中華民國計算語言學學會

通訊 ACLCLP

發行人：吳宗憲
主　編：王新民
執行編輯：黃　琪

The Association for Computational Linguistics and Chinese Language Processing 第二十二卷第一期

E-Mail：aclclp@hp.iis.sinica.edu.tw
地 址：台北市研究院路二段128號中研院資訊所
Website：http://www.aclclp.org.tw
電 話：(02)2788-3799 ext:1502
傳 眞：(02)2788-1638
劃 撥：19166251

本 期 要 目

獎助學生出席國際會議

獎助會議：

1. COLING　2. ACL
3. ACM SIGIR　4. ICASSP

獎助說明：

1. 申請人須同時具備下列資格：
   (1) 被接受論文之第一作者(指導教授不計)。
   (2) 本會會員。
   (3) 投稿時爲國內在學學生。
2. 獎助金額：由審查委員會依地區別及論文等級審定獎助金額，每名獎助金額上限爲美金 1,000 元。
3. 獎助名額：每個會議獎助一～二名。

申請辦法：

1. 申請期限：論文被接受發佈日起兩週內提出。
2. 申請手續：申請人需將論文接受函、審查意見、學生證、論文全文及申請書等相關資料郵寄至本會秘書處。(申請書請至 http://www.aclclp.org.tw/doc/fundreg.htm 下載)

受獎助人義務：

1. 出席會議發表論文。
2. 論文全文必須以書面同意投稿至本會期刊。
3. 代學會攜去宣傳品及帶回相關資料。

Oriental COCOSDA 2011 Call for Papers

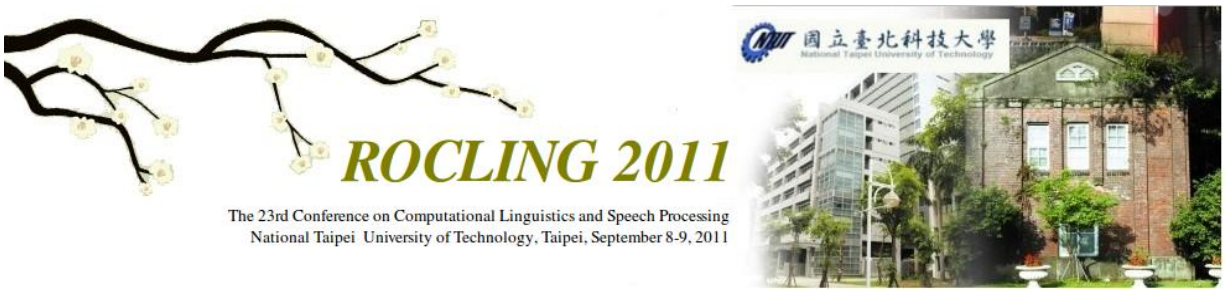
The oriental chapter of COCOSDA (The International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques) is pleased to announce that the 14th Oriental COCOSDA Conference will be held on Oct. 26-28, in Hsinchu, Taiwan hosted by the National Chiao Tung University, Taiwan. Oriental COCOSDA is an international conference held annually by the oriental chapter of COCOSDA. The first preparatory meeting was held in Hong Kong in 1997 and then the past thirteen workshops were held in Japan, Taiwan, China, Korea, Thailand, Singapore, India, Indonesia, Malaysia, Vietnam, Japan, China and Nepal.

Oriental COCOSDA conference in Taiwan will help in boosting the research and development in the field of Speech Technology and will help in enthusing the interest towards Speech Technology in East and Southeast Asia.

Important Dates

Full Paper Submission	July, 1, 2011
Notification of Acceptance	Aug. 5, 2011
Final Manuscript	Aug. 26, 2011
Conference	Oct. 26-28, 2011

Conference Website
http://ococosda2011.cm.nctu.edu.tw/



The 23rd Conference on Computational Linguistics and Speech Processing

http://sites.google.com/site/rocling2011/

Call for Papers

The 23rd ROCLING Conference will be held at 科技大樓國際會議廳
National Taipei University of Technology, Taipei, on September 8-9, 2011

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 XXIII will be hosted by the Department of Electronic Engineering, National Taipei University of Technology. The two-day conference will feature invited talks, paper, and poster sessions. ROCLING XXIII 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.

Topics of Interest:

- cognitive/psychological linguistics
- discourse/dialogue modeling
- information extraction/text mining
- information retrieval
- language understanding/generation
- lexicon/morphology
- machine translation/multilingual processing
- named entity recognition
- NLP applications/tools/resources
- phonetics/phonology
- question answering
- semantic web
- semantics/pragmatics
- speech analysis/synthesis
- speech recognition/understanding
- spoken dialog systems
- spoken language processing
- syntax/parsing
- text/speech summarization
- web knowledge discovery
- word segmentation/POS tagging
- others

Important Dates:

Preliminary paper submission deadline:	July 2, 2011
Notification of acceptance:	August 6, 2011
Camera-ready due:	August 15, 2011
Conference date:	September 8-9, 2011

Submission Guidelines:

Prospective authors are invited to submit full papers of no more than 15 A4-sized pages, single-spaced, in PDF format. 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 on-line at https://www.easychair.org/account/signin.cgi?conf=rocling2011

Contact:

For any submission and publication problems, please contact Prof. Wei-Ho Tsai, whtsai@ntut.edu.tw

「中央研究院口語韻律語料庫暨工具平台」開放申請

「中央研究院口語韻律語料庫暨工具平台」(Sinica Continuous Speech Prosody Corpora & Toolkit，簡稱 COSPRO & Toolkit)，係中研院語言所鄭秋豫教授多年從事語流韻律研究所收集的國語連續語流語料及依研究需要所發展的工具平台(1994-2005)。基於學術資源共享之理念與促進語音科學研究與技術能有突破性發展之初衷，於 2006 年即釋出本語料庫與工具平台，原由民間公司—艾爾科技公司(L Labs Inc.)發行，現基於語料管理與學術能見度考量，於今(2011)年 2 月重新授權予學會發行，供國內外學術或民間機構非營利使用。

COSPRO 包含九個子語料庫，每個子語料庫都針對不同的語流韻律現象設計而成，COSPRO 01-08 屬於麥克風朗讀語音，COSPRO 09 則爲麥克風自發性語音(76MB)。內容包羅了不同長度的語料，短至孤立詞組(1-4 字詞)，長至段落語篇(85-996 音節)。共包括 114 位語者(61 位女性，53 位男性)，可供語音研究、語音合成與語者辨識等多方面應用。

