中華民國計算語言學學會
通訊 ACLCLP
The Association for Computational Linguistics and Chinese Language Processing
發行人：陳信希
主　編：王新民
執行編輯：黃　琪
第二十卷第二期

E-Mail：aclclp@hp.iis.sinica.edu.tw
Website : 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

《中文計算語言學期刊》線上投稿系統開放使用

《中文計算語言學期刊》投稿方式開始採用線上投稿，以方便作者投稿後隨時上線查看文章審查、出版進度，也提供審查者更便利的審查過程。該系統是由 Public Knowledge Project 發展的開放源碼，名爲「Open Journal Systems」，已獲得國外 2000 餘種期刊採用。

在使用者投稿或審稿前，必須先註冊取得帳號，才能登入系統進行後續動作。自行上網註冊時，可選擇作者(author)、審查者(reviewer)或讀者(reader)等任何一種或多種身份。目前本系統僅供投稿、審稿之用，本期刊出版之著作，仍以原有方式公開於學會網頁上，瀏覽閱讀時無須註冊。歡迎有意投稿者及擔任審查者上線註冊。(http://www.aclclp.org.tw/journal/submit.php)

IJCLCLP Online Submission is Now Available

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is now adopting an online submission system to assist both the authors and reviewers with the review process. The system developed by the Public Knowledge Project was an open source system and has been adopted by more than 2000 journals world wide.

Before using the system, users are required to register as an author, a reviewer, or both. With the user account, users can submit papers, review articles when asked, and monitoring the whole submission and reviewing process at any time.

Any users intended for paper submission or reviewing are all welcomed.
The online submission system is at:
http://www.aclclp.org.tw/journal/submit.php.

ROCLING-2009

「第二十一屆自然語言與語音處理研討會」將由國立中興大學資訊工程系、中興大學理學院及學會聯合主辦，謹訂於九月一日(週二)～二日(週三)假台中市國立中興大學理學大樓舉行，徵稿啓事及專題演講主講人介紹請參閱本刊第二～三頁。

學會職務異動

學會「口語處理組」召集人成功大學資訊工程系簡仁宗教授請辭，由「會員委員會」主任委員台北科技大學電子工程系廖元甫教授接任；「會員委員會」主任委員由暨南國際大學電機工程學系洪志偉教授接任，並均自3/27起生效。

Conference on Computational Linguistics and Speech Processing
第二十一屆自然語言與語音處理研討會

September 1-2, 2009, National Chung Hsing University, Taichung, Taiwan, ROC

http://rocling2009.cs.nchu.edu.tw/

CALL FOR PAPERS

The 21st ROCLING Conference will be held at National Chung Hsing University, Taichung, on September 1-2, 2009. Sponsored by Association for Computational Linguistics and Chinese Language Processing (ACLCLP), ROCLING is the most historic and major conference in the broad field of computational linguistics, speech processing, and related areas in Taiwan.

ROCLING XXI will be hosted by the Department of Computer Science and Engineering, National Chung Hsing University. The two-day conference will feature invited talks, papers, and poster sessions.

ROCLING XXI invites submissions of original and unpublished research papers on all areas of computational linguistics, natural language processing, and speech processing, including, but not limited to, the following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
(b) discourse/dialogue modeling
(c) information extraction/text mining
(d) information retrieval
(e) language understanding/generation
(f) lexicon/morphology
(g) machine translation/multilingual processing
(h) named entity recognition
(i) NLP applications/tools/resources
(j) phonetics/phonology
(k) question answering
(l) semantic web
(m) semantics/pragmatics
(n) speech analysis/synthesis
(o) speech recognition/understanding
(p) spoken dialog systems
(q) spoken language processing
(r) syntax/parsing
(s) text summarization
(t) web knowledge discovery
(u) word segmentation/POS tagging
(v) others

Paper Submission

Prospective authors are invited to submit full papers of no more than 25 A4- sized pages in PDF format (please go to the conference website to download the paper guideline and template). Papers will be accepted only by electronic submission through the conference website. The submitted papers should be written in either Chinese or English, and in single column, single-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, followed by a concise abstract of the paper.
Papers should be made in PDF format and submitted online at:
http://rocling2009.cs.nchu.edu.tw/papers/

Best Paper Award

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

Important Dates

Preliminary paper submission due:	**July 2, 2009**
Notification of acceptance:	August 1, 2009
Final paper due:	August 10, 2009
Conference date:	September 1-2, 2009

Sponsors

Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
Department of Computer Science and Engineering, National Chung Hsing University
College of Science, National Chung Hsing University

ROCLING-2009 Keynote Speakers

1. Prof. Chin-Hui Lee (Georgia Institute of Technology)

Chin-Hui Lee received his undergraduate degree in Electrical Engineering from National Taiwan University in 1973, his master's in Engineering and Applied Science from Yale University in 1977, and his doctorate in Electrical Engineering with a minor in Statistics from the University of Washington in 1981.

After graduation, Dr. Lee joined Verbex Corporation where he conducted research on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation where he was engaged in research and product development in speech coding, speech synthesis, speech recognition and signal processing. From 1986 to 2001, he was with Bell Laboratories where he became a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department.

In August 2001, he accepted a one-year appointment as a visiting professor at the National University of Singapore's School of Computing. He joined the ECE faculty in September 2002.

Dr. Lee has published more than 250 papers and holds 25 patents. He edited the book Automatic Speech and Speaker Recognition: Advanced Topics, and has contributed chapters to ten books.
Homepage: http://www.ece.gatech.edu/faculty-staff/fac_profiles/bio.php?id=125

2. Dr. Keh-Jiann Chen (Institute of Information Science, Academia Sinica)

Keh-Jiann Chen obtained a B.S. in mathematics from National Cheng Kung University in 1972. He received a Ph.D. in computer science from the State University of New York at Buffalo in 1981. Since then he joined the Institute of Information Science as an associate research fellow and became a research fellow in 1989. He was the deputy director of the institute from August 1991 to July 1994.

His research interests include Chinese language processing, lexical semantics, lexical knowledge representation, and corpus linguistics. He had been and continued in developing the research environments for Chinese natural language processing including Chinese lexical databases, corpora, Treebank, lexical analyzer and parsers.

Dr. Chen is one of the founding members of the Association of Computational Linguistic and Chinese Language Processing (also known as ROCLING). He had served as 2nd term president of the association from 1991 to 1993. Currently he is the board member of the Chinese Language Computer Society, the advisory board member of the International Journal of Computational Linguistics and Chinese Language Processing, and the editor of journal of Computer Processing of Oriental Language.
Homepage: http://www.iis.sinica.edu.tw/pages/kchen/index_zh.html

Call for Papers
International Journal of
Computational Linguistics & Chinese Language Processing

Special Issue on
Computer Assisted Language Learning

Applications of natural language processing (NLP) techniques in language learning and assessment have drawn much attention in recent years. Several successful systems have been reported, including bilingual concordaners, automated essay scoring, essay critiquing, writing aid, collocation checkers, among others. The goal of this special issue is to report the state-of-the-art NLP applications in language learning. Prospective authors are invited to submit their innovative works and review articles to this special issue. We are soliciting paper submissions in topical areas including, but not limited to:

- Context-sensitive spelling checking
- Speech technology and language learning
- Automatic item generation
- Automatic essay scoring
- Automatic identification of grammatical errors
- Applications of chatbot in language learning
- Context-sensitive writing aid
- Context-sensitive glossing and translation
- Semantic web applications in language learning
- Adaptive text selection, testing, and course sequencing using NLP techniques
- Corpus-based NLP techniques in language learning

Schedule

Submission deadline:	July 1, 2009
Notification of acceptance:	September 1, 2009
Final manuscript due:	November 1, 2009
Tentative publication date:	December 15, 2009

Instructions for Authors

All manuscripts are subject to anonymous peer review. The template file for manuscripts is available at the homepage of the International Journal of Computational Linguistics & Chinese Language Processing (http://www.aclclp.org.tw/journal/index.php). Authors should submit their papers in PDF format via the aforementioned web page by registering new accounts.

