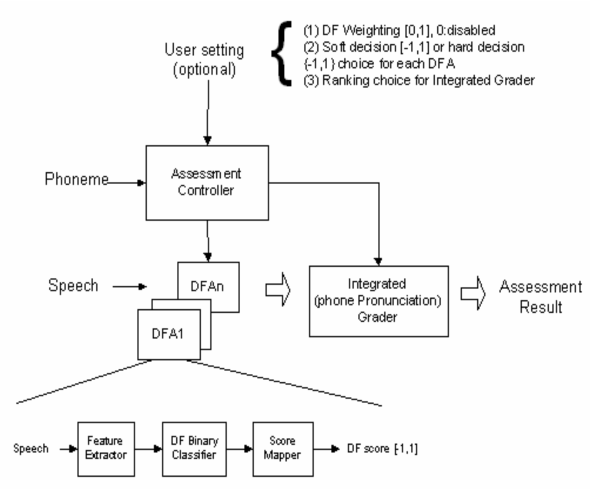

User setting
(optional)
(1) DF Weighting [0,1], 0:disabled
(2) Soft decision [-1,1] or hard decision
{-1,1} choice for each DFA
(3) Ranking choice for Integrated Grader
Phoneme
Assessment Controller
Speech
DFAn
DFA1
Integrated (phone Pronunciation) Grader
Assessment Result
Speech
Feature Extractor
DF Binary Classifier
Score Mapper
DF score [-1,1]

Figure 2: Phone-level assessment system architecture

3.3. Layer 3—Continuous Speech Assessment

The process for the overall system of continuous speech pronunciation assessment is shown in Figure 3. Inputs are continuous speech and its corresponding text. A text-to-phoneme (T2P) converter converts the text to phoneme string. Then the system uses the phoneme string to align the speech waveform to the phoneme sequence by the phone aligner. Then by using the phone assessor, we can obtain the score of each phoneme and integrate them to get the final pronunciation score for a word or a sentence. It should be noted that the T2P could be done by manually prepared information or by computer automatically on the fly. Phone alignment can be done by HMM alignment or any other means of alignment. The DF detection results can be optionally fed back to the aligner to adjust the alignment into a finer and more precise segmentation of speech waveform.

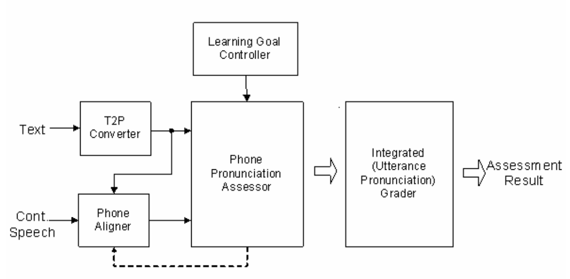


Figure.3: Continuous Speech Pronunciation Assessor.

4. Experiment

We conducted a rudimentary experiment to validate the feasibility of this new framework. English is adopted for experiment because it is the most popular second language being learned in the world. Thus, we constructed 16 DFAs for 40 English phonemes.

4.1.Database and Features

The speech corpora used for training and testing in the experiment are shown in Table 2. All speech data are of 8KHz sampling rate. In this rudimentary experiment, we adopt only MFCC as the acoustic features for all DFs. A feature vector is generated for each frame per 10 ms. The feature vector is of 27 dimensions, including 12 cepstra, 12 delta-cepstra, 1 energy, 1 delta-energy and 1 delta-delta energy.

		speakers	utterances	seconds
Training	male	49	11,191	
	female	23	11,621	
	total	72	22,812	146,869
Testing	male	5	686	
	female	5	699	
	total	10	1,385	10,215

Table 2: The speech corpora for experiment

4.2. Binary Classifiers Design

Two types of binary classifiers are designed for experiment: GMM and SVM. We trained two GMM models for each DF. One (positive model) was trained by the phone segments with positive value of DF; the other (negative model) was trained by the phone segments with negative DF value. Both GMM models are of 32 mixtures. All phone segments were generated by forced alignment of each speech utterance by HMM models. Likelihood scores were computed for each frame on the two GMM models. Finally, an average likelihood over all frames was computed for each GMM model. The likelihood ratio of positive model to negative model would result in the classification (or assessment) output.

For an alternative method of constructing the classifier, we also implemented Support-Vector Machine (SVM) [8]. We implemented four basic kernels: linear, polynomial, radial basis function, and sigmoid function for SVM. We found in a preliminary experiment that the SVM with linear and polynomial kernels out-perform the other two in our case. Finally, for computation simplicity we chose linear kernel to construct the binary classifier for DFAs. As in the case of GMM, all phone segments were generated by HMM forced alignment. Each SVM was trained by feature vectors on a frame-basis. However, the output for a phone segment is obtained by averaging outputs of all frames of the segment.