釋出語料庫容量共 10.5GB，其中大部分語料(7.7 GB)已經過處理，並附說明，而非僅是原始音檔。每個子語料庫，除了 wav 檔案之外，還包括每位語者的朗讀(轉寫)文本(*.txt)、人工調整音標檔(*.adjusted / *.syl)以及停延韻律標記檔(*.break)。其餘未經處理之原始語料，則僅釋出 wav 檔案、語者的朗讀(轉寫)文本(*.txt)以及程式處理過後的音標檔(*.phn)。

本語料庫與其他語音資料庫最大的差異在於：包含(1)**人工調整音標檔(*.adjusted / *.syl)**：不只是 HTK 處理過的音段標註檔案(*.phn)，絕大多數釋出語料的人工調整音標檔，均以人工方式對齊語音音段邊界，標註子音與母音的時間碼。(2)**停延韻律標記檔(*.break)：**訓練有素的標音員依感知所獲得的韻律標註檔案，是以聽感爲基礎，並通過標註一致性檢驗。人工感知韻律標註的主要意義在於，以本語料庫所提供的韻律標記做爲語音信號分析的標準答案，而非得自文本分析結果，是符合語音事實的韻律單位，目的是凸顯語音與文本並不完全比配的事實。

COSPRO Toolkit 係一視窗介面(Window-based)，好操作(user-friendly)的語音分析暨合成之工具平台，集合了 Adobe® Audition®, Praat© and Speech Viewer®等常見語音分析(合成)軟體之特點，其主要功能有：聲學訊號分析功能、標記口語語流功能以及重新合成語音訊號(re-synthesizing speech signals)功能，特別適合作爲教學工具。

進一步申請資訊請參閱學會網頁：http://www.aclclp.org.tw/corp_c.php。

同時說話聲之內含語者的辨識問題探討

A Preliminary Study of Speaker Recognition in Overlapping Speech

蔡偉和

國立台北科技大學電子工程系暨電腦與通訊研究所

whtsai@ntut.edu.tw

1. 前言

自動語者辨識(Automatic Speaker Recognition)研究[8,16]在「語音處理」(Speech Processing)與「生物測定」(Biometrics)領域中已有數十年歷史，其目的在於判斷一段語音是由誰所說(稱爲「語者識別」，Speaker Identification，SID)或是否爲某人所說(稱爲「語者確認」，Speaker Verification)。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)，評比項目也隨著技術的發展而不斷地更新，但其中「語者辨識」評比自 1996 年起迄今仍持續進行，顯示這項研究議題的重要性與可發展性。近幾年的語者辨識評比著重於對話語音中的語者偵測判斷(Speaker Detection)，並與另一項評比「Rich Transcription」中的子項目相結合成一個特別的研究議題，稱爲「語者分段標記」（Speaker Diarization），又稱「Who Spoke When」[10]，其目標是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，並給予一個說話者識別身分。

然而，儘管語者辨識技術已不斷地提升，且所處理的語音資料也愈趨多元，但目前大多數研究所探討的問題都僅止於單語者的辨識問題，即假設同一個時間下只有一位說話者，這並無法完全符合真實情況，因爲許多場合可能存在著多位語者同時說話的情形。研究統計[1]，在一般兩人的對話中，有超過 10%的機率發生同時說話的情形。而同時說話的情形在小型會議中更爲普遍。筆者將多人同時說話的語音稱爲「重疊語音」(Overlapping Speech)，而重疊語音所包含的語者稱爲「同時語者」(Simultaneous Speakers)。現階段僅有少數的研究討論重疊語音的問題，且多著重於語音辨認用途上。在[1,5]中曾分析自動語音辨認系統在重疊語音區段的辨認錯誤情形。研究結果發現目前的自動語音辨認系統對重疊語音區段的辨認能力十分薄弱，但該文獻中仍未提出可行的解決之道。在[6,7]中曾探討如何自動偵測會議對話語料中的重疊語音區段，該論文的訴求是如果能有效地找出對話語料中的重疊語音區段，則可將這些區段排除於語音辨認之外，避免影響整體的語音辨認結果。

截至目前，在重疊語音中辨識語者的研究更是缺乏，唯一文獻是筆者實驗室所發表的一項初步研究[17]。雖然辨識同時語者的困難性仍高，但相較於辨識重疊語音的內容

而言，應該較爲容易可行。本文僅針對筆者在這項研究的初步結果進行說明，期望藉此提供感興趣的讀者些許靈感，在相關的研究上有所助益。

直覺上，若能將重疊語音中各語者的聲音訊號分離開來，則辨識同時說話者的問題就可拆解爲兩次傳統的單語者辨識。然而分離重疊語音是件極爲困難的事，即使人耳能夠分辨混合音中各種不同聲音來源，但也未必具備分離重疊語音訊號的能力。現階段有許多盲訊號分離(Blind Signal Separation)的研究[18-21]，但大多數是在多通道訊號的前提下探討，例如多支麥克風錄音或立體聲音訊。少數的研究也針對單聲道(Monaural)混合音分離進行探討，其中較具代表性者爲計算聽覺場景分析(Computational Auditory Scene Analysis；CASA)研究[22-27]。在若干文獻中，CASA 成功地用於改善人聲吵雜環境下的語音辨識效能。它的原理是利用 Gammatone Filter banks 將混合音訊號拆解爲時間與頻率(Time-Frequency；TF)上小單元的能量強度，並判斷每一個小單元應歸屬於哪一個訊號源，因此各訊號源在某些小單元上被判定爲不具有能量，而在某些小單元上則被判定爲佔有所有能量，最後，再依各訊號源所屬的各小單元上能量合成訊號。如圖 1 所示爲文獻[22]所舉的一個例子。圖 1(a)爲一男性語者的說話片段，內容爲「Primitive tribes have an upbeat attitude」，經由 TF 分析後以灰階表示，白色代表能量較強，而黑色代表能量較弱。圖 1 (b)爲一女性語者的說話片段，內容爲「Only the best players enjoy popularity」。圖 1 (c)是將這兩位語者的說話聲以 1:1 能量混合。若在任一 TF 小單元中男性語者語音能量強於女性語者語音能量，則令一項運算元 Mask 值爲 1，代表男性語者語音在該 TF 小單元上佔有所有能量；否則 Mask 值設爲 0，代表男性語者語音在該 TF 小單元上不具有任何能量。所得 Mask 畫成圖 1 (d)，其中白色代表 1，黑色代表 0。將圖 1 (d)中每一 Mask 小單元與圖 1 (c)相乘，可獲得圖 1 (e)。若將圖 1 (e)重新合成訊號波形，則可聽到大部分聲音爲男性語者的語音。而圖 1 (d)中的 Mask 稱爲 Ideal Binary Mask，也就是當給定一個混合音時，若知道 Ideal Binary Mask，則將混合音乘上 Ideal Binary Mask 後，將可得近似的標的 (Target)音。但可想而知，近似標的音必然與真實標的音不同。而在實際的使用上，混合音所隱含的 Ideal Binary Mask 並不知道，因此需再進行估計，圖 2 所示爲另一個例子，圖 2 (a)爲一真實標的語音的 TF 表示，經疊加其他聲音後的混合音如圖 2 (b)所示，而圖 2 (c)爲 Ideal Binary Mask，圖 2 (d)是經由估算所得的 Mask，圖 2 (e)與圖 2 (f)分別是混合音乘上 Ideal Binary Mask 與所估算 Mask 的結果。文獻中指出經由 Mask 後的結果對訊號音質及在語音辨認上皆有大幅地改善。