Guest Editors

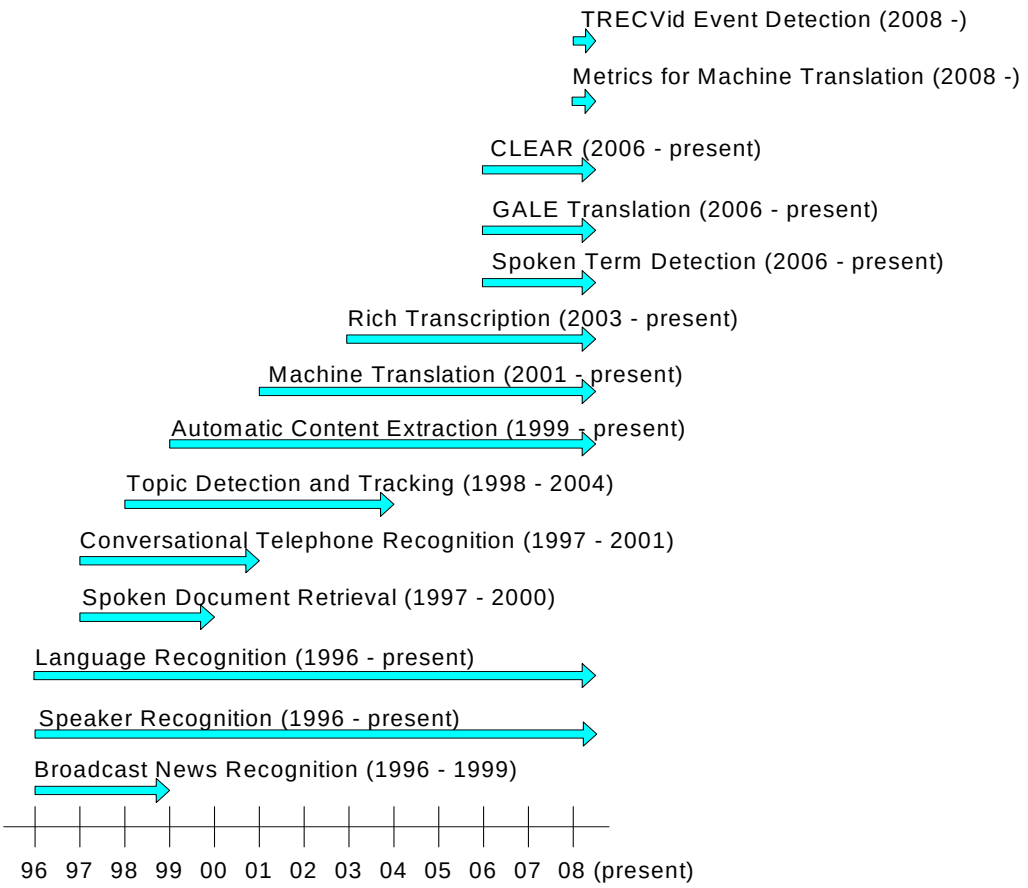
Dr. Chao-Lin Liu
Dept. of Computer Science, National Chengchi University
Email: chaolin@nccu.edu.tw

Dr. Zhao-Ming Gao
Dept. of Foreign Languages and Literatures
National Taiwan University
Email: zmgao@ntu.edu.tw

音訊內容之語者自動分段標記技術簡介
Introduction to Speaker Diarization

[1]蔡偉和、[2]鄭士賢、[2]王新民
[1]國立台北科技大學電子工程系、[2]中央研究院資訊科學研究所

自1996年起，美國國家標準與技術研究院（National Institute of Standard and Technology，NIST）舉辦了無數次的語音辨識相關技術評比(Benchmark Tests)，藉由訂定標準的效能量測方法與建立特定任務(task)測試語料庫，比較世界各研究單位之辨識系統效能，以促進state-of-the-art技術的不斷提昇。NIST所舉辦的評比項目可歸納於圖一，其中除了「語者辨識」(Speaker Recognition)與「語言辨識」(Language Recognition)評比自1996年起迄今仍持續進行外，許多評比項目常隨著實際應用需求的變化而被其他新的評比項目所取代。例如最早的語音辨認技術評比爲「廣播新聞辨認」(Broadcast News Recognition)；自1999年後，技術較量的舞台已轉移至「口語文件檢索」(Spoken Document Retrieval)與「電話對話辨認」(Conversational Telephone Recognition)。自2002年後，語音辨認、語者辨識、與文件檢索等概念更進一步整合爲一共同的評比項目，稱爲Rich Transcription (RT)。其目的是希望使自動語音辨識結果有更高的可讀性，讓人們更有效地運用大量的語音資料。

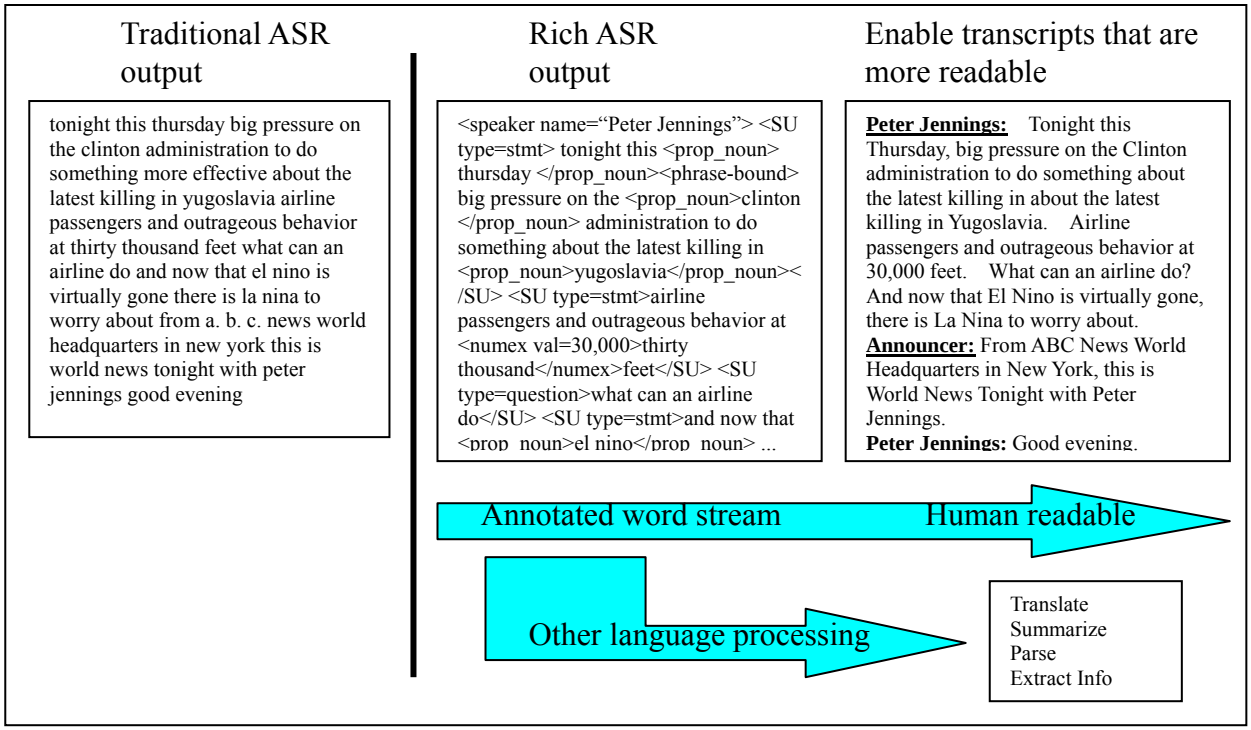


圖一：NIST Benchmark Tests

RT的任務定義如圖二所示，其重點包括兩項：一爲「音轉字謄寫」(Speech-to-Text Transcription, STT)，另一爲「後設資料擷取」(Metadata Extraction, MDE)。圖三說明RT和傳統自動語音辨識輸出的具體差異。目前RT任務使用的標準語料庫包括廣播新聞語料、電話對談語料及會議錄音語料。STT評比涵蓋英文、中文和阿拉伯文，但MDE評比目前只考慮英文。在MDE評比類別中，有一個項目稱作「語者分段標記」(Speaker Diarization)，又稱「Who Spoke When」[Canseco-Rodriguez2004] [Tranter2006]，顧名思義，就是要在一段錄音資料中區分出不同說話者的說話區段，並一一標示出來。這項工作主要涉及三個步驟：(1) 將音訊自動切割成爲很多小區段，目標是每一小區段只包含一個說話者；(2) 對這些小區段進行自動分群，希望每一群集都只包含一個說話者的聲音；(3) 判別每一群集的性別，給予一個說話者識別身分，最後與STT產生的自動語音辨識結果整合。

