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：王駿發
主　編：曾淑娟
執行編輯：黃　琪
第十六卷第四期
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

ROCLING-2005

「第十七屆自然語言與語音處理研討會」已開放線上報名 (http://www.aclclp.org.tw/)，報名截止日 8/31 日。會議時間：9/15～9/16；地點：國立成功大學電機系館；專題演講：主講人：加拿大蒙特婁大學 Prof. Jian-Yun Nie 演講，講題：Statistical Language Modeling and Information Retrieval；議程(暫訂)請參閱第二～三頁。

博碩士論文獎申請

第五屆博碩士論文獎申請已於 7/31 日截止，本次共收到 10 篇碩士論文及兩篇博士論文申請，評審結果預計九月初公布於學會網頁，並將於 9/15 日之會員大會舉行頒獎儀式。

「台灣口音英語語料庫」開放申請

EAT 語料從 2004 年 5 月起分別由台大、師大、成大、清大及交大等五個單位開始進行收集，至 2005 年 1 月初步完成收集，從各單位回收之語料經由工研院電通所匯整並請專人做語料庫整理，整理後之語料依音檔之品質及所唸內容之正確性分爲可用(usable)及不可用(unusable)兩大類，可用之語料再依英語系及非英語系細分，然後再依性別做最後的分類，綜合所得的語料。EAT 語料分爲電話及麥克風語料，電話語料部份是透過 Dialogic 電話語音介面卡，以所錄得的 8KHz,8Bits,Mulaw 格式的取樣點，經程式轉成 8khz, 16bits,pcm 格式的取樣點，然後將所有取樣點存放一.wav 格式的音檔，麥克風語料則是由各錄音單位所準備的個人電腦及麥克風，直接從 pc 的音效卡錄製 16khz,16bits 的聲音訊號，然後將所有取樣點存成一.wav 格式的音檔。語料庫申請辦法請參閱本會網頁，相關說明請參閱第四～五頁。

「研究院雙語詞網」開放申請

「中央研究院中英雙語詞網」(The Academia Sinica Bilingual Wordnet)，簡稱「研究院雙語詞網」(Sinica BW)，爲一涵蓋約十萬英文同義詞集（SynSet）之中英雙語電子資料庫。本資料庫以英文 WordNet 架構爲基礎，並以以台灣地區的語言使用爲經驗基礎。提供的訊息包含中英雙語跨語言資訊轉換、詞義的區分與詞義關係的連結以及使用領域（含使用之頻率）。讓不同來源的典藏知識內容，可以轉換成互通的 (inter-operable) 訊息。所引用的資料主要爲中央研究院文獻語料庫（語言所），詞庫小組（資訊所）開發的資料外。另外引用了普林斯頓大學 的 WordNet（http://wordnet.princeton.edu/），以及遠見科技股份有限公司與中研院共同開發資料。本資料庫之原始內容（除英語 WordNet 單語資料庫由普林斯頓大學開發擁有並公開授權。）其智財權由中央研究院與遠見科技股份有限公司共同持有。公開授權資料分別以純文字以及 XML 檔案格式儲存。網址：http://bow.sinica.edu.tw/ 。申請辦法請參閱本會網頁，資料庫說明請參閱第六～七頁。

Tentative Program of ROCLING XVII

September 15		
Time	**Session**	**Chair**
08:30-09:00	Registration	
09:00-09:10	Opening Ceremony 王駿發 教授	
09:10-10:10	Keynote Speech – I Prof. Jian-Yun Nie	簡立峰 教授
10:10-10:40	Coffee Break	
10:40-12:00	S1: Language/Speaker Identification	陳信宏 教授
	S2: Information Retrieval/Extraction & Summarization	陳信希 教授
12:00-13:20	Lunch、中華民國計算語言學學會會員大會	
13:20-14:00	Invited Speech – I 陳克健 教授	吳宗憲 教授
14:00-14:10	Break	
14:10-15:10	S3: Language Learning	張俊盛 教授
	Tutorial: Sinica BOW	黃居仁 教授
15:10-15:30	Coffee Break	
15:30-16:50	Panel Discussion	余孝先 博士
17:00-18:30	Tainan City Tour	
18:30-20:30	Banquet	

September 16		
Time	**Session**	**Chair**
09:00-09:40	Invited Speech – II 張俊盛 教授	簡仁宗 教授
09:40-10:00	Coffee Break	
10:00-12:00	S4: Speech Analysis/Synthesis	鄭秋豫 教授
	S5: Syntax/Semantics	許聞廉 教授
12:00-13:00	Lunch	
13:00-14:40	S6: Speech Recognition/Enhancement	王小川 教授
	S7: Multilingual/Multimedia Processing	張景新 教授
14:40	Closing	

94/9/15(四)

S1: Language/Speaker Identification

- A Novel Algorithm for Speaker Change Detection Based on Support Vector Machine
  王駿發,林博川,王家慶,宋豪靜
- Speaker Segmentation and Clustering for the Recorded Speech
  Hsiao-Chuan Wang, Chun-Ching Su
- 結合聲學與韻律訊息之強健性語者辨認方法
  Yuan-Fu Liao, Zhi-Xian Zhuang, Zi-He Chen, Yau-Tarng Juang
- Language Identification based on Gaussian Mixture Model Tokenizer and Language Model
  Hsiao-Chuan Wang, Zhi-Jie Chang

S2: Information Retrieval/Extraction & Summarization

- A Practical Passage-based Approach for Chinese Document Retrieval
  Szu-Yuan Chi, Chung-Li Hsiao, Lee-Feng Chien
- Web Information Extraction for the Creation of Metadata in Semantic Web
  Ching-Long Yeh, Yu-Chih Su
- 以概念分群爲基礎之新聞事件自動摘要
  劉政璋,葉鎮源,柯皓仁,楊維邦
- 中文句子相似度之計算與應用
  鄭守益,梁婷

S3: Language Learning

- 日本學生學習華語的聲調偏誤分析:以二字調爲例
  張可家,陳麗美
- 電腦輔助閱讀測驗自動出題
  楊媛茜,楊捷扉,張嘉銘,張俊盛
- FAST：電腦輔助英文文法出題系統
  陳佳吟,柯明憲,吳紫葦,張俊盛

94/9/16(五)

S4: Speech Analysis/Synthesis

- 應用錯誤型態分析於英語發音輔助學習
  湯士民,莊則敬,吳宗憲
- A Mandarin Text-to-Speech System Using Prosodic Hierarchy and a Large Number of Words
  余明興,張唐瑜,許燦煌,蔡育和
- Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification
  包蒼龍,陳育得,葉俊亨,張原豪
- 閩南語語句基週軌跡產生: 兩種模型之混合與比較
  古鴻炎,黃維
- 台灣閩南語聲調評分系統評估與研究
  蔡岳廷,廖嘉新,呂道誠,呂仁園
- Statistical Analysis of Two Polarity Detection Schemes in Speech Watermarking
  Bin Yan, Zhe-Ming Lu, Jeng-Shyang Pan, Sheng-He Sun