然而，CASA 雖能達到某種程度地萃取混合音中的標的語音，但所獲得的語音仍與原始未混合前的語音相去甚遠。若就語者辨識的觀點來看，CASA 處理後雖能維持若干程度之語音內容詞的可辨性(Intelligibility)，但其語者的特徵性卻已嚴重遭到破壞，因此並不適合用於語者辨識上。有鑑於此，筆者所探討的同時語者辨識方法暫不考慮從訊號分離的觀點出發。

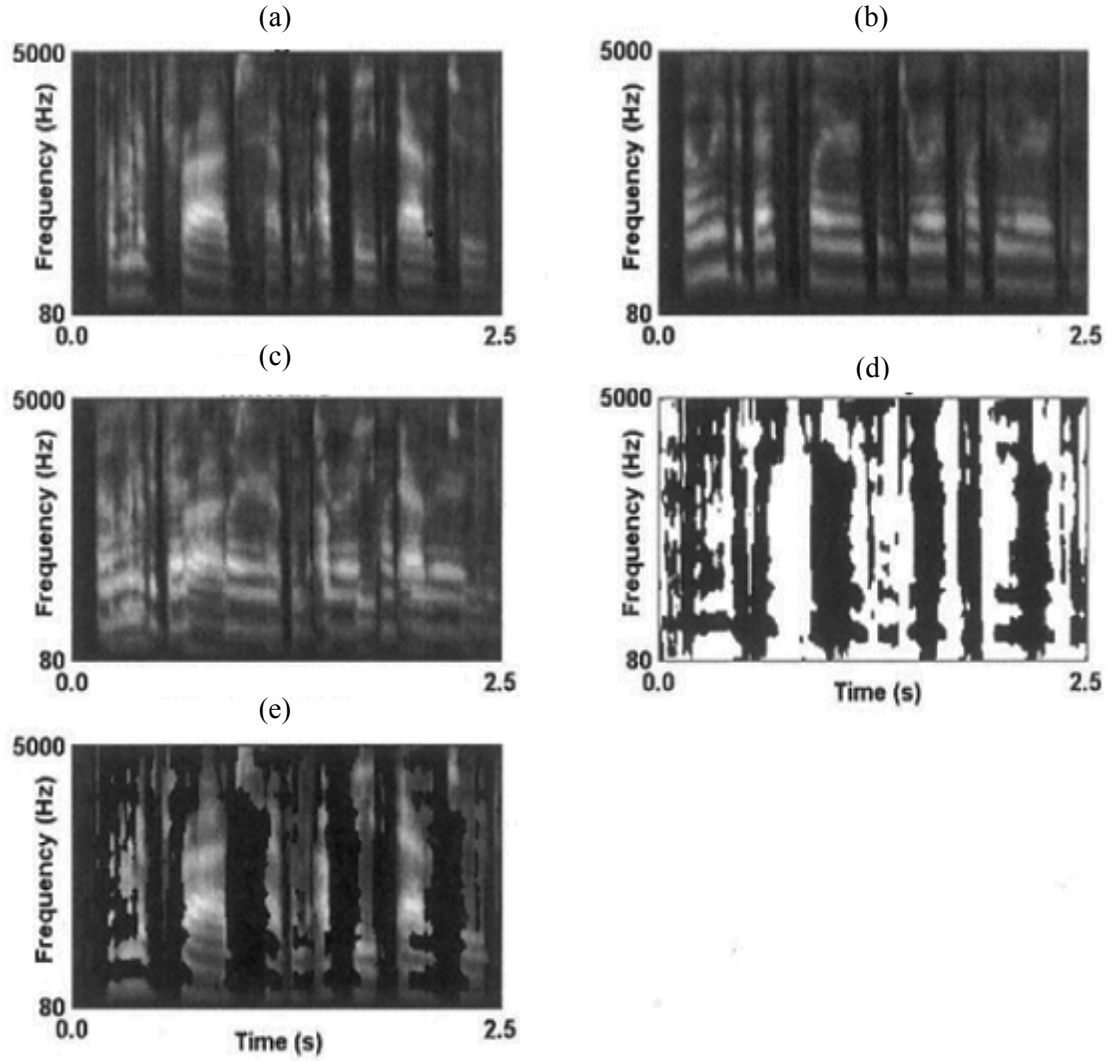


圖 1：CASA 處理實例(摘自[22])。(a)爲一男性語者的說話片段，內容爲「Primitive tribes have an upbeat attitude」，經由 TF 分析後以灰階表示，白色代表能量較強，而黑色代表能量較弱；(b)爲一女性語者的說話片段，內容爲「Only the best players enjoy popularity」；(c)是將這兩位語者的說話聲以 1:1 能量混合；(d)爲 Ideal Binary Mask，其中白色代表 1，黑色代表 0；(e)是將(d)中每一 Mask 小單元與(c)相乘，藉以獲得男性語者的說話訊號。