Input: Human-human speech (e.g., broadcasts and conversations)
Output: Rich transcript (words + metadata) accurate enough for
Machines to detect, extract, summarize, translate
Humans to read & understand easily

圖二：RT 的任務定義

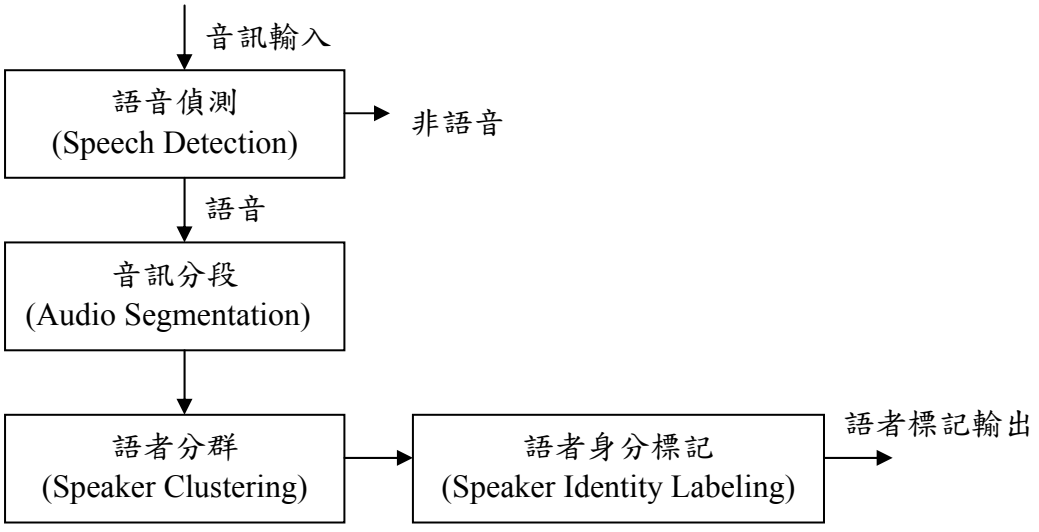


圖三：RT 和傳統自動語音辨識的差異

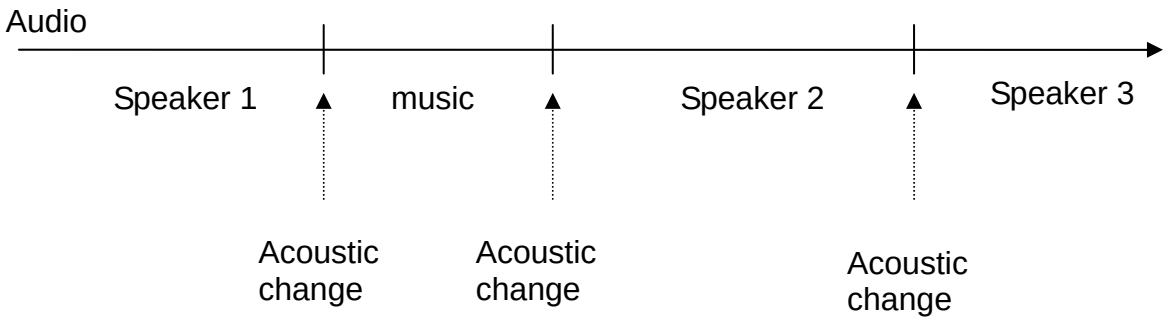
國際上之 Speaker Diarization 研究情況

目前，國際上投入此項研究較具代表性的機構包括MIT Lincoln Laboratory, CMU, Cambridge University, University of Washington, Brown University, LIMSI, IBM, SRI International, International Computer Science Institute等。根據NIST所訂定的規則，評估Speaker Diarization之效能好壞是計算Diarization Error Rate (DER)，其定義爲輸出結果中錯誤的說話者區段長度與所有說話者總長度的比例。由於Speaker Diarization只是對切割出來的小區段分群，然後給每一群集一個說話者識別身分，並不一定辨識出每一群集真正的說話者身分，所以是在輸出結果與人工標記獲得最佳對應下計算錯誤說話者區段長度，這些錯誤包括一小區段語音內包含不同說話者聲音的分段錯誤、區段被分到錯誤說話者群集的分群錯誤及同一說話者的不同區段被分到兩個或兩個以上群集的假警報（False Alarm）錯誤。計算DER的工具可以從NIST網站（http://www.nist.gov/speech/tests/rt/）下載。

圖四是一般Speaker Diarization系統的基本步驟。語音偵測之目的在於區分一音訊串流（Audio Stream）之何處屬語音片段與何處屬非語音片段，而音訊分段的目的在偵測聲學特性變換點(Acoustic Change)，例如：語者的變換、背景聲音的變換等等，如圖五所示。另外語者分群的目的是將屬於相同語者之語音片段合爲一群，屬於不同語者之語音片段分爲不同群。此項技術自1997年DARPA 舉辦廣播新聞自動文字轉寫評比之後即吸引許多研究人員投入研究，主要原因是其在語音處理領域有很多應用，例如(1)語者追蹤 [Bonastre2000][Lu2002a]：一般的作法是先利用音訊分段技術將輸入之音訊串流分割成同質音段(Homogeneous Segment)，然後再對輸出之音段進行語者識別；(2)語者調適語料之自動收集[Chen1998b]：對一音訊串流進行音訊分段之後，再實行語者分群或者語者識別，則可自動地收集串流中各語者之語料；(3)語音檢索[Wang2004]：音訊串流在經音訊分段與語者分群後，可將其分割成更有意義之檢索單位。

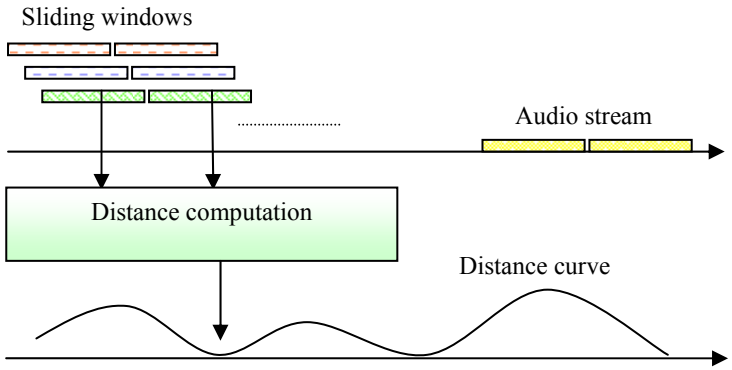


圖四：Speaker Diarization系統基本步驟



圖五：音訊串流中的聲學特性變換點(Acoustic Change)

目前「音訊分段」的方法大概可分成三類：(1)以模型爲基礎（Model-based）的方法[Bakis1997]：這是一種監督式(Supervised)的方法，需事先假設音訊串流中包含之聲音類別爲已知，且需要事先收集這些聲音類別的訓練語料以訓練其統計模型，然後藉由對音訊區段分類的結果來判斷聲學特性變換點；(2)以距離爲基礎（Metric-based）的方法[Zhou2000][Cettolo2003][Delacourt2000]：這是一種非監督式(Unsupervised)的方法，如圖六所示，藉由定義兩音段之間的距離量測，以及發展基於此量測的變換點偵測演算法來進行音訊分段。通常使用倒頻譜係數（Cepstral Coefficients）作爲信號之特徵值，透過統計模型來計算兩音段間的距離量測；(3)以信號特徵爲基礎的方法[Zhang2001][Lu2002b]：這類方法首重發展具鑑別性之信號特徵，以期更能鑑別兩音段之差異。



圖六：以距離為基礎（Metric-based）的聲學特性變換點偵測法

關於「語者分群」，（或可稱「非監督式的語者辨認」[Lapidot2002] [Makhoul2000]），如圖七所示，其目的是將未知語音片段按其所屬未知語者進行分群，使相同語者之語音片段合爲一群，而不同語者之語音片段分開爲不同群。目前最普遍採用的方式[Gish1991][Jin1997][Solomonoff1998][Chen1998a][Reynolds1998][Zhou2000][Moh2003][Liu2005]是先量測兩兩音訊片段間的相似度，再利用「階層式分群法」(Hierarchical