S5: Syntax/Semantics

- A Probe into Ambiguities of Determinative-Measure Compounds
  Shih-Min Li, Su-Chu Lin, Keh-Jiann Chen
- Machine Learning Approach to Robust Chinese Shallow Parsing
  Shih-Hung Wu, Cheng-Wei Shih, Chia-Wei Wu, Tzong-Han Tsai, Wen-Lian Hsu
- 異體字語境關係分析和建立
  周亞民,黃居仁
- 台語變調系統實作研究
  楊允言,李盛安,劉杰岳,高成炎
- 從雙語學術名詞庫中抽取中文語意
  白明弘,陳克健,張俊盛
- 利用向量支撐機辨識中文基底名詞組的初步研究
  張席維,高照明,劉昭麟

S6: Speech Recognition/Enhancement

- Perceptual Factor Analysis for Speech Enhancement
  Chuan-Wei Ting, Jen-Tzung Chien
- 國語廣播新聞語料轉述系統之效能評估
  Yih-Ru Wang, 張隆勳, Sin-Horng Chen
- An Approach of Using the Web as a Live Corpus for Spoken Transliteration Name Access
  Ming-Shun Lin, Chia-Ping Chen, Hsin-Hsi Chen
- 風險最小化準則在中文大詞彙連續語音辨識上之研究
  郭人瑋,劉士弘,陳柏琳

S7: Multilingual/Multimedia Processing

- 基於統計與迭代的中英雙語語料詞與小句對應演算法
  黃子桓,高照明
- Cross-Linguistic Comparison of the MARKET Metaphors
  Siaw-Fong Chung
- 電視新聞語料場景的自動切割與分類
  姜柏亘,謝鴻文,呂仁園
- Improving Translation of Low-Frequency Unknown Proper Names Using a Two-Stage Hybrid Translation Extraction Method
  Min-Shiang Shia, Jiun-Hung Lin, Scott Yu, Wen-Hsiang Lu
- Translation Divergence Analysis and Processing for Mandarin-English Parallel Text Exploitation
  Shun-Chieh Lin, Jia-Ching Wang, Jhing-Fa Wang

以上暫訂議程若與大會網頁有異，則以大會網頁爲準。
大會網頁：http://www.aclclp.org.tw/rocling2005.html

台灣口音英語語料庫
English Across Taiwan(EAT)

一. EAT 錄音計畫說明

EAT 錄音計畫共發出 600 份錄音提示卡，每份提示卡皆含 80 個錄音句，其中包含英文長句，英文短句，英文單詞及中英夾雜句等.600 份提示卡由五個單位合力完成錄音，每個單位負責 120 份提示卡，而每一份提示卡需分別由英語系及非英語系學生各錄製一份，每一學生需錄製麥克風語料及電話語料各一份，麥克風語料錄製 16khz 取樣頻率 16bits 的取樣點音檔，電話語料錄製 8khz 取樣頻率 16bits 的取樣點音檔.其中電話語料又可細分爲 600 份(英語系+非英語系)固定式電話(PSTN)語料 及 600 份(英語系+非英語系)行動電話(GSM)語料，歸納如下表所列：

600份提示卡
- 600個英語系學生(提示卡編號100000-100599)
  - 麥克風語料
    - 600份(發給學生自行錄製或集中錄音)
  - 電話語料
    - PSTN語料300份(各校架站收集)
    - GSM語料300份(統一撥至0800351151收集)
- 600個非英語系學生(提示卡編號101000-101599)
  - 麥克風語料
    - 600份(發給學生自行錄製或集中錄音)
  - 電話語料
    - PSTN語料300份(各校架站收集)
    - GSM語料300份(統一撥至0800351151收集)

各單位將負責收集 120 份 PSTN 及 120 份 GSM 的語料，其中各單位的 PSTN 語料將由各單位自行架錄音站收集，而 GSM 語料則統一由工研院的語料錄音站收集.每個單位錄音完成後，計含 240 份麥克風語料，120 份 PSTN 及 120 份 GSM 語料各提示卡號的分配如下:

提示卡分配
- 師大: (100000-100119, 101000-101119)
  - 陳柏琳 老師
- 交大: (100120-100239, 101120-101239)
  - 陳信宏老師, 王逸如老師
- 清大: (100240-100359, 101240-101359)
  - 張俊盛老師, 張智星老師
- 成大: (100360-100479, 101360-101479)
  - 簡仁宗老師
- 台大: (100480-100599, 101480-101599)
  - 李琳山老師

二. 錄音設備及環境

EAT 語料分爲電話及麥克風語料，電話語料部份是透過 Dialogic 電話語音介面卡，以所錄得的 8KHz、8Bits、Mulaw 格式的取樣點，經程式轉成 8khz、16bits、pcm 格式的取樣點，然後將所有取樣點存放一.wav 格式的音檔，麥克風語料則是由各錄音單位所準備的個人電腦及麥克風，直接從 pc 的音效卡錄製 16khz、16bits 的聲音訊號,然後將所有取樣點存成一.wav 格式的音檔。

注意：所有音檔內容皆屬於 raw 格式，也就是沒有經過 dc-offset 及 silence removal 的處理。

三. EAT 語料統計

EAT 語料從 2004 年 5 月開始收集，至 2005 年 1 月初步完成收集，從各單位回收之語料經由工研院電通所匯整並請專人做語料庫整理,整理後之語料依音檔之品質及所唸內容之正確性分爲可用(usable)及不可用(unusable)兩大類，可用之語料再依英語系及非英語系細分，然後再依性別做最後的分類，綜合所得的語料，依 PSTN，MIC 及 GSM 分類,得到如下的統計結果：

MIC16K語料				
	可用			
	英語系		非英語系	
	男性	女性	男性	女性
句數	11977	30094	25432	15540
人數	166	406	368	224

GSM語料				
	可用			
	英語系		非英語系	
	男性	女性	男性	女性
句數	6168	15681	12721	8048
人數	85	216	192	122

PSTN語料				
	可用			
	英語系		非英語系	
	男性	女性	男性	女性
句數	5582	14244	10584	6685
人數	82	206	160	103

中央研究院中英雙語詞網
The Academia Sinica Bilingual WordNet

開放授權的資料包括下列所示的二大類訊息：

A. 領域分類樹：

參考「中國圖書分類法」爲基準，並參考各知識分類與實際研究經驗，提出：包含九大類的知識分類（Knowledge Content），涵蓋 438 個領域，並因應語言資源特性加入下列語言使用（Language Usage）的各類訊息：專名（說明文字符號的指涉）（Proper Name）、語體（說明文字符號的使用）（Genre/Strata）、各種語言／詞源（Language/Etymology）、各國地名（Country Name）。知識分類（Knowledge Content）的九大類分別是：人文學科（Humanities）、社會科學（Social Science）、形式科學（Formal Science）、自然科學（Natural Science）、醫療科學（Medical Science）、工程科學（Engineering Science）、應用產業（Production Industry）、藝術（Fine Arts）以及休閒娛樂（Recreation）。