4.3. Classification Error

Although the assessment result should be a continuous value, we'd like to observe the performance of the DFA by the classification error. A threshold for each DFA was determined by equal classification error rate after the training phase. The total binary classification error rate was then computed in the testing phase. The result is shown in Table 3.

Destinctive Feature	GMM (%)	SVM (%)	Hybrid (%)
sonorant	34.83	17.38	17.38
sylllabic	62.04	29.37	29.37
consonantal	14.26	27.06	14.26
coronal	22.45	50.51	22.45
anterior	59.42	36.59	36.59
high	54.46	31.13	31.13
low	65.36	20.02	20.02
back	39.86	49.36	39.86
round	25.24	21.51	21.51
nasal	65.31	30.09	30.09
lateral	19.28	9.92	9.92
continunant	41.89	33.39	33.39
delayed release	45.83	36.59	36.59
tense	41.77	32.01	32.01
voiced	22.48	19.61	19.61
strident	69.56	17.32	17.32
AVG	42.75	28.87	25.72

Table 3: Classification error rates for all DFs

Clearly, SVM gets a better performance than GMM in average, however, not in all DFs. As we have mentioned, heterogeneous modules can be designed for different DFs. Therefore, if we chose the method (GMM or SVM) that give better performance (the shaded part in Table 3) for each DFA, the overall error rate dropped to 25.72%

4.4. Discussion and Future Work

Frankly speaking, we are still in a very beginning stage for the realization of this new framework. The design of the system described above is quite straightforward and thus requires further elaboration in the future.

- First of all, the acoustic features must be carefully designed according to each specific DF. MFCC is definitely not the best choice for most of the DFs. This is a big job because there are so many DFs. Another related issue is to define new and better DF from the acoustic view instead of linguistic view as mentioned before.

- Secondly, the simple average of binary classifier outputs over all frames is not a proper design. Similarly, the phone segmentation by HMM alignment is particularly a problem. We need to conceive an implicit alignment method, which imbeds in and fully in synergy with the phone assessment mechanism, just like that in the speech recognition with HMM. That is, the integrated output of all DFAs should be done per frame. The segmentation and assessment of each phone will then be done simultaneously.
- Finally, the phonological rule must be added to convert from phonemes to phones. That is, the Table 1 should be expanded into a distinctive feature table for all possible allophones in the target language. The values of DFs should depend on the realized phone in the context of continuous speech, not the original canonical phoneme. This is actually one of the advantages of this framework because the phonological rule can be easily expressed by the value change of distinctive features. Even an optional phonological rule can also be easily implemented by disabling the DFs that distinguish the allophones.

5. Conclusions

This paper presents a novel framework for pronunciation assessment based on distinctive features. We accidentally heard a talk given by Prof. C.-H. Lee (Georgia Institute of Technology) [9], who was making propaganda for his new speech research paradigm for next-generation ASR. We were surprised at the similar philosophy although his idea is even broader and more complete. All the more we were encouraged and convinced that this should be a promising direction either for pronunciation assessment or for ASR.

6. Acknowledgements

This paper is a partial result of the Project B34BPN2300 in OP23 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C. The authors would like to thank Mr. Nien-Chih Wang for the help in SVM.

7. References

[1] Clark, John and Yallop, Colin, *An Introduction to Phonetics & Phonology,* Basil Blackwell Ltd, Oxford UK, 1990.

[2] Witt, S. M., *Use of Speech Recognition in Computer-assisted Language learning,* Ph.D. dissertation, University of Cambridge, UK, 1999.

[3] Franco, H., Neumeyer L., Kim, Y., and Ronen, O. " Automatic Pronunciation Scoring for Language Instruction", *Proc. of ICASSP 97,* pp. 1471-1474, Munich, Germany, 1997.

[4] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, “Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech”, *Proc. of ICSLP 96*, pp. 1457-1460, Philadelphia, Pennsylvania, 1996.

[5] Jakobson, R., Fant, C. G. M. and Halle, M, *Preliminaries to Speech Analysis: the Distinctive Features and Their Correlates,* MIT Press, Cambridge, Mass., 1952.

[6] Jakobson, R. and Halle, M, *Fundamentals of Language,* Mouton, The Hague, 1956.

[7] Chomsky, N. and Halle, M, *The Sound Pattern of English,* Harper & Row, New York, 1968.

[8] Cortes C. and Vapnik. V., “Support-vector network”, *Machine Learning,* 20(3):273–297, 1995.

[9] Lee, C.-H., “From Knowledge-Ignorant to Knowledge-Rich Modeling: A Anew Speech Research Paradigm for Next-Generation ASR”, an invited talk in Academia Sinica, Taipei, Taiwan, Nov. 27, 2004.

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。