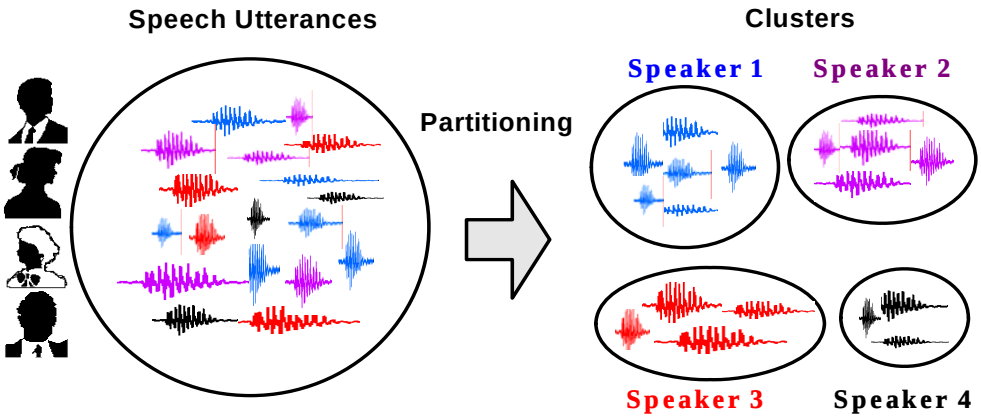
Clustering)逐步將相似度高的音訊片段進行合併(Agglomeration)，然後建立一「群樹」(Cluster Tree)，輸出各種不同群數的分群結果，最後再決定出最佳的群數。圖八所示爲一階層式凝聚分群 (Hierarchical Agglomerative Clustering, HAC)之概念圖。假設有*N*個欲進行分群之音訊片段$\mathbf{X}_1, \mathbf{X}_2,\ldots, \mathbf{X}_N$，則首先產生*N*個群集，在每一群集中不重複地放入一個音訊片段。然後，將*N*個群集中任意兩個最爲相似的群集合併，此時群集總數成爲*N*-1，而這裡因每一群集中只含一個音訊片段，因此任兩群集之間的相似性即是任兩音訊片段之間的相似性。接著，將*N*-1個群集中任意兩個最爲相似的群集合併，此時群集總數成爲*N*-2，而這裡因每一群集中可能不只含有一個音訊片段，所以群集之間的相似性必須重新估算，一般是藉由音訊片段之間的相似性來估算，方法大致包含下列幾種：

(i) *complete linkage* $S_c(c_i, c_j) = \min_{\mathbf{x}_n \in c_i, \mathbf{x}_k \in c_j} S(\mathbf{X}_n, \mathbf{X}_k)$,

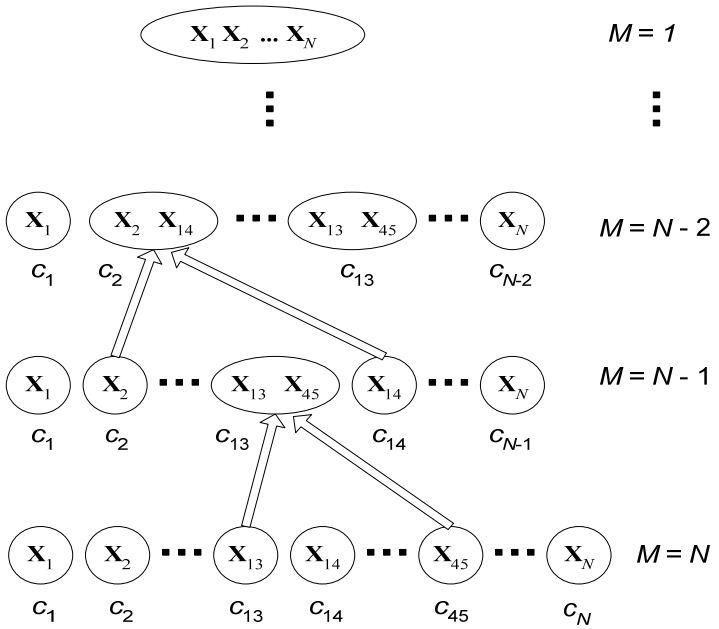
(ii) *single linkage* $S_c(c_i, c_j) = \max_{\mathbf{x}_n \in c_i, \mathbf{x}_k \in c_j} S(\mathbf{X}_n, \mathbf{X}_k)$,

(iii) *average linkage* $S_c(c_i, c_j) = \frac{1}{\#(\mathbf{x}_n \in c_i, \mathbf{x}_k \in c_j)} \sum_{\mathbf{x}_n \in c_i, \mathbf{x}_k \in c_j} S(\mathbf{X}_n, \mathbf{X}_k)$,

其中 $S(\mathbf{X}_n, \mathbf{X}_k)$ 代表兩音訊片段$\mathbf{X}_n$與$\mathbf{X}_k$之間的相似性， $S_c(c_i, c_j)$ 代表兩群集*ci*與*cj*之間的相似性，而 $\#(\mathbf{X}_n \in c_i, \mathbf{X}_k \in c_j)$ 代表所有 $\mathbf{X}_n \in c_i, \mathbf{X}_k \in c_j$ 的數目。另外，目前最普遍用來計算兩音訊片段間之相似性的方法包括KL距離(Kullback Leibler Distance)或泛似然率比(Generalized Likelihood Ratio)。



圖七：語者分群概念圖



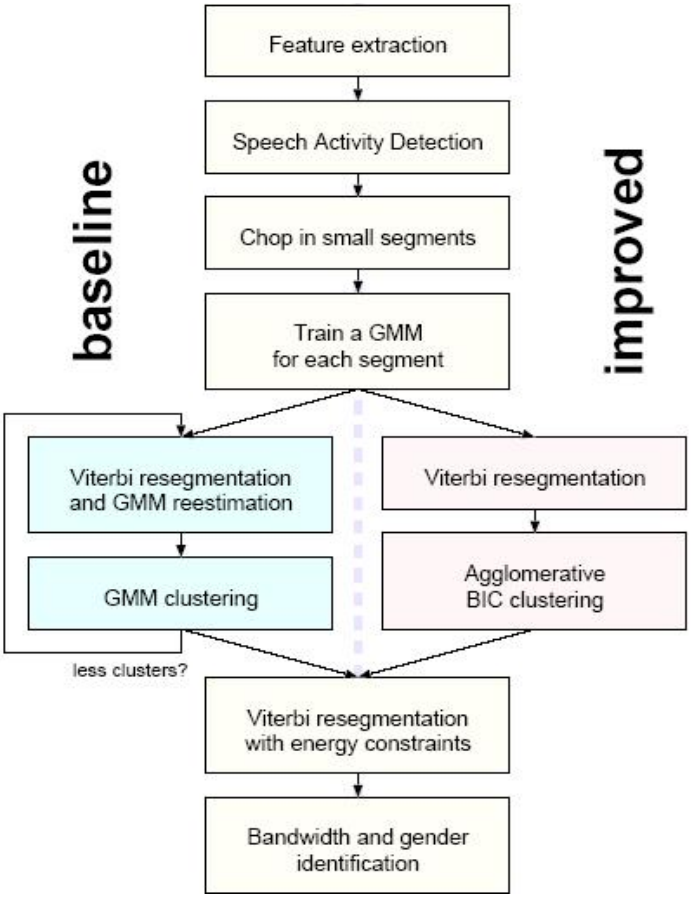
圖八：階層式凝聚分群法示意圖

以 LIMSI [Barras2004][Zhu2005]爲例，如圖九所示，他們的基本作法是事先利用訓練語料，對語音、音樂、背景聲三類信號分別訓練出一個高斯混合模型（Gaussian Mixture Model, GMM）；然後第一步先利用上述三個 GMMs，以 Viterbi 演算法對音訊信號進行初始分段；第二步先假設每一小段自成一群，然後反覆進行 Viterbi 重新分段、GMM 重評估和 GMM 分群，直到結果收斂不變爲止，最後輸出信號分段和分群結果；第三步是參考信號能量曲線，利用最後的群 GMMs 重新調整邊界；最後進行信號傳輸通道種類（例如麥克風或電話語音）和說話者性別識別。在其隨後所發表的研究中，第二步改採 Viterbi 重分段和 BIC（Bayesian Information Criterion）凝聚式分群。所謂的 BIC 凝聚式分群是先將每一小段視爲一群，各自訓練一個 GMM，然後計算任兩群間的Δ*BIC* 值，將Δ*BIC* 值最小的兩群合併，反覆進行，直到任兩群都無法合併爲止。爲了運算效率考量，通常 GMM 是以「單一高斯密度模型」(Uni-Gaussian Model)來實現。給定任意兩群 *ci 與 cj*，假設它們的樣本向量數分別是 n_i 和 n_j，總和是 n，Σ、Σ_i 和 Σ_j 分別是 n、n_i 和 n_j 個樣本向量求得的共變異數矩陣（Covariance Matrix），c_i 和 c_j 間的Δ*BIC* 值可由下式求得：

$$\Delta BIC = (n_i+n_j)\log(|\Sigma|) - n_i\log(|\Sigma_i|) - n_j\log(|\Sigma_j|) - \lambda P ,$$