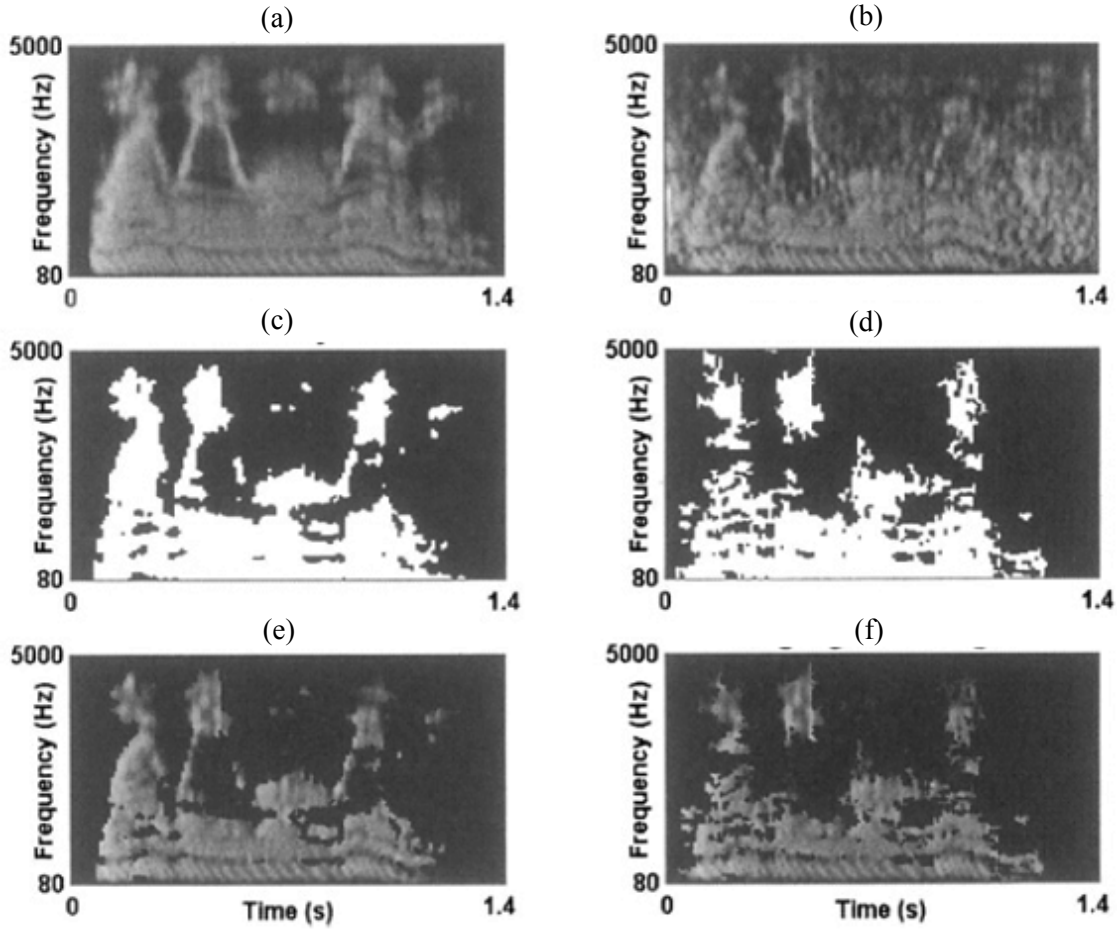


圖 2：CASA 處理實例(摘自[22])。(a)一真實標的語音的 TF 表示，(b)爲標的語音疊加其他聲音後的混合音，(c)爲 Ideal Binary Mask，(d)爲估計之 Binary Mask，(e)是混合音乘上 Ideal Binary Mask 的結果，(f)是混合音乘上估計之 Binary Mask 的結果。

2. 問題定義與分析

重疊語音中的語者辨識問題，除了可依前述觀點分爲「語者識別」(Speaker Identification)與「語者確認」(Speaker Verification)之外，其研究主題又可依下列幾項因素來定義。

- **同時語者數目：**一般，愈多人同時說話時，愈難辨識說話者爲何。在筆者的初步研究中，僅考慮重疊語音中包含兩位同時說話的語者。
- **重疊音量比例：**當發生多人同時說話時，有些人音量較大，有些人則音量較小。一般，音量較大的語者較容易辨識，音量較小者則不容易辨識。而音量大小有時與麥克風所架設的位置有關，離麥克風較近者通常呈現較大的音量。因此在重疊語音中，有些語者可能聽起來像是「背景語者」(Background Speakers)一樣，聲音較爲弱且模糊，而有些語者則具有「前景語者」(Foreground Speakers)的角色，像是主要發言人一樣。當然，辨識前景語者比較容易，而辨識背景語者則十分困難。另外，若多位語者的音量相當

時，是否能夠讓系統辨識所有語者更是值得探討的問題；因此，筆者的初步研究即是鎖定這樣的問題。

- **語音內容相同與否：**重疊語音中的語者可能說相同的話，即異口同聲，也可能說不同話，即各說各話。對人耳來說，說不同話的情形似乎較容易分辨出不同的語者，但對自動辨識系統而言可能未必如此。筆者的初步研究中對於異口同聲與各說各話皆曾進行探討。
- **開放集(Open-set)與封閉集(Close-set)問題：**對於「語者識別」而言，待測的語音可能是系統所登記(Enroll)語者群中的某一位語者所說的，也可能不是系統所登記語者群中的任何一位語者所說的。若系統僅能處理前者的問題，則其屬於封閉集(Close-set)的語者識別；若系統亦能處理後者的問題，則其屬於開放集(Open-set)的語者識別。顯然，開放集的語者識別較封閉集的語者識別困難。當考慮重疊語音下的封閉集語者識別時，同時語者將是封閉集內各語者的所有可能組合。例如封閉集共有 N 位語者，且若假設重疊語音皆只包含兩位同時說話的語者，則將形成 $C_2^N = N! / [2!(N-2)!]$種可能的識別結果。然而，當考慮重疊語音下的開放集語者識別時，待測的語音可能包含全都是系統所登記的語者；也可能包含局部屬於系統所登記的語者，以及局部不屬於系統所登記的語者；另外也可能包含全都不屬於系統所登記的語者。處理此類問題將十分困難。因此，筆者的初步研究先侷限於封閉集的語者識別。
- **聲音品質：**就像各種語音辨識問題一樣，欲辨識的重疊語音可能在不良環境下所收錄，包括麥克風品質不良、聲音經編解碼與傳輸造成失真、背景噪音干擾等，因此將使辨識問題更爲棘手。在筆者的初步研究中，暫時不考慮複雜的音質問題，僅以乾淨的麥克風語音進行研究。
- **內含語者均勻性：**待測的語音內可能 1)任何瞬間或區段皆包括相同的同時語者；2)某些瞬間或區段包括同時語者 A 與 B，而另外某些瞬間或區段包括同時語者 B 與 C，或僅包括語者 B。在 1)的情形下，辨識系統只需判斷整個待測語音的屬性爲何，但在 2)的情形下則需決定出各瞬間或區段的語音屬性爲何。筆者的初步研究考慮 2)的情形。