B. 英漢雙語詞網對應資料庫：英文詞形、對應的 WordNet1.6 之同義詞集（synset）以及詞類爲基準，每一筆紀錄皆標示以下訊息：

（1） 英文詞形。

（2） 英文 WordNet1.6 同義詞集 offset：該英文詞形對應的 WordNet 1.6 版本同義詞集 offset，詳細資料請請參閱參考文件一。

（3） 詞類：詞類，包含名詞（Noun）、動詞（Verb）、形容詞（Adjective）以及副詞（Adverb），詳細資料請請參閱參考文件一。

（4） 英文詞形搭配詞類所屬的頻率等級：據詞彙分佈情況可分爲三層次依序爲核心詞彙、通用詞彙以及參考詞彙。英文區分核心、通用與參考詞彙的原則爲：（1）、核心詞彙是 BNC, Brown, CIDE 等幾個語料庫取累計頻率 66%時的所有詞，再把各語料庫的個別詞表交集，得到的結果便是核心詞彙。（2）、通用詞彙：同上，但累計頻率取到 80%。（3）、參考詞彙：其餘在各英文資源有收的詞。

（5） 中文對譯詞形：在 WordNet1.6 同義詞集中該英文詞形對譯的中文詞形，詳細資料請請參閱參考文件二。

（6） 中文對譯詞形搭配詞類所屬的頻率等級：根據中文詞彙分佈情況可分爲四層次依序爲核心詞彙、通用詞彙、參考詞彙以及一般詞彙。參考詞彙包含了通用詞彙與核心詞彙，通用詞彙則包含了核心詞彙。（1）、核心詞彙是指所參考的五本辭典都列的詞且出現在中研院平衡語料庫語料庫十次以上。（2）、通用詞彙則收錄在任意三本辭典以上的詞且出現在中研院語料庫四次以上。（3）、參考詞彙指的是收錄在三本以上辭典的詞，或收錄在五本辭典中任一本且出現在中研院語料庫一次以上，或者是同義詞詞林的標題詞。區分原則的詳細資料請請參閱參考文件三。

（7） 詞彙領域分類：針對該英文詞形在 WordNet 1.6 同義詞集給與相對應於領域分類樹之領域訊息，詳細資料請請參閱參考文件四、五。

WordNet1.6 同義詞集原本共 99,642 筆，英文詞形爲 122,045 個，中文詞形爲 109,970 個，以英文詞形、對應的 WordNet1.6 之同義詞集（synset）以及詞類爲基準，總共有 173,941 筆資料，以英文詞義爲基礎其中 24,222 筆有領域訊息。

範例

A. 領域分類樹：檔案名稱 20050311domain.xls

人文學科Humanities

　　語言學　linguistics

B. 漢英雙語詞網對應資料庫：以英文詞形、對應的 WordNet1.6 之同義詞集（synset）以及詞類爲基準，每一筆紀錄皆標示以下訊息：

a. 純文字檔：檔案名稱 SinicaBOW_License1.0 英漢雙語詞網.txt

65　exercise　Noun　　00469856通用詞彙例題 ◎ 通用詞彙

66　exercise　Noun　　00411620通用詞彙體操 ◎ 核心詞彙 、 運動 ◎ 核心詞彙體操 ◎ gymnastics

b. XML 檔：檔案名稱 SinicaBOW_License1.0 英漢雙語詞網.xml

```
<Record Conut="65">
    <EnglishLemma>exercise</EnglishLemma>
    <POS>Noun</POS>
    <WordNetSynsetOffset Version="1.6">00469856</WordNetSynsetOffset>
    <EnglishFrequancyRank>通用詞彙</EnglishFrequancyRank>
    <ChineseTransList>
        <ChineseTrans>
            <ChineseLemma>例題</ChineseLemma>
            <ChineseFrequancyRank>通用詞彙</ChineseFrequancyRank>
        </ChineseTrans>
    </ChineseTransList>
</Record>
<Record Conut="66">
    <EnglishLemma>exercise</EnglishLemma>
    <POS>Noun</POS>
    <WordNetSynsetOffset Version="1.6">00411620</WordNetSynsetOffset>
    <EnglishFrequancyRank>通用詞彙</EnglishFrequancyRank>
    <ChineseTransList>
        <ChineseTrans>
            <ChineseLemma>體操</ChineseLemma>
            <ChineseFrequancyRank>核心詞彙</ChineseFrequancyRank>
        </ChineseTrans>
        <ChineseTrans>
            <ChineseLemma>運動</ChineseLemma>
            <ChineseFrequancyRank>核心詞彙</ChineseFrequancyRank>
        </ChineseTrans>
    </ChineseTransList>
    <DomainList>
        <Domain>
            <ChineseDomain>體操</ChineseDomain>
            <EnglishDomain>gymnastics</ChineseDomain>
        </Domain>
    </DomainList>
</Record>
```