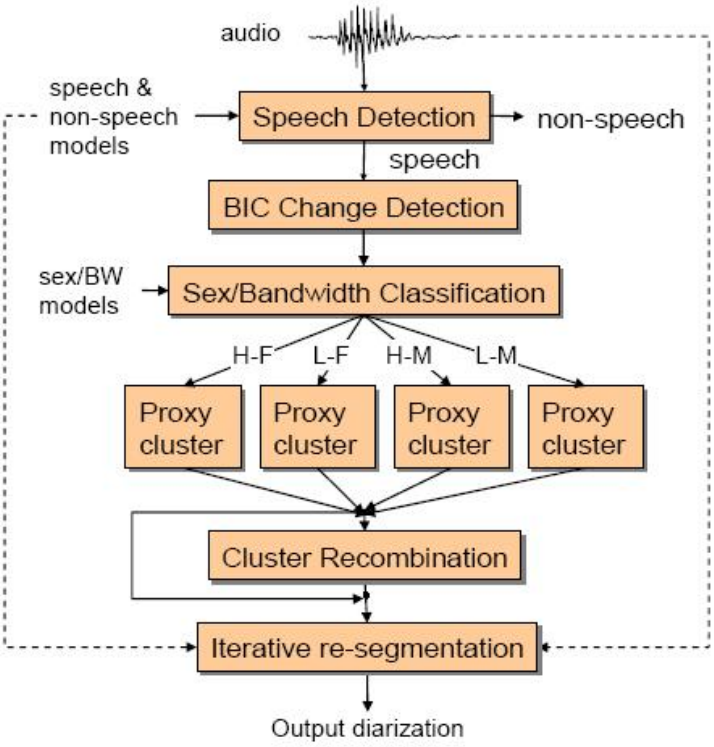
其中 P 是懲罰值（penalty），計算如下：

$$P = \frac{1}{2}(d + \frac{1}{2}(d)(d+1))\log n \ 。$$



圖九：LIMSI 之 Speaker Diarization 系統架構

圖十則是 MIT Lincoln Laboratory 採用的 Speaker Diarization 系統架構[Reynolds2004][Tranter2006]，他們先用分別代表純語音、有背景音樂語音、有其他背景聲音語音、音樂及其他背景聲音的 5 個 GMMs 來剔除音訊中不含語音信號的區段；再用 IBM 提出的 BIC 音訊分段方法[Chen1998a]將長語音段分割成同質的（Homogeneous）小區段；然後利用 H-F（高頻寬女性語音）、L-F（低頻寬女性語音）、H-M（高頻寬男性語音）及 L-M（低頻寬男性語音）4 個 GMMs 將上述小區段分別歸到四類語音；最後針對四類語音分別進行自動分群。他們的方法中最大的特色在於分群的方式，作法主要是參考 Sturim 等人在 2001 年提出來的 anchor model 方法[Sturim2001]，這個方法的基本概念是事先找 N 組說話者的語音，每人訓練一個 GMM，所以一共有 N 個 GMMs，待分群的每一小段語音，需分別對這 N 個 GMMs 計算 likelihood，因此，每一小段語音可以表示成一個 N 維的特性向量（Characteristic Vector），所以，音段分群的問題便轉換成向量分群的問題。

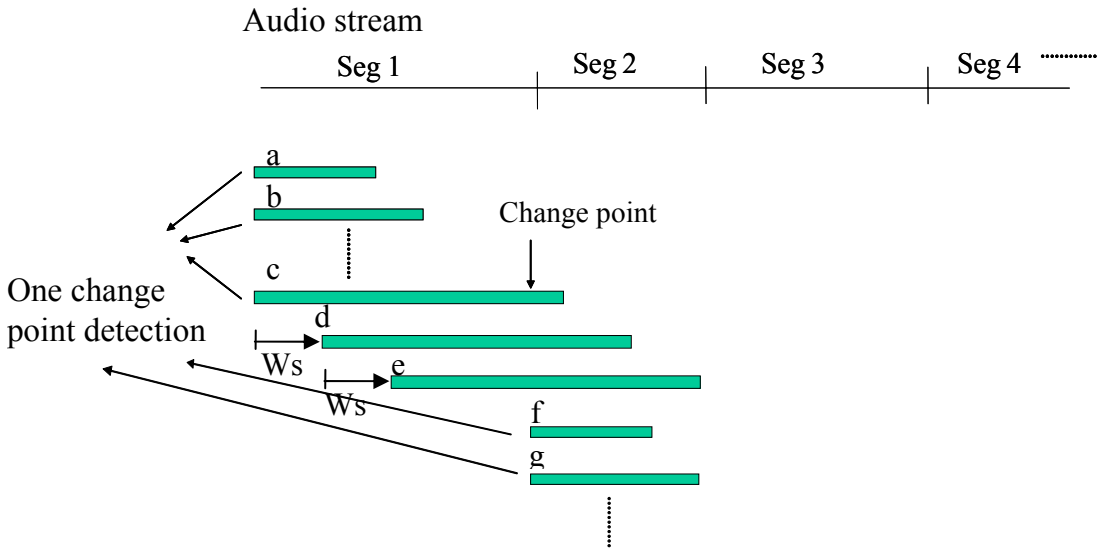


圖十：MIT Lincoln Laboratory 之 Speaker Diarization 系統架構

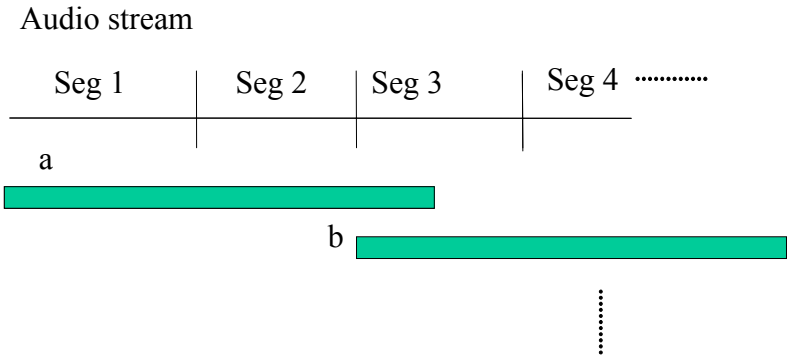
本研究團隊之研究情況

在「音訊分段」研究上[Cheng2003] [Cheng2004] [Cheng2008]，我們考慮到以模型爲基礎的方法有其實用限制，因爲實際應用中通常無法事先確知輸入的音訊串流中可能包含的聲音類別有哪些；而以信號特徵爲基礎的方法，目前尚不容易針對其發展出有效之聲學特性變換點偵測演算法。所以，本研究團隊致力於改良以距離爲基礎的音訊分段技術。回顧圖六之以距離爲基礎（Metric-based）的基本方法[Siegler1997]，此法計算兩個固定長度的滑動音窗(Sliding Window)的距離，產生一距離曲線(Distance Curve)，高於所設定之閥值(Threshold)的峰點(Peak)將被設定爲變換點。此法的優點是速度快，缺點是正確率不高。因爲音訊串流中同質音段長短不一，爲了能夠將短音段對應的變換點也偵測出來，此法所設定的音窗長度通常不大於 2 秒，以致於一個特定時間點僅由少量的音訊資料來判斷其是否爲變換點。圖十一是 IBM 團隊提出的一種由下而上(Bottom-Up)的成長音窗聲學特性變換點偵測法[Chen1998a]。與圖六方法相較，這個方法用了更多的音訊資料，或是進行多次評估，來判斷某一特定時間點是否爲變換點，正確率因此較高。如圖十一所示，它先試著在短音窗內（音窗 a）偵測轉換點，若偵測不到，則放大音窗（音框 b），直到找到轉換點或是音窗寬度已達到預設最大值爲止。圖十一的例子中，音窗 c 寬度已達上限，雖然其中包含一個轉換點，但並未被偵測出來，所以將目前音窗的起點往前移動 Ws，重新偵測，如果還是沒有偵測到轉換點，則再往前移動 Ws，結果在音窗 e 中偵測到轉換點，所以將小音窗起點移至偵測到的轉換點（音窗 f），重新啓動音窗成長偵測。