圖 3 所示爲本文所探討的語者辨識問題。在一段高品質麥克風所錄下之長的音訊資料中，某些區段包括單一語者的說話，某些區段包括兩位同時語者，而所有可能的語者皆爲已知，系統的目標是決定出各瞬間(或區段)語音的內含語者爲何。

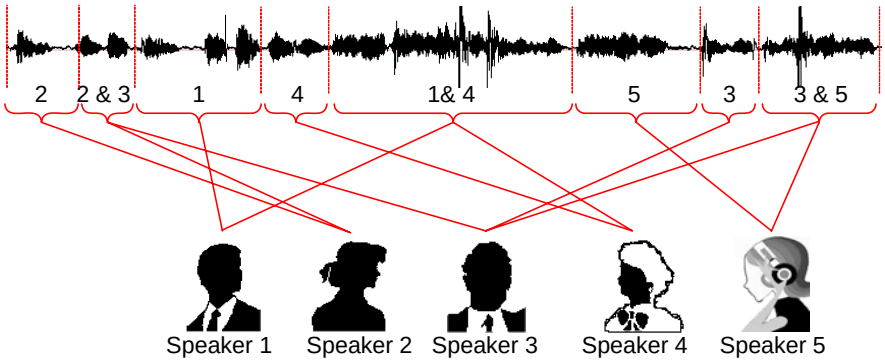


圖 3：本文所探討的語者識別問題。

3. 研究方法

圖 4 所示爲筆者所採取的語者識別(SID)基本策略。利用一個滑動視窗(Sliding Window)將長的音訊資料連續且不重複地切割爲固定長度的小區段，然後分別判斷各小區段的屬性。判斷的方法又可分成兩種：一爲「兩階段的方法」，另一爲「單一階段的方法」。

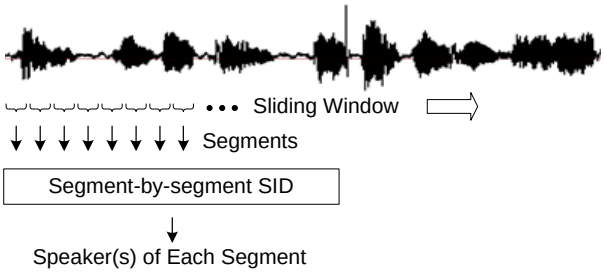


圖 4：本文所採取的語者識別基本策略。

3.1 兩階段的識別系統

如圖 5 所示，由於每一小區段的內含語者可能是單一語者，也可能是兩位同時語者，因此系統先進行一項「重疊語音偵測」(Overlapping Speech Detection)，判斷每一小區段是否爲重疊語音，若是，則執行「雙語者識別」(Two-Speaker Identification)，若否，則執行傳統的「單語者識別」(Single-Speaker Identification)。

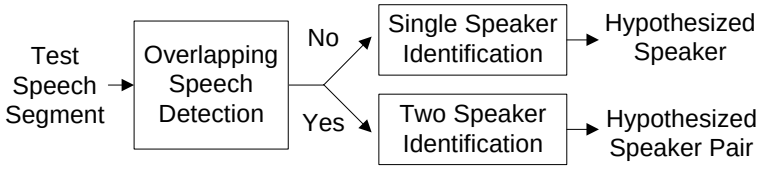


圖 5：兩階段的語者識別系統。

3.1.1 重疊語音偵測

圖 6 爲重疊語音偵測之架構圖。基本原理是利用統計分類器來區別重疊語音與非重疊語音。它包括兩個操作階段，一爲訓練、另一爲判斷。在訓練階段，我們建立兩個高斯混合模型(Gaussian Mixture Models，GMMs)，λ^n 與 λ^o，其中λ^n 用以表示非重疊語音的聲學特徵，而λ^o 則表示重疊語音的聲學特徵。高斯混合模型的參數包括期望值向量、變異矩陣、與混合權重，可藉由 expectation-maximization (EM) [14]演算法來估算。而估算時必須透過大量的資料來完成，也就是利用大量的非重疊語音來估算λ^n，且利用大量的重疊語音來估算λ^o。但由於重疊語音較非重疊語音難取得，我們可以透過波形疊加的方式來產生重疊語音。另外，在訓練高斯混合模型前，我們將語音訊號轉成 Mel-scale Frequency Cepstrum Coefficients (MFCCs)。

在測試階段，我們將一個未知待測的音訊轉成 MFCCs 後，分別送入λ^n 與λ^o進行匹配，即計算似然率(likelihood probabilities), $\Pr(\mathbf{X}|\lambda^n)$與 $\Pr(\mathbf{X}|\lambda^o)$。如果 $\log\Pr(\mathbf{X}|\lambda^o)-\log\Pr(\mathbf{X}|\lambda^n)$ 較一個預設的門檻η大，則判定該音訊爲重疊語音，否則爲非重疊語音。