Sinica BOW: http://BOW.sinica.edu.tw/　　WordNet: http://wordnet.princeton.edu/

現階段大詞彙連續語音辨識研究之簡介

陳柏琳

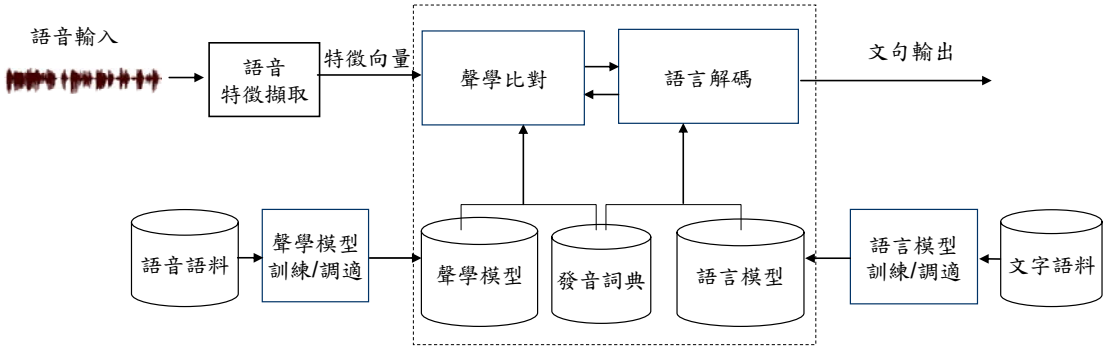
台灣師範大學資訊工程研究所

berlin@csie.ntnu.edu.tw

1. 前言

語音長久以來是人與人之間最自然且最方便的溝通方式[1]。隨著電子數位科技的蓬勃發展以及無線通訊與網際網路的創新普及，傳統的桌上型電腦不再是人們唯一主要的資訊存取平台，有可能取而代之的是各式各樣的手攜式設備(如 PDA、Mobile Phone、Tablet PC 等)以及更多的行動載具與家電產品，這些設備將變成是可以計算、通訊與上網的智慧型設備，而且朝輕薄短小的趨勢演進發展。同時，將不是每種設備都具有螢幕、鍵盤和滑鼠等這些人們習以為常的輸出入裝置；就算是有，它們也將不若過去在桌上型電腦使用時那樣地方便。於是「語音」這種人類最自然且最容易使用的溝通媒介，可能會在未來扮演著人類與各式智慧型設備間最主要的人機介面，徹底改變人類長久以來與其之互動方式，進而擴展人類對各式智慧型設備的使用層面與資訊存取的效率。另一方面，日常生活中可以存取與使用的多媒體影音資訊愈來愈多，例如廣播電視節目、語音信件、演講錄影和數位典藏等。這些多媒體資訊可以從網路上大量地取得，已經成為傳統文字資訊外社會大眾廣泛使用的資訊來源。顯而易見的是，在上述的絕大部分多媒體資訊中，語音可以說是最具語意的主要內涵之一，當播放出多媒體的語音資訊或是顯示出對應的正確轉寫文字時，我們就可以大概地瞭解其中所要傳達的主題或概念[2]。因此，語音辨識技術對多媒體資訊處理也扮演著相當重要的角色，近年來在國際上有相當多從事多媒體語音內涵自動轉寫的研究被發表，其中常以廣播新聞[3, 4]、電話交談式語音[5]、演講[6]與口述歷史典藏[7]的大詞彙連續語音辨識的研究為主。

大詞彙連續語音辨識基本上包括了三個主要模組：前端處理、聲學比對、語言解碼，如圖



圖一、大詞彙連續語音辨識流程。

一所示。前端處理將數位語音訊號切割成重疊的音框(Frames)，進行語音特徵向量擷取；聲學比對將已建立好的聲學模型(如隱藏式馬可夫模型(Hidden Markov Models, HMM)，可以是音素、音節、詞爲單位)與輸入語句中每個可能語音段落的特徵向量作比對；語言解碼則是依據所有可能候選詞段落的聲學相似度與候選詞間的語言模型(如 N連語言模型等)限制，進行解碼找出機率最大(最有可能)的文句。以下，將針對當前大詞彙連續語音辨識的相關技術發展作簡介。

2. 語音特徵(Speech Features)

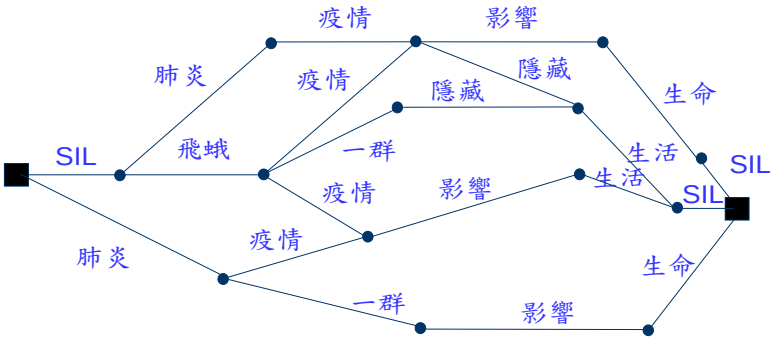
以考量人耳聽覺感知出發的梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC)[8]或是感知線性預測係數(Perceptual Linear Prediction Coefficients, PLPC)[9]已成爲目前主流的語音特徵向量擷取方法之一，配合上它們的一階與二階時間軸導數(Time Derivatives)、以及特徵平均值與變異數正規化(Mean and Variance Normalization)的強健性(Robustness)處理後，可以在一般大詞彙連續語音辨識問題上得到不錯的效果。近幾年則陸續有研究嘗試針對這些語音特徵向量作進一步處理，最常見的是對語音特徵向量作線性轉換並降低維度只保留具有鑑別力的特徵成分，例如使用線性鑑別分析(Linear Discriminant Analysis, LDA)[10]、異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[11, 12]、異質性鑑別分析(Heteroscedastic Discriminant Analysis, HDA)[13]等。其中線性鑑別分析是假設所有類別特徵向量的分佈變異是相同的；而異質性線性鑑別分析與異質性鑑別分析則是打破這樣的假設。同時，也有許多的研究嘗試以核函數線性鑑別分析(Kernel Linear Discriminant Analysis, Kernel LDA)[14]對語音特徵向量做進一步處理，希望藉由核函數將特徵向量投射到高維度特徵空間作線性鑑別分析，解決在原特徵空間可能存在的非線性鑑別問題。

另一方面，由於在聲學模型(例如隱藏式馬可夫模型狀態觀測機率分佈)常使用具對角化共變異矩陣(也就是假設特徵向量維度間彼此爲無關的)的高斯分佈，但是上述的語音特徵向量或是鑑別分析並不保證此一特性，因而有學者提出以最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)，嘗試讓轉換過後的共變異矩陣的值集中在對角線上，在對聲學模型相似度影響最小的條件下，儘量滿足對角化共變異矩陣的要求。因此，目前在大詞彙連續語音辨識的語音特徵擷取上常見到以結合線性鑑別分析與最大相似度線性轉換(LDA-MLLT)[4]或是異質性線性鑑別分析與最大相似度線性轉換(HLDA-MLLT)[3, 5]等的一些作法。

3. 聲學模型(Acoustic Models)

語音辨識的聲學模型通常使用隱藏式馬可夫模型，在傳統上常採用最大相似度訓練方法[16]，而此種模型訓練並沒有考慮到語音辨識時模型間彼此的關係，在模型參數訓練完成後有可能使

得語音特徵向量落在對應的聲學模型與非相關模型的相似度值同時變大，產生辨識上的混淆。因此近十幾年來有所謂的鑑別式聲學模型訓練(Discriminative Acoustic Model Training)方法被提出來，不以最大化訓練聲學語料的相似度爲目標，而以最小化分類(或辨識)錯誤爲目標，常見的有最小化分類錯誤(Minimum Classification Error, MCE)[17]、最大化交互資訊(Maximum Mutual Information, MMI)[18]、全面風險估測法(Overall Risk Criterion Estimation, ORCE)[19]、最小化貝式風險(Minimum Bayes Risk, MBR)[20]、最小化音素錯誤(Minimum Phone Error, MPE)[21]等。其中尤以英國劍橋大學在 2002 年左右所提出的最小化音素錯誤(MPE)聲學模型訓練方法，它結合了語音辨識產生的詞圖(Word Graph or Lattice)(如圖二所示)、音素正確率計算方法、及一些訓練參數與輔助函數設定等，是當今在大詞彙連續語音辨識研究上最佳的聲學模型訓練方法之一，在若干研究上都陸續驗證了其優越的表現。最小化音素錯誤訓練方法近來也被推廣至聲學模型調適[22]、語音特徵向量之異質性鑑別分析[23]，也能夠獲得相當大的成效；同時，它也可以應用在語言模型的訓練上[24]，藉由估測語言模型參數來最大化訓練語料中每一句語音對應詞圖的期望正確率，對於語言模型的參數估測也提供了新的研究視野。