因爲這個方法需不斷放大偵測音窗，而造成高計算成本。爲了提升運算效率，我們發展圖十二的由上而下(Top-Down)的偵測演算法[Cheng2008]，以減少計算成本。基本想法是設定大音窗（音窗 a），在大音窗內以各個擊破（Divide-and-Conquer）的方式依序找出最可能的轉換點，然後將大音窗移至最後面的轉換點（音窗 b），繼續偵測。配合更具鑑別性的兩音段間距離量測，我們希望開發一高效率與高正確率之音訊分段技術。



圖十一：由下而上(Button-Up)音窗成長聲學特性變換點偵測法

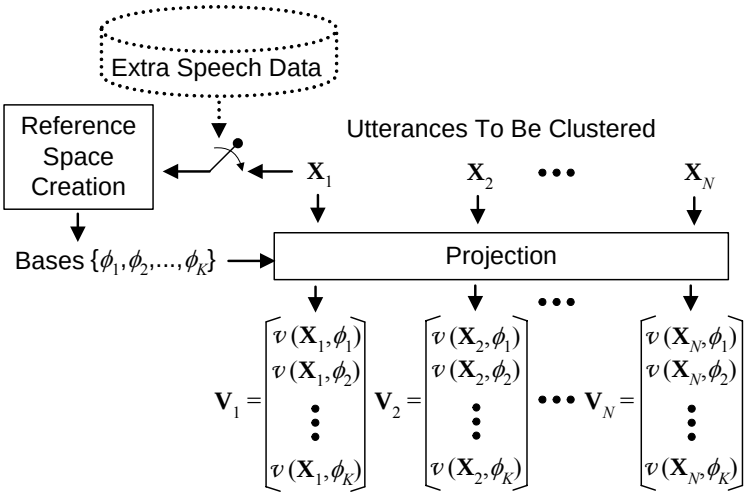


圖十二：由上而下(Top-Down) 聲學特性變換點偵測法

在「語者分群」方面，首先，我們考慮到傳統上語音片段間之相似度量測大多採頻譜特徵之直接比較，例如計算 KL 距離(Kullback Leibler Distance)或泛似然率比(Generalized Likelihood Ratio)，這種作法往往不能有效地反應語音片段是否屬於相同語者。主要原因是語音頻譜的差異一般不僅表示語者的差異，更反應其音韻訊息上的差異，甚至是錄音環境與傳輸通道上的特性差異，因此所獲得之群集未必代表語者。有鑑於此，我們提出一種語者聲音特徵空間的建構方法[Tsai2004][Tsai2005a][Tsai2007a]，利用最大

事後機率(Maximum A Posteriori)估計之語者模型訓練方法以及主成分分析(Principal Component Analysis)技術，將各語音片段的頻譜參數投影至一種泛語者聲音特徵空間中，所獲得的投影座標值可有效地反應出各語音片段所屬語者的差異，使分群結果不致成爲其他聲音的種類群。

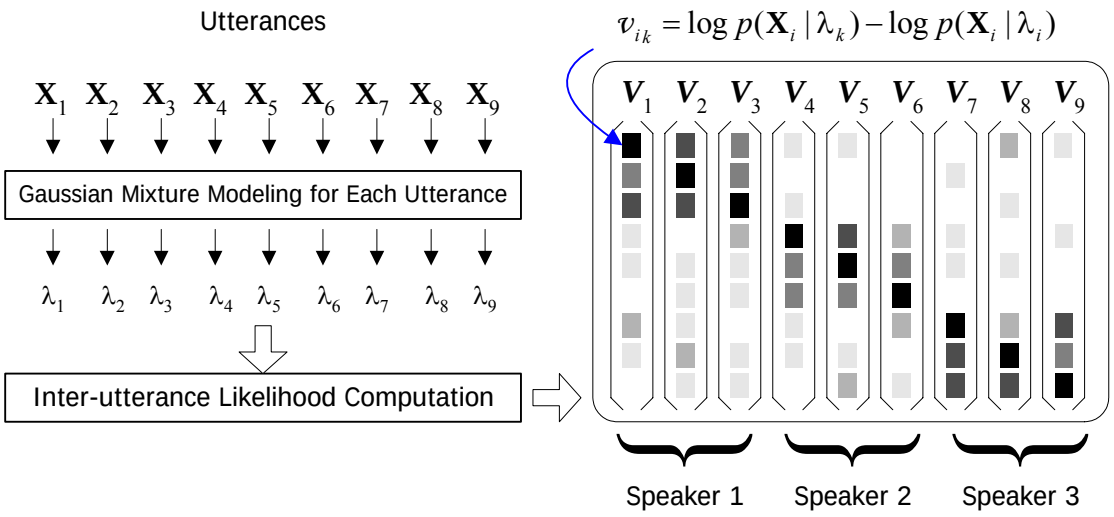
如圖十三所示，假設有 N 個欲進行分群之語音片段 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$，我們利用這些語音片段來產生一個泛語者聲音特徵空間，或透過一個外部資料庫來建立特徵空間，而這個參特徵空間是由 K 個基底(bases)所組成，每一個基底表示一種語音特性。然後每一個語音片段對每一個基底進行投影，因此語音片段就從原先的頻譜特徵向量序列(例如 MFCCs)變成一個 K 維度的投影向量 $\mathbf{V}$。於是，語音片段間彼此的相似性就可以用向量間的「Cosine Measure」來表示，即 $S(\mathbf{X}_i, \mathbf{X}_j) = \dfrac{\mathbf{V}_i \cdot \mathbf{V}_j}{\| \mathbf{V}_i \| \| \mathbf{V}_j \|}$，其值越大代表彼此越像。



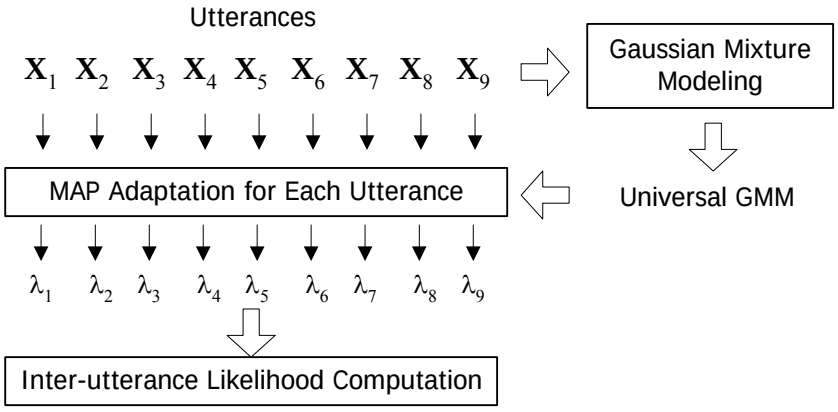
圖十三：以泛語者聲音特徵空間表示語音片段

而欲產生上述之泛語者聲音特徵空間，我們發展了如圖十四所示之架構。首先，將每一語音片段分別表示爲一高斯混合模型，因此若有九個語音片段，則我們產生九個高斯混合模型，視之爲九個基底。接著，將每一語音片段分別送入九個模型計算似然率，則所得之九個似然率可以串成一個向量。由於語音片段若與模型屬於相同語者，所獲得之似然率一般較所屬不同語者之情況來得高，因此圖所舉例之語音片段中，其中 $\mathbf{X}_1$、$\mathbf{X}_2$ 與 $\mathbf{X}_3$ 屬相同語者，我們可看到 $\mathbf{V}_1$、$\mathbf{V}_2$ 與 $\mathbf{V}_3$ 具有很高的相似性，即向量中具有較大值(顏色深者)的元素位在相似的位置。但由於語音片段可能長度很短，使用少量資料所產生的高斯混合模型可能甚不可靠，因而將造成不良的特徵空間。爲了解決此問題，我們進一步發展如圖十五所示之架構，其中每一高斯混合模型並非利用 Expectation Maximum (EM)