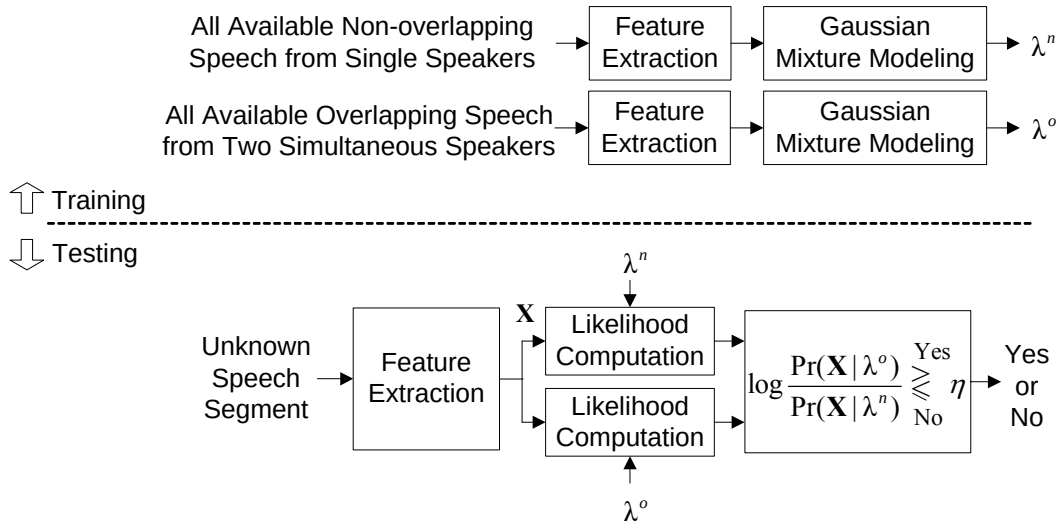


圖 6：重疊語音偵測。

3.1.2 單一語者識別

單一語者識別即傳統的語者識別問題。如圖 7 所示，假設系統需要識別 N 個語者，則在訓練階段分別利用各語者的說話資料來產生 N 個高斯混合模型$\lambda_1, \lambda_2, \ldots, \lambda_N$，代表這些語者的聲學特徵。當給定一未知受測的音訊 **X** 時，其中 **X** 爲 MFCCs，系統可根據最大似然率法則(maximum likelihood，ML)來判斷其所屬語者。

$$I^* = \arg\max_{1 \le i \le N} \Pr(\mathbf{X} \mid \lambda_i). \quad (1)$$

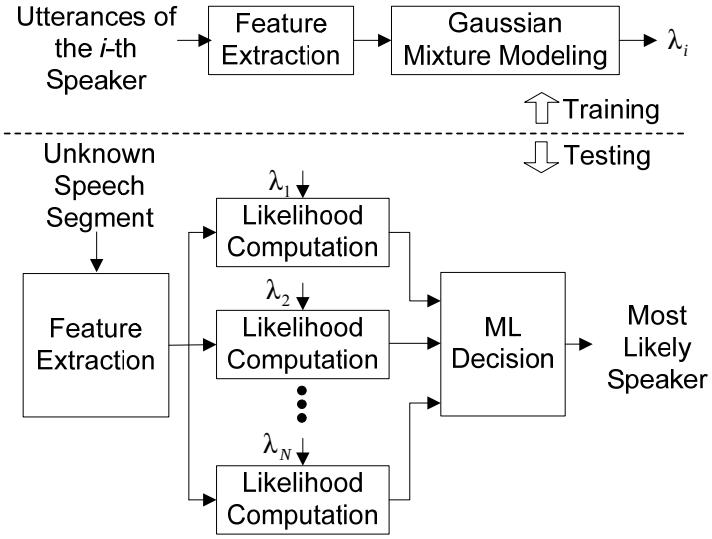


圖 7：單語者識別單元。

3.1.3 雙語者識別

雙語者識別原理類似於單語者識別，除了其中的高斯混合模型由原來所代表的單語者更改爲雙語者，即利用高斯混合模型來表示任兩位語者同時說話的聲學特徵。因此，對於語者群數爲 N 的識別系統而言，將需產生 $C_2^N = N! / [2!(N-2)!]$ 對的雙語者模型 $\lambda_{i,j}$, $i \neq j$, $1 \le i, j \le N$，亦即識別種類變爲 C_2^N 種。然而，在訓練語料的收集上若要求所有語者彼此同時說話，將太費時費力。因此我們設計兩種方式來取得雙語者的訓練語料。第一種是利用波形疊加的方式，如圖 8 所示，即直接將兩語者的任兩語音波形疊加，且兩波形能量調整爲相當，以此模擬重疊語音。當給定一未知受測的音訊 $\mathbf{X}$ 時，系統將決定出最可能的雙語者爲：

$$(I^*, J^*) = \arg\max_{1 \le i, j \le N, i \ne j} \Pr(\mathbf{X} \mid \lambda_{i,j}). \quad (2)$$