圖二、詞圖為語音辨識所有可能候選詞與詞句的簡潔表示。

另一方面，由於可取得的多媒體影音來源愈來愈多，但大多數都沒有正確的人工轉寫(Manual Transcription)資訊來供作聲學模型訓練使用。近年來已有一些研究，嘗試發展以非監督式(Unsupervised)聲學模型訓練方式，從大量龐雜的語料中擷取較可靠的語句片段供聲學模型訓練，使大詞彙連續語音辨識的準確性更爲提升或能於新的應用領域中迅速建立起新的雛形系統。例如有研究嘗試比對語音辨識自動轉寫(Automatic Transcription)與廣播或電視新聞節目對應字幕(Closed Caption)，擷取其中正確或一致的語音段落作爲訓練語料[25, 26]；或者以發音確認(Utterance Verification)的技術來克服訓練語料沒有正確人工轉寫的問題，先使用大詞彙連續語音辨識器對龐大且無人工轉寫的語料進行語音辨識，利用信心度評估(Confidence Measure)對自動轉寫的語料進行篩選，擷取較爲正確可靠的自動轉寫語料片段，達到非監督式聲學模型

訓練的目的[27, 28]。

4. 語言模型(Language Models)

絕大部分的大詞彙連續語音辨識是使用統計式語言模型(Statistical Language Models)，而以 *N* 連語言模型(*N*-gram Language Models)是最常被使用的(尤其是二連及三連語言模型)，它主要根據前面的 *N*-1 詞歷史(Word History)來決定下一個詞可能出現的機率。*N* 連語言模型的機率分佈通常由最大相似度(Maximum Likelihood Estimation, MLE)來估測[29]，近來也由學者提出以最大熵值(Maximum Entropy, ME)準則[30]、或是最小化分類錯誤[31]及最小化詞錯誤[24]等方法來訓練語言模型。然而，訓練 *N* 連語言模型時，常遭遇資料稀疏的問題(Data Sparseness Problems)，過去幾年已經有一些像平滑(Smoothing)或插補(Interpolation)等方法陸續被提出，達到不錯的效果[32]。同時也有學者提出以語句中潛藏的語意、主題或者語法資訊來做爲語言模型限制，常見的有潛藏語意分析(Latent Semantic Analysis, LSA)[33]、機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)[34]、潛藏向量狀態(Hidden Vector State, HVS)模型[35]等，這些模型也可以進一步地以機率差補方式或是透過最大熵值準則[36]等方式來與 *N* 連語言模型結合而達到不錯的效果。值得一提的是，近年來統計式語言模型(例如上述的 *N* 連語言模型)常被用於資訊檢索模型，而資訊檢索模型亦反過來被推廣至語音辨識的語言模型使用上(例如上述的潛藏語意分析、機率式潛藏語意分析等)，讓語音辨識與資訊檢索技術有了一個成功的研究交流[37]。

另一方面，在處理一些較複雜困難的大詞彙連續語音辨識課題上如廣播及電視新聞自動轉寫，由於新聞播報的主題和語言內容的詞彙使用，常具多變性與時效性，會使得統計式語言模型往往很難做到準確的估測，於是便有了所謂的語言模型調適(Language Model Adaptation)的研究[38]。語言模型調適通常會結合背景文字語料庫(Background Text Corpus)與測試語音同一時期(Contemporary)或者是同一領域(In-domain)的文字語料庫來訓練出較具強健性的調適後語言模型，以得到較佳的詞接連預測能力。例如，常見的 *N* 連語言模型調適技術有基於傳統最大事後機率(Maximum a Posteriori, MAP)估測發展出的模型插補(Model Interpolation)及詞頻數混合(Count Merging)調適技術[39]等。

5. 搜尋(Search)

大詞彙連續語音辨識的聲學比對與語言解碼通常合而爲一，以使用搜尋演算法而達成。大部分的搜尋演算法是採用一階段(One-pass)、音框同步(Frame-synchronous)的詞彙樹複製搜尋(Tree-Copy Search)方式[40]。在詞彙樹中每個分枝代表一個聲學模型(隱藏式馬可夫模型)，由

樹根到任一個樹梢的路徑代表一個詞或一些發音相同的詞，路徑上的分枝就是代表這個詞或這些詞會使用到的聲學模型。搜尋時，在每一個語音音框會同時存在數棵詞彙樹，每個詞彙樹代表不同的語言模型歷史或限制。另一方面，由於詞彙樹中存活的隱藏式馬可夫模型狀態節點可能會隨音框數呈指數倍增加，因此必須以光束剪裁(Beam Pruning)技術適當地在搜尋時剪裁分數較低的詞彙樹內部狀態節點(Internal Nodes)或不完全路徑(Partial Paths)。近年來陸續有一些學者提出以語言模型前看(Language Model Look-ahead)[41]及聲學模型前看(Acoustic Model Look-ahead)[42, 28]來預先估算尚未搜尋到語音段落的語言與聲學模型分數，再加上詞彙樹內部狀態節點本身實際搜尋時所累積的分數，當成剪裁比較的依據，在語音辨識的正確率與速度的提昇上得到不錯的結果。此外在每個音框，我們可以紀錄存活的詞彙樹的樹梢節點中分數較高者之相關資訊(這些樹梢節點本身代表著可能的候選詞)，諸如它們的語言模型歷史、對應候選詞開始與結束的音框、以及搜尋分數，然後再依此資訊建立起一個詞圖(如圖二所示)[43]，並在詞圖上使用更高階的語言模型，如詞三連(Trigram)、四連(Fourgram)語言模型等，重新進行一次詞圖搜尋(Word Graph Rescoring)，找出最佳的文句。

另一方面，也有學者以使用有限狀態轉換機(Finite State Transducer, FST)來建構大詞彙連續語音辨識，成爲一個新興的研究議題[44]。其概念主要源自於在自然語言處理(Natural Language Processing, NLP)領域，常使用有限狀態轉換機來模擬語言的文法結構與特性；用在語音辨識則融入了聲學處理的特性，將聲學模型、發音詞典以及語音模型等透過有限狀態轉換機而緊密地整合在一起，並在有限狀態轉換機上進行語音辨識之搜尋。語音辨識的有限狀態轉換機可以透過一些最佳化的演算法，諸如確定化(Determinization)演算法、最小化(Minimization)演算法、加權值推移(Weight Pushing)演算法等[45]，求得等價且確定的有限狀態轉換機。經由一系列大詞彙連續語音辨識實驗的比較發現使用有限狀態轉換機可達到與上述詞彙樹複製搜尋和詞圖搜尋一樣或者更好的辨識率與辨識速度[46]。

6. 結論

不論是多媒體資訊中語音內涵的處理、或是口語對話和語音資訊檢索系統中使用者的語音輸入都需要使用大詞彙連續語音辨識技術將多媒體中的語音內涵或使用者的語音輸入自動轉寫成文字(或音節)資訊後再作進一步處理，因此大詞彙連續語音辨識技術的角色可說是相當地重要。大詞彙連續語音辨識在過去近 20 年來一直是語音辨識研究領域中一個相當困難卻也是相當熱門的研究主題，它必須同時考慮到語音特徵擷取、聲學與語言模型訓練、搜尋比對、甚至是語音強健處理等問題，因此也爲這些研究提供了一個具挑戰性的發展與測試平台。

7. 參考文獻

[1] Juang, B. H., Furui, S., “Automatic recognition and understanding of spoken language—a first step toward natural human-machine communication,” *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1142-1165, 2000.