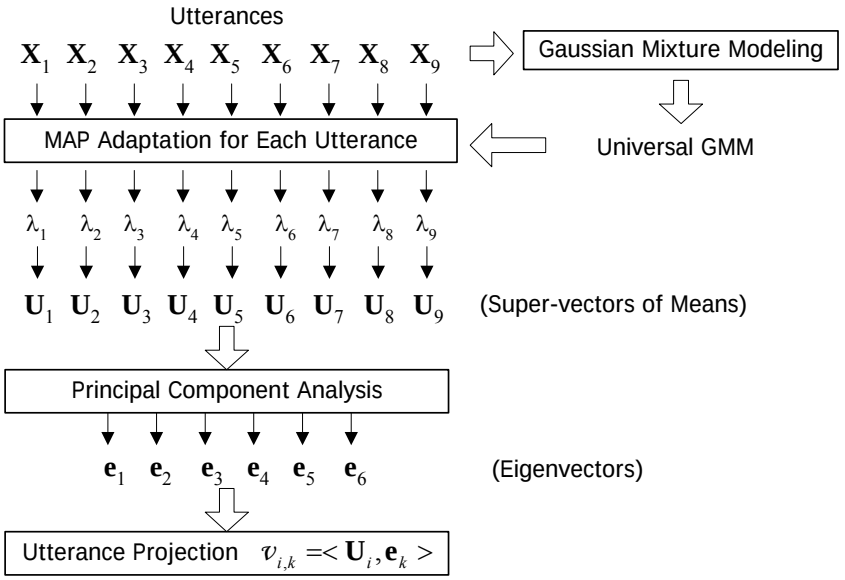
演算法所求出，而是透過一個通用模型(Universal Model)的參數調適(Adaptation)所獲得，此通用模型是將所有的語音片段合併訓練而得出。參數調適的方法爲最大事後機率估計。這種利用參數調適來提昇模型可靠度的方法是源自於語音辨認中的語者調適技術，其對於訓練資料不足的情況下效果特別顯著。然而，若考慮特徵空間應具備正交基底的本質，上述建立特徵空間的方法明顯非最佳，因爲基底之間的特性重疊非常嚴重。爲求更進一步改善分群效能，我們接著提出圖十六之以 Eigenvoice [Kuhn2000]爲基礎的架構。首先，對每一語音片段產生一高斯混合模型。接著將各模型中之所有混合(Mixture Component)的平均植向量(Mean Vectors)取出並串成一長向量 **U**，再透過主成分分析法求出所有長向量的基底。由於本徵值(Eigenvalue)較小的基底代表其重要性較低，因此我們先捨棄部分本徵值較小的基底，再計算出各長向量在各剩餘基底上的投影量。



圖十四：以高斯混合模型為基礎之泛語者聲音特徵空間表示



圖十五：以 GMM-MAP 為基礎之泛語者聲音特徵空間表示

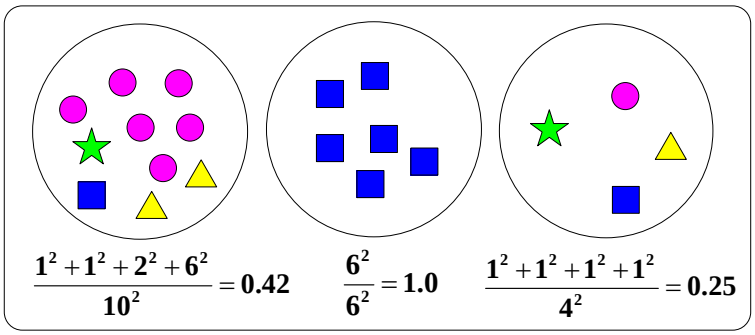


圖十六：以 Eigenvoice 為基礎之泛語者聲音特徵空間表示

另一方面，我們也探討最佳化的群集產生方式，以改進普遍所使用之階層式分群法的缺點。首先，由階層式分群法的逐群進行(Cluster-By-Cluster)方式可發現，該分群法雖盡力使任何新產生群集之同質性(Homogeneity)達到最高，但卻無法保證累加所有群集之同質性可達最高，其原因是該方法並未考慮新產生群集與原存在群集之間的關係。因此，當某次合併過程或分離過程中，若發生某些不同語者的語音片段被誤置於同一群或某些相同語者的語音片段被誤置於不同群時，這種錯誤在之後的合併或分離過程中將持續地蔓延，造成整體系統效能不佳。有鑑於此，我們提出一種「最高純度分群法」(Maximum Purity Clustering, MPC) [Tsai2005b] [Tsai2007a]，同時考慮所有群集的內部同質性，以「純度」表示，並求取可使所有群集純度達最高的最佳解。該最佳解是從各種語音片段與群集的組合方式中找出，亦即考慮每一語音片段應歸屬哪一群集才能使整體群集純度達最高。其中，「純度」的定義是：從任一群集中取出一個語音片段，取出後放回，連續取兩次，兩次取出之語音片段屬相同語者的機率，以數學方式表示爲：

$$\rho_m = \frac{1}{n_{m*}^2} \sum_{p=1}^{P} n_{mp}^2,$$

這裡ρ_m爲第m群的純度，n_{m*}爲第m群的語音片段個數，n_{mp}爲第m群中屬於第p位語者的語音片段個數，P爲語者總數。圖十七爲一個計算實例，在第一群中屬於圓狀者有六個、屬於星狀者有一個、屬於方形者有一個、屬於三角形者有兩個，因此純度爲$(6^2+1^2+1^2+2^2)/10^2 = 0.42$；在第二群中，所有元素皆屬於方形，因此純度爲$6^2/6^2 = 1$；在第三群中，所有元素皆爲不同的屬性，因此純度甚低。我們利用語音片段間的相似性來估算群集中的純度，同時應用基因演算法(Genetic Algorithm)來決定每一語音片段應歸屬哪一群集才能使整體群集純度達最高。

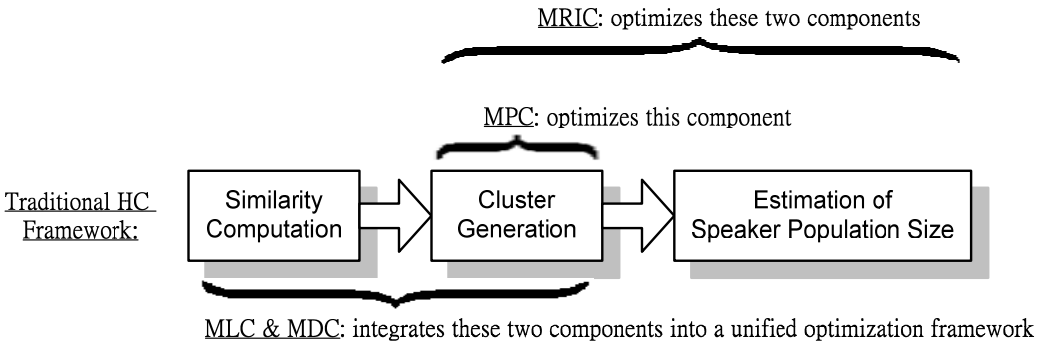


圖十七：分群純度計算實例

除了最高純度分群法以外，我們更分別提出了「最大似然率分群法」(Maximum Likelihood Clustering, MLC)與「最小歧異分群法」(Minimum Divergence Clustering, MDC) [Tsai2006]來達到所有群集的內部同質性最高的目標。而這兩種方法與前述之最高純度分群法的差異在於它們將「相似性量測」與「群集產生」整合成單一的最佳化過程。如圖十八所示爲相關方法之運作比較。傳統上語者分群的做法包括三個獨立單元，分別是「相似性量測」、「群集產生」、與「語者數目估計」。「最大似然率分群法」與「最小歧異分群法」同時最佳化前兩項單元。而另外有一項方法，稱爲「最小芮氏指標分群法」(Minimum Rand Index Clustering, MRIC) [Tsai2007b]則同時最佳化後兩項單元。所謂芮氏指標(Rand index)，是一種量測分群錯誤程度的指標，考慮同一群集中屬於不同語者的語音片段數目，及同一語者被分到不同群集的語音片段數目，數學定義爲

$$R(M)=\sum_{m=1}^{M}n_{m*}^2+\sum_{p=1}^{P}n_{*p}^2-2\sum_{m=1}^{M}\sum_{p=1}^{P}n_{mp}^2,$$