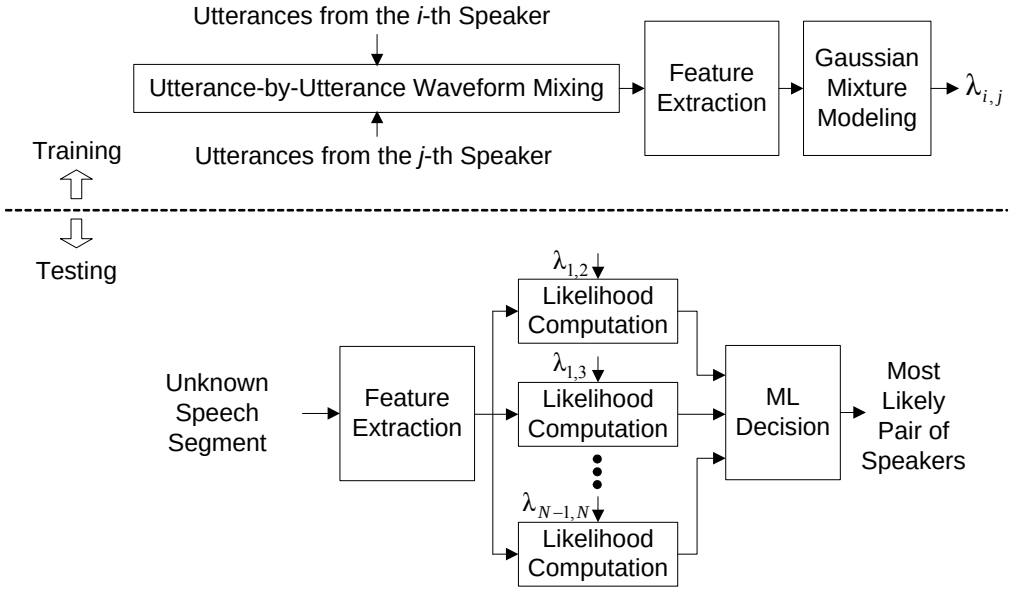


圖 8：基於直接波形疊加方式的雙語者識別系統。

然而，上述利用波形疊加來產生訓練語料的方式有一項缺點，就是當語者的數目非常龐大時，或有新的語者要加入系統時，一一疊加波形便顯得麻煩。因此我們考慮另一種訓練雙語者模型的方式，透過所謂的 Parallel Model Combination (PMC)技術[15]，如圖 9 所示。假設系統已建立了 N 個單語者高斯混合模型，則利用這些模型的兩兩組合，將可產生 C_2^N 個雙語者模型。由於重疊語音相當於兩語者說話聲在時域上進行疊加，也相當於在頻域或頻譜上疊加，但在倒頻譜上(模型是由倒頻譜特徵所產生的)則非單純的疊加，因此模型的組合並非單純的個別參數相加。我們需先將高斯混合模型的參數從倒頻譜轉至頻譜，然後才能相加。另外，考慮到兩個混和數爲 K 的 GMMs 組合將產生一個混和數爲 $K \times K$ GMM，造成運算量過於龐大，因此我們利用 UBM-MAP [16]技術來維持所組合出之 GMM 的混和數爲 K。作法是先透過所有語者的訓練語料一同訓練出一個混和數爲 K 的「通用模型」(Universal GMM)，然後再藉由個別語者的訓練語料，將通用模型調整爲各語者專屬模型(單語者模型)，調整方式爲 maximum a posterior (MAP) estimation。由於所有的語者專屬模型皆出自通用模型，每個模型的每個高斯混合順序可相對應，因此當我們要組合兩個單語者模型時，只需考慮一個模型之第 k 個高斯混合與另一模型之第 k 個高斯混合的組合，不需考慮與第ℓ個高斯混合的組合。也就是說，當組合第 i 語者與第 j 語者的模型時，所產生新模型之第 k 個高斯混合的期望值向量與變異矩陣爲：

$$\boldsymbol{\mu}_{i,j}^{k} = \mathbf{D}\{\log[\exp(\mathbf{D}^{-1}\boldsymbol{\mu}_{i}^{k}) + \exp(\mathbf{D}^{-1}\boldsymbol{\mu}_{j}^{k})]\}, \tag{3}$$

$$\boldsymbol{\Sigma}_{i,j}^{k} = \mathbf{D}\{\log[\exp(\mathbf{D}^{-1}\boldsymbol{\Sigma}_{i}^{k}(\mathbf{D}^{-1})') + \exp(\mathbf{D}^{-1}\boldsymbol{\Sigma}_{j}^{k}(\mathbf{D}^{-1})')]\}, \tag{4}$$

其中 $\boldsymbol{\mu}_i^k$ 與 $\boldsymbol{\Sigma}_i^k$ 分別是第 i 語者模型λ_i之第 k 個高斯混合的期望值向量與變異矩陣；$\mathbf{D}$ 爲 Discrete Cosine Transform Matrix；而(′)爲轉置。至於混合權重則維持原來的值。

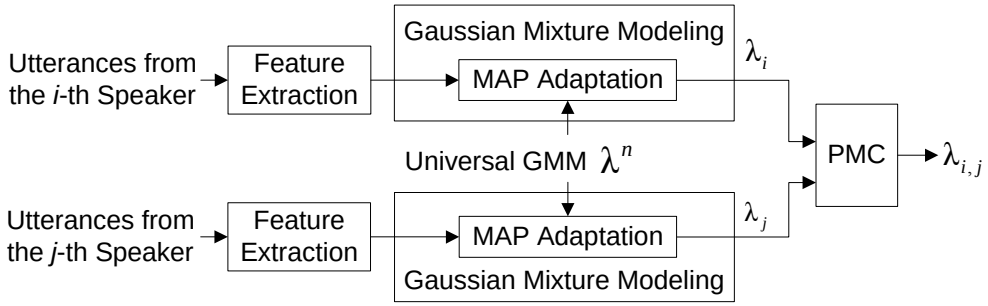


圖 9：基於 parallel model combination 之模型產生方式的雙語者識別架構。

3.2 單一階段的識別系統

由於上述兩階段的識別系統共牽涉 N 個單語者模型與 C_2^N個雙語者模型，我們應該也可以將這兩組模型放在一起比對測試音訊，亦即執行(N+C_2^N)種類別的辨識。更明確地說，若將單語者模型λ_i記爲$\lambda_{i,i}$, $1 \le i \le N$ ，則系統可決定一未知受測音訊的語者爲：

$$(I^*, J^*) = \arg\max_{1 \le i,j \le N} \Pr(\mathbf{X} \mid \lambda_{i,j}). \tag{5}$$

而這裡如果 $I^* = J^*$，代表該音訊僅包含單語者聲音；若 $I^* \neq J^*$，則代表該音訊包含第 I^* 語者與第 J^*語者的聲音。單一階段識別系統的好處是省去了重疊語音偵測的步驟，因此也少了這步驟所產生的誤判，但需付出的代價是辨識的類別較多，且混淆性也較高，因此其效能不一定較好。

4. 識別效能評估

有關上述研究方法的效能評估結果可參考文獻[17]上所討論，筆者在此不做詳細說明，但其中效能評估的指標則值得特別一提。傳統上，語者識別使用精確度(Identification Accuracy)來評估系統效能，定義爲：

$$\text{Acc (in \%)} = \frac{\text{正確識別出的語句數目}}{\text{測試語句數目}} \times 100\%$$

但由於我們考慮的重疊語音中包含兩位語者，因此「正確識別出的語句數目」並不能反應語音中已有部分語者已被識別出的情形。例如，有一個重疊語音內含語者 s_1 與 s_2，而系統識別的結果爲 s_1 與 s_4，則雖然結果並不正確，但已部分正確。因此，我們可定義兩種精確度來評估系統的效能：

$$\text{Acc.1 (in \%)} = \frac{\text{正確識別出的雙語者數目}}{\text{測試語句數目}} \times 100\%\ ,$$

$$\text{Acc.2 (in \%)} = \frac{\text{正確識別出的語者數目}}{\text{測試語者數目}} \times 100\%\ 。$$