[2] Lee, L.S., Chen, B., “Spoken Document Understanding and Organization - The Key to Future Efficient Retrieval/Browsing Applications,” to appear in *IEEE Signal Processing Magazine*, vol. 22, no. 5, 2005.

[3] Woodland, P.C., “The Development of the HTK Broadcast News Transcription System: An Overview,” *Speech Communication*, vol. 37, 2002, pp. 47-67.

[4] Beyerlein, P., Aubert, X., Haeb-Umbach, R., Harris, M., Klakow, D., Wendemuth, A., Molau, S., Ney, H., Pitz, M., Sixtus, A., “Large Vocabulary Continuous Speech Recognition of Broadcast News – The Philips/RWTH Approach,” *Speech Communication*, vol. 37, 2002, pp. 109-131.

[5] Hain, T., Woodland, P.C., Evermann, G., Gales, M.J.F., Povey, D., Moore, G., Wang, L. Liu, X., “Automatic Transcription of Conversational Telephone Speech,” to appear in *IEEE Transactions on Speech and Audio Processing*.

[6] Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C., “Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, 2004, pp. 401-408.

[7] Byrne W., Doermann D., Franz, M., Gustman, S., Hajic, J., Oard D., Picheny M., Psutka, J., Ramabhadran B., Soergel, D., Ward, T., Zhu, W. J., “Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, 2004, 420-435.

[8] Davis, S.B., Mermelstein, P., “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans. on Acoustic, Speech, and Signal Processing*, vol. 28, no. 4, 1980, pp. 357-366.

[9] Hermansky, H., “Perceptual Linear Predictive (PLP) Analysis ff Speech,” *Journal of the Acoustical Society of America*, vol. 87, 1999, pp. 1738-1752.

[10] Duda, R.O., Hart P.E., *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973.

[11] Kumar, N., “Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition,” *Ph.D. thesis, John Hopkins University, Baltimore*, 1997.

[12] Gales, M.J.F., "Maximum Likelihood Multiple Subspace Projections for Hidden Markov Models," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, 2002, pp. 37-47.

[13] Saon, G., Padmanabhan, M., Gopinath, R., Chen, S,. “Maximum Likelihood Discriminant Feature Spaces,” in Proc. *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. II, 2000, pp. 1129-1132.

[14] Mika, S., “Fisher Discriminant Analysis With Kernels”, in Proc. *IEEE International Workshop on Neural Networks for Signal Processing*, 1999, pp. 41-48.

[15] Gopinath, R.A., “Maximum likelihood modeling with Gaussian distributions,” in Proc. *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. II, 1998, pp. 661-664.

[16] Rabiner, L., “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, 1989, pp. 257-286.

[17] Juang, B.H., Chou, W., Lee, C.H., “Minimum Classification Error Rate Methods for Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, 1997, pp. 257-265.

[18] Woodland, P.C., Povey, D., “Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition,” *Computer Speech and Language*, vol. 16, 2002, pp.25-47.

[19] Kaiser, J., Horvat, B., Kacic, Z., “Overall Risk Criterion Estimation of Hidden Markov Model Parameters,” *Speech Communication*, Vol. 38, 2002, pp.383-398.

[20] Doumpiotis, V., Byrne, W., “Lattice Segmentation and Minimum Bayes Risk Discriminative Training for Large Vocabulary Continuous Speech Recognition,” to appear in *Speech Communication*.
[21] Povey, D., “Discriminative Training for Large Vocabulary Speech Recognition,” *Ph.D Dissertation, Peterhouse, University of Cambridge*, July 2004.
[22] Wang, L., Woodland, P.C., “MPE-Based Discriminative Linear Transform for Speaker Adaptation,” in Proc. *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2004, pp. 321-324.
[23] Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H., Zweig, G. “fMPE: Discriminatively Trained Features for Speech Recognition,” in Proc. *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2005, pp. 961-964.
[24] Kuo, J.W., Chen, B., “Minimum Word Error Based Discriminative Training of Language Models," to appear in *the European Conference on Speech Communication and Technology*, September 2005.
[25] Lamel, L., Gauvain, J.L., Adda, G., “Lightly Supervised and Unsupervised Acoustic Model Training,” *Computer Speech and Language*, vol. 16, no.1, 2002, pp. 115-229.
[26] Nguyen, L., Xiang, B., “Light Supervision in Acoustic Model Training,” in Proc. *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2004, pp. 185-188.
[27] Wessel, F. and Ney, H., “Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, 2005, pp. 257-265.
[28] Chen, B., Kuo, J.W., Tsai, W.H., “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 10, no. 1, 2005, pp. 1-18.
[29] Rosenfeld, R., “Two Decades of Statistical Language Modeling: Where Do We Go from Here,” *Proceedings of the IEEE*, vol. 88, no. 8, 2000, pp. 1270-1278.
[30] Berger, A., Della Pietra, S., Della Pietra, V., “A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, vol. 22, no. 1, 1996, pp. 39-71.
[31] Kuo, H.K., Fosler-Lussier, E., Jiang, H., Lee, C.H., “Discriminative Training of Language Models for Speech Recognition,” in Proc. *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2002, pp. 325-328.
[32] Chen, S.F., Goodman, J., “An Empirical Study of Smoothing Techniques for Language Modeling,” *Computer Speech and Language*, vol. 13, 1999, pp. 359-394.
[33] Bellegarda, J.R., “Latent Semantic Mapping: Dimensionality Reduction via Globally Optimal Continuous Parameter Modeling,” *IEEE Signal Processing Magazine*, Vol. 22, No. 5, September 2005.
[34] Mrva, D. and Woodland., P.C., “PLSA-Based Language Model for Conversational Telephone Speech,” in Proc. *International Conference on Spoken Language Processing*, 2004, pp. 2257-2260.
[35] Seneviratne, V., Young, S., “The Hidden Vector State Language Model,” to appear in *the European Conference on Speech Communication and Technology*, September 2005.
[36] Wang, S., Schuurmans, D., Peng, F. and Zhao, Y., “Combining Statistical Language Models via the Latent Maximum Entropy Principle,” *Machine Learning Journal*, vol. 59, 2005, pp. 1-22.
[37] Croft, W. B. (editor), Lafferty, J. (editor), Language Modeling for Information Retrieval. Kluwer-Academic Publishers, 2003.
[38] Bellegarda, J. R., “Statistical Language Model Adaptation: Review and Perspectives,” *Speech Communication*, vol. 42, 2004, pp. 93-108.
[39] Bacchiani, M., Roark B., “Unsupervised Language Model Adaptation,” in Proc. *IEEE International Conference on Acoustics, Speech, Signal processing*, vol. I, 2003, pp. 224-227.
[40] Aubert, X. L., “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” *Computer Speech and Language*, vol. 16, 2002, pp. 89-114.
[41] Ney, H., Ortmanns, S., “Dynamic Programming Search for Continuous Speech Recognition,”