這裡n_{*p}爲第p位語者的語音片段個數，而$R(M)$是在分群群數爲M之情況下的錯誤程度。直覺上，芮氏指標與純度有著異曲同工之處，前者考慮分群的錯誤程度，而後者考慮分群的正確程度。但由於純度易受到群集內的語音片段數目的多寡而造成數值上的偏頗，其並不適合用於比較具有不同群數之分群結果。例如，有一分群系統產生群數爲10，另一分群系統產生群數爲20，則一般情況下後者所獲得的純度較前者爲高，最極端的例子是將每一語音片段視爲一群，則其純度將爲1，但明顯可知此數值爲假象，分群結果並非完美。相對地，芮氏指標則具有比較不同群數之分群結果的能力。主要是因爲該指標的值並不會因群數增加而不斷減少，而是在群數等於語者數目的情況下達到最小，當群數與語者數目相差愈多時，指標值將愈大。因此，我們可以利用這項指標來估算最佳的分群數。同時，我們可從各種語音片段與群集的組合方式中找出使芮氏指標值達最小的最佳解，亦即考慮每一語音片段應歸屬哪一群集才能使芮氏指標值達最小。此最佳解求取方式亦使用基因演算法來實現。



圖十八：相關分群方法之策略比較

結語

隨著網際網路與多媒體應用概念的發展，語音辨認研究已從過去的發展聽寫機與聲紋辨識器演變爲現今以資訊或資料檢索爲主的課題。Speaker Diarization 是一項因應資料檢索而生的新興研究題目，目前相關研究文獻皆是來自少數幾個研究群，例如 MIT Lincoln Laboratory、CMU、Cambridge University、LIMSI 等幾乎已主導了該項研究。然而，由以上所介紹之各項方法可知，目前這些研究群所普遍採用的各項 Speaker Diarization 元件皆非最佳，因而我們嘗試由最佳化的觀點來提升效能。未來仍值得努力的目標是將「音訊分段」與「語者分群」進行最佳化整合，以求達到最低的 Diarization Error Rate。

參考文獻

[Bakis1997] R. Bakis et al, “Transcription of broadcast news shows with the IBM large vocabulary speech recognition system,” in Proc. *DARPA Speech Recognition Workshop*, 1997.

[Barras2004] C. Barras, X. Zhu, S. Meignier, and J. L. Gauvain, “Improving speaker diarization,” in Proc. *DARPA RT04*, Palisades NY, November 2004.

[Bonastre2000] J. F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, and C. Wellekens, “A speaker tracking system based on speaker turn detection for NIST evaluation,” in Proc. *ICASSP2000*.

[Canseco-Rodriguez2004] L. Canseco-Rodriguez, J. L. Gauvain, and L. Lamel, “Towards using STT for broadcast news speaker diarization,” in Proc. *DARPA RT04 workshop*, Palisades, NY, November 2004.

[Cettolo2003] M. Cettolo and M. Vescovi, “Efficient audio segmentation algorithms based on the BIC,” in Proc. *ICASSP2003*.

[Chen1998a] S. Chen and P. Gopalakrishnam, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion,” in Proc. *DARPA Broadcast News Workshop*, 1998.

[Chen1998b] S. Chen et al., “IBM's LVCSR system for transcription of broadcast news used in the 1997 HUB4 English evaluation,” in Proc. *DARPA Broadcast News Transcription and Understanding Workshop*, 1998

[Cheng2003] S. S. Cheng and H. M. Wang, “A Sequential Metric-based Audio Segmentation Method via The Bayesian Information Criterion,” in Proc. *Eurospeech*, 2003.

[Cheng2004] S. S. Cheng and H. M. Wang, “METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation,” in Proc. *ICSLP*, 2004.

[Cheng2008] S. S. Cheng, H. M. Wang, H. C. Fu, “Bic-based Audio Segmentation by Divide-and-Conquer,” in Proc. *ICASSP*, 2008.

[Delacourt2000] P. Delacourt, C. J. Wellekens, “DISTBIC: a speaker-based segmentation for audio data indexing,” *Speech Communication*, v.32, pp. 111-126, 2000.

[Gish1991] H. Gish, M. H. Siu and R. Rohlicek, “Segregation of speakers for speech recognition and speaker identification,” in Proc. *ICASSP1991*.

[Jin1997] H. Jin, F. Kubala, and R. Schwartz, “Automatic speaker clustering,” In: Proc. *DARPA Speech Recognition Workshop*, 1997.

[Kuhn2000] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, “Rapid speaker adaptation in eigenvoice space,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 695-707, 2000.

[Lapidot2002] I. Lapidot, H. Guterman, and A. Cohen, “Unsupervised speaker recognition based on competition between self-organizing maps,” *IEEE Trans. on Neural Network*, pp. 877-887, July 2002.

[Liu2005] Z. Liu, “An efficient algorithm for clustering short spoken utterances,” in Proc. *ICASSP*, 2005.

[Lu2002a] L. Lu and H. J. Zhang, “Speaker change detection and tracking in real-time news broadcasting analysis,” in Proc. *ACM Multimedia2002*.

[Lu2002b] L. Lu, H. J. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 7, pp. 504-516, 2002.

[Makhoul2000] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, “Speech and language technologies for audio indexing and retrieval,” in Proc. *IEEE*, vol. 88, no. 8, pp. 1338-1353. 2000.

[Moh 2003] Y. Moh, P. Nguyen, and J. C. Junqua, “Towards domain independent speaker clustering,” in Proc. *ICASSP*, 2003.

[Reynolds 1998] D. A. Reynolds, E. Singer, B. A. Carson, G. C. O’Leary, J. J. McLaughlin, and M. A. Zissman, “Blind clustering of speech utterances based on speaker and language characteristics,” In Proc. *ICSLP*, 1998.

[Reynolds2004] D. A. Reynolds and P. Torres-Carrasquillo, “The MIT Lincoln Laboratory RT-04F diarization systems: applications to broadcast audio and telephone conversations,” in Proc. *DARPA EARS RT-04F Worksho*p, White Plaines, NY, Nov 2004.

[Siegler1997] M. Siegler, U. Jain, B. Raj and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in Proc. *DARPA Speech Recognition Workshop*, 1997.

[Solomonoff 1998] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," In: Proc. *ICASSP*, 1998.

[Sturim2001] D. E. Sturim, D. A. Reynolds, E. Singer, and J. P. Campbell, "Speaker indexing in large audio databases using anchor models," in Proc. *ICASSP2001*.

[Tranter2006] S. Tranter and D. A. Reynolds, "An overview of automatic speaker diarisation systems," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557-1565, Sept 2006.

[Tsai2004] W. H. Tsai, S. S. Cheng and H. M. Wang, "Speaker clustering of speech utterances using a voice characteristic reference space," in Proc. *ICSLP*, 2004.

[Tsai2005a] W. H. Tsai, S. S. Cheng, Y. H. Chao, and H. M. Wang, "Clustering speech utterances by speaker using eigenvoice-motivated vector space models," in *Proc. ICASSP,* Philadelphia, USA, March 2005.

[Tsai2005b] W. H. Tsai and H. M. Wang, "Speaker clustering of unknown utterances based on maximum purity estimation," in Proc. *Interspeech-Eurospeech*, 2005.

[Tsai2006] W. H. Tsai and H. M. Wang, "Speech utterance clustering based on the maximization of within-cluster homogeneity of speaker voice characteristics," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1631-1645, September 2006.

[Tsai2007a] W. H. Tsai, S. S. Cheng, and H. M. Wang, "Automatic speaker clustering using a voice characteristic reference space and maximum purity estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1461-1474, May 2007.

[Tsai2007b] W. H. Tsai and H. M. Wang, "Speaker Clustering Based on Minimum Rand Index," in Proc. *ICASSP*, 2007.

[Wang2004] H. M. Wang, S. S. Cheng, and Y. C. Chen, "The SoVideo Mandarin Chinese broadcast news retrieval system," *International Journal of Speech Technology*, 3(2), pp. 128-145, June 2004.

[Zhang2001] T. Zhang, C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 441 - 457, 2001.

[Zhou2000] B. W. Zhou and J. H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion," in Proc. *ICSLP2000*.

[Zhu2005] X. Zhu, C. Barras, S. Meignier, and J. L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in Proc. *Interspeech-Eurospeech*, 2005.