對上述例子而言，「正確識別出的雙語者數目」爲 0，「測試語句數目」爲 1，因此 Acc.1 = 0%。但「正確識別出的語者數目」爲 1，「測試語者數目」爲 2，因此 Acc.2 = 50%。顯然，Acc.2 必高於 Acc.1，但 Acc.2 較能夠反應部份語者被正確識別的情形。另外，當未知測試語句是否爲重疊語音時，若測試語句僅含單語者語音，而系統誤判爲雙語者；或測試語句包含兩位語者語音，而系統誤判爲單語者語音，此時識別精確度該如何定義? 我們可以考慮將每一語音的內含語者都設爲 2，當測試語句僅含單語者時，例如 s_1，則可將其所含語者視爲 s_1 與 s_1，因此不違背語者數爲 2 的設法。亦即，若此時系統識別的結果爲 s_1 與 s_4，則 Acc.2 = 50%。同理，若測試語句所含語者爲 s_1 與 s_2，而系統識別的結果爲單語者 s_1，則 Acc.2 亦爲 50%。在文獻中，實驗資料包含 10 位男性語者，測試 150 個單語者語句及 1305 個重疊語音之結果如表 1。我們發現兩階段的系統較單一階段系統的效能爲佳。有興趣的讀者可參閱文獻[17]上的詳細結果。

表 1：識別 150 個單語者語句及 1305 個重疊語音之結果。

方法	Acc.1	Acc.2
兩階段系統	91.1%	94.3%
單一階段系統	89.3%	92.1%

5. 結語

目前大部分的語者辨識研究仍專注於單語者的辨識，這並不能滿足許多場合可能有多位語者同時說話的情形。本文探討了 1)如何判斷未知語音中是否包括兩位語者同時說話、2)如何判斷一未知重疊語音是由哪兩位語者所說、3)如何由一長串音訊資料中判斷各區段所含語者或雙語者。然而這項研究仍處於初步探討的階段，因此對於整個重疊語音辨識問題僅考慮了冰山的一角，許多更實際情況的挑戰尚待克服。未來首要工作是先擴充語音資料庫，包括大量語者的資料庫須先建立起來，且該資料庫應考慮能同時作爲語者辨識與語音辨認之用，甚至能支援其他相關研究，這項工作將有賴各界參與來共同推動。

6. 參考文獻

1. E. Shriberg, A. Stolcke, and D. Baron, “Observations on overlap: findings and implications for automatic processing of multi-party conversation”, In *Proc. European Conference on Speech Communication and Technology*, 2001.
2. A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, “Meeting Browser: Tracking and summarizing meetings”, In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
3. H. G. Okuno, T. Nakatani, T. Kawabata, “Listening to two simultaneous speeches”, *Speech Communication*, Vol. 27, pp. 299-310, 1999.
4. N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, “Meetings about meetings: research at ICSI on speech in multiparty conversations”, In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.
5. Ö. Çetin and E. Shriberg, “Speaker overlaps and ASR errors in meetings: effects before, during, and after the overlap”, In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.
6. K. Boakye, B. Trueba-Hornero, O. Vinyals, G. Friedland, “Overlapped speech detection for improved speaker diarization in multiparty meetings”, In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.
7. K. Yamamoto, F. Asano, T. Yamada, and N. Kitawaki, “Detection of overlapping speech in meetings using support vector machines and support vector regression”, *IEICE Trans. Fundamentals*, Vol. e89-a, No. 8, pp. 2158-2165, 2006.
8. D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models”, *IEEE Trans. Speech Audio Proc.*, Vol. 3, No. 1, pp. 72-83, 1995.
9. S. E. Johnson, “Who spoke when?–Automatic segmentation and clustering for determining speaker turns”, In *Proc. European Conference on Speech Communication and Technology*, 1999.
10. S. E. Tranter and D. A. Reynolds, “An overview of automatic speaker diarization systems”, *IEEE Trans. Audio Speech Language Proc.*, Vol. 14, No. 5, pp. 1557 – 1565, 2006.
11. J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, “Speech and language technologies for audio indexing and retrieval”, *Proc. IEEE* Vol. 88, No. 8, pp. 1338-1353, 2000.
12. A. Pikrakis, T. Giannakopoulos, S. Theodoridis, “A speech/music discriminator of radio recordings Based on Dynamic Programming and Bayesian Networks”, *IEEE Trans. Multimedia*, Vol. 10, No. 5, pp. 846 - 857, 2008.
13. N. Mesgarani, M. Slaney, S. Shamma, “Discrimination of speech from non-speech based on multiscale spectro-temporal modulations”, *IEEE Trans. Audio Speech Language Proc.*, Vol. 14, No. 6, pp. 920–930, 2006.
14. A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society*, Vol. 39, pp. 1–38, 1977.
15. M. Gales and S. Young, “Robust continuous speech recognition using parallel model combination”, *IEEE Trans. Speech Audio Proc.*, Vol. 4, No. 5, pp. 352 - 359, 1996.
16. D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models”, *Digital Signal Processing*, Vol. 10, pp. 19-41, 2000.
17. W. H. Tsai and S. J. Liao, "Speaker Identification in Overlapping Speech", Journal of Information Science and Engineering, 26(5): 1891-1903, 2010.
18. A. Hyvärinen and E. Oja, “ Independent Component Analysis: Algorithms and Applications,” *Neural Networks*, Vol.13, no 4-5, pp. 411-430, 2000.

19. C. J. Lin "On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization", *IEEE Transactions on Neural Networks*, 18 (6): 1589–1596, November, 2007.
20. J. F. Cardoso, "Blind signal separation: statistical principles", *Proc. of the IEEE*, vol. 9, no. 10, pp. 2009-2026, October 1998.
21. J. T. Chien and B. C. Chen, "A new independent component analysis for speech recognition and separation", *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 14, no. 4, pp. 1245-1254, July 2006.
22. D. L. Wang, Speech Separation By Humans And Machines, Chapter 12, pp. 181-197, 2005.
23. Y. Li and D. L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1475-1487, 2007.
24. T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, Lang. Proc.*, vol. 15, no. 3, pp. 1066 - 1074, 2007.
25. G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *Journal of the Acoustical Society of America*, vol.124, no. 2, pp. 1306-1319, 2008.
26. D. L. Wang and G. J. Brown, "Computational Auditory Scene Analysis: Principles, Algorithms and Applications," *Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 13-13, 2008.
27. C. L. Hsu and J. S. Jang , "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 18, no. 2, pp. 310-319, 2010.