IEEE Signal Processing Magazine, vol. 16, no. 5, 1999, pp. 64-83.
[42] Seide, F., "The Use of Virtual Hypothesis Copies in Decoding of Large-Vocabulary Continuous Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, 2005, pp. 520-533.
[43] Ortmanns, S., Ney, H., Aubert, X., "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language*, vol. 11, 1997, pp. 43-72.
[44] Mohri M., Pereira, F.C.N., Riley, M., "Weighted Finite-State Transducers in Speech Recognition," *Computer Speech and Language*, vol. 16, no. 1. 2002, pp. 69-88.
[45] Mohri, M., "Weighted Finite-State Transducer Algorithms: An Overview," *Formal Languages and Applications*, vol. 148, 2004.
[46] Kanthak, S., Ney, H., Riley, M., and Mohri, M., "A Comparison of Two LVR Search Optimization Techniques," in Proc. *International Conference on Spoken Language Processing*, 2002, pp. 1309–1312.

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。

雜訊對於語音辨認之影響及語音辨認之強健性技術的介紹

洪志偉
暨南大學資訊工程學系

『雜訊』，或其較通俗的名稱，噪音，通常指的是不受歡迎的訊號，亦即雜訊本身而言並不帶有可用的資訊，且會對於原本人們所需含有資訊的信號產生干擾與破壞的作用，因而使人們無法接收到與發送端相同的訊號，而較不容易精準地擷取出想要得到的資訊，廣義的雜訊是泛指對於任何形式、種類的訊號（例如影像、圖片、樂音、聲音、震波等）的干擾源，而在這裡，我們的討論將侷限於干擾聲音訊號的雜訊。

雜訊的種類

當聲音信號從發音端傳播至接收端時，其中所經過的路徑（或稱通道）即是雜訊的來源，我們可以根據許多不同的角度來對雜訊作分類，以下是其中幾種分類的角度：

1. 對於原聲音信號的干擾形式

   有些雜訊是來自於傳輸通道本身對於不同頻率的聲音訊號，造成選擇性放大縮小(scaling)或時間延遲(time delay)的效應，這樣形式的雜訊通常稱爲摺積性雜訊(convolutional noise)，這是因爲接收到的信號是此類雜訊與原信號在時間上的摺積，而如果從頻率的角度來看，則接收到的信號頻譜由雜訊的頻譜與原信號頻譜相乘積所得。摺積性雜訊通常又稱爲通道失真(channel distortion)。顧名思義，這類的雜訊是來自傳遞聲音訊號的通道，諸如電話線、無線通訊、麥克風、一個具有迴音效應的房間等，都會對於原傳輸聲音信號造成此類摺積性地干擾作用。

   相對於摺積性雜訊，另一種雜訊則與原聲音訊號在時間上成加成性的作用，這種雜訊即稱爲加成性雜訊，此也是我們一般大眾最常想到的雜訊形式，通常稱之爲噪音，在我們日常生活裡，無時無刻都充斥著此類的雜訊，例如在房間裡的冷氣機、電腦所發出的機器運作聲、街道上車輛的引擎或喇叭聲、菜市場上鼎沸人聲等，加成性雜訊通常又稱爲背景雜訊(background noise)，這些雜訊跟我們原先欲傳達的聲音訊號疊加在一起，接收者不僅聽到原傳送的聲音訊號，也同等地接收了這些干擾源。

2. 雜訊的頻譜位置特性

   雜訊是屬於隨機信號，因此欲估計出特定時間點上雜訊的精確值是有困難的，然而我們仍然能夠根據雜訊本身的統計特性，來獲取雜訊的一些資訊，例如根據雜訊在時間上的相關序列，我們就能得到雜訊的功率頻譜，進而得知雜訊對於聲音訊號所影響的頻帶。最廣爲

人知的白色雜訊(white noise)，其功率頻譜在每個頻率上的大小都是一樣的，也就是它同時等量地干擾到聲音訊號的低頻、中頻與高頻成分，白色雜訊其時間上的特性是，不同時間點的雜訊彼此毫無相關性，亦即我們完全無法由某個時間點的雜訊值去估測另一個時間點的雜訊值。相對於白色雜訊的全頻帶特性，有些雜訊的功率頻譜則可能比較集中於低頻成分，例如俗稱的粉紅雜訊(pink noise)其功率頻譜值即與頻率值成反比。常見的汽車引擎聲、冷氣空調聲大都屬於此類低頻帶的雜訊。另外，也有特別集中於高頻帶的雜訊，例如警報器聲、救護車的警鈴聲、哨音等。

3. 雜訊特性隨時間的變化度

   如果雜訊的統計特性（例如平均值、變異數、相關序列等）並不隨時間而改變，或能夠在較長的時間之內維持不變的話，我們即稱這種雜訊爲穩定性(stationary)雜訊，如前述的白色雜訊、人聲嘈雜的雜訊、汽車引擎聲雜訊等大致都屬於這種穩定性雜訊，而當雜訊的統計特性會隨著時間而改變時，我們就稱之爲非穩定性(nonstationary)雜訊，例如忽然而來的開門聲響、咳嗽聲、敲鍵盤聲、或飛機起降的聲音等，這些雜訊的統計特性相對而言都隨時間而快速變化，因此我們較難對它們的特性有較精確的估測。

雜訊對人的影響

在介紹人對雜訊的反應之前，我們首先介紹人耳對聲音的反應，首先，對聲音的頻率而言，人耳對於不同頻率聲音的大小，有著不同的感知，相對而言，人耳對於越高頻率（約 4000 至 5000Hz 左右）的聲音感覺越敏銳，也就是當頻率越高時，稍微的聲響人便能感應到，更精確地說，人對於高低頻率聲音能量的大小有著感知的不同，例如相同能量的 1000Hz 與 100Hz 的聲音，人們就會感覺 1000Hz 的聲音會比較大聲。此外，人耳對於聲音大小的感知也不是線性的，而是類似以取對數的形式，例如當聲音大小變化 10 倍時，人的聽覺卻不認爲有變化到 10 倍這麼多，特別是當聲音越大聲時，人耳對於聲音大小的變化越不敏感，因此衡量聲音的大小時，我們是仿效人耳的效應，將聲音壓力值取以 10 爲底對數再乘上一定值 20，單位是以分貝來表示。也因此，如聲音壓力級增加 10 分貝（10 倍），人耳大約只感覺聲音強度增加了一倍（即原來的兩倍），同樣地，若聲音壓力級如增加至 20 分貝（100 倍），人耳卻只覺得聲音的強度增加爲原先的四倍。無疑地，這樣的效應對人耳具有保護作用，避免人耳因爲過大的聲音而損傷。

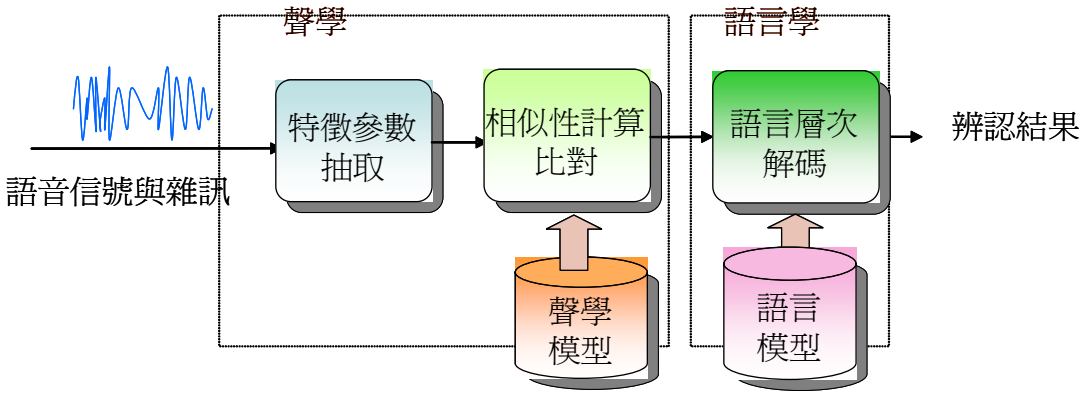
再回到雜訊的討論，如前所述，雜訊爲不悅耳、不受歡迎的聲音，它會干擾發聲者的聲音，以致接收者可能無法精確得到發聲者藉由聲音所要傳達的訊息。同時，它也會造成聽者的困擾與不悅，干擾人們的身心狀態，如前所述，可能由於人耳對於高頻率的聲音比較敏感，因此當我們聽到較高頻率的雜訊，例如警鈴聲、工地的機械聲等，通常會立即造成我們煩躁、緊張與

焦慮，然而低頻的雜訊，常見於一般居家的冷卻水塔、抽風機、冷氣機及具重低音喇叭場所，其雖然對人的生理影響並不立即且明顯，但長久對人的傷害仍不容小覷，據研究顯示，低頻雜訊對人造成的壓迫感，容易造成人們失眠、無法專心、心悸、頭痛、頭昏眼花等，也因此，近年來政府環保署相關機構開始對於此類的低頻雜訊展開研究與管制。

在雜訊的環境下，人因爲受到干擾，其講話的方式也會有所不同，例如爲了使對方聽清楚，說話者通常會試著減慢說話的速度、提高說話的音調、增加音量等等，也因此語音型態在雜訊環境與靜音環境之下的型態也有顯著的差異。然而人靈敏的聽覺系統，在雜訊環境下仍然能夠有效地擷取出說話者的語音，把雜訊『聽而不聞』，進而達到獲取資訊的目的，這是因爲巧妙的人類善於決定何者爲想要留下的資訊，而何者是要摒除的資訊，進而作有效的篩選。例如有些父母親能夠在工地的嘈雜聲或警車呼嘯聲中熟睡，然而他們的嬰兒簡短的的嗚咽或呢喃，卻能使他們驚醒。

雜訊對於語音辨認系統的影響及語音強健性技術

語音辨認系統的基本目的，即是希望能夠精確地辨認出所接收到的聲音，狹義的語音辨認，是將接收到的聲音轉換到它所對應的文字，廣義而言，則是希望擷取出接收聲音中所要得到的資訊。一個較完整的語音辨認系統，通常包含了聲學處理與語言學處理兩部分：如下圖



圖一、一個語音辨認模型的示意圖

爲了使一個語音辨認系統『聽的懂』語音，我們通常必須先蒐集許多語音資料，利用這些語音資料連帶它們所代表的文字資訊，『訓練』出語音辨認的模型，也就是先讓語音辨認系統『學習』哪些聲音是對應到哪些文字或是語意，用以訓練語音模型的這些語音資料，通常是在某特定的環境下錄製而成，如一個安靜的實驗室，或是有線或無線電話系統等，然而當此語音辨認系統應用在另一個環境時，我們無法保證此應用環境與原先訓練語料的錄製環境是相同的，此時這兩個環境的不匹配(mismatch)就自然會影響語音辨識的精確度。例如，在一個人聲嘈雜的

環境裡，語音辨認系統同時接收到欲辨認的人聲及其他人的語音，當然它無法如靈敏的人耳聽覺系統精確地篩選出要辨認的部分，因此語音辨認的效果自然就不如原先在相同訓練語料錄製環境來的好。明顯地，雜訊是造成環境不匹配、辨識率下降的主因之一。爲了克服這個問題，這方面的相關研究已經有相當長的一段歷史，且目前仍方興未艾，提出的改進方法通常稱爲強健性(robustness)技術。當一個語音辨認系統在不匹配的環境下，其辨認精確度如果沒有衰減許多，我們便稱此系統具備強健性。

強健性技術主要目的就是要降低模型訓練環境與測試環境之間的不匹配現象，從圖一之語音辨認模型的示意圖來看，由於雜訊干擾主要是在輸入的信號源部分，因此諸多強健性技術皆是建立在聲學處理的階段，如前所述，由於雜訊造成了輸入的語音信號特性與事先訓練成的語音模型的特性並不匹配，所發展的強健性技術爲了降低此不匹配，大致分爲兩大類，第一大類是試圖改變輸入語音信號或其轉換成的特徵參數的特性，使其儘量與語音模型的特性吻合，另一大類則是試圖更新事先訓練成的語音模型，使其特性逼近於輸入信號的特性。以下分別簡單介紹這兩大類技術的一些較有名的方法。

第一類方法，簡單來說是建構在信號的層次，試圖降低雜訊對於純語音信號的影響，根據這樣的目標，這一類方法又可分爲兩個角度來發展，第一個角度是儘量把雜訊成分從訊號中降低，凸顯純語音信號的部分，換句話說，就是提升信號與雜訊的比值(訊雜比，Signal to Noise Ratio, SNR)，這個角度的方法通常稱爲語音強化法，代表的技術諸如：

(1) 陣列式的接收器：藉由多重接收語音源彼此的相關性提升訊雜比
(2) 頻譜消去法：估測出雜訊的頻譜，然後試著將訊號中雜訊成分扣除掉
(3) 韋納濾波器：藉由最小化接收訊號與乾淨訊號之間的差值功率所設計的濾波器，進而把雜訊加以過濾。

第一類方法的第二個角度，則是建立不容易受到雜訊影響的特徵參數，使其本身即具備強健性。這個角度的技術諸如：

(1) 倒頻譜平均消去法：此方法是假設乾淨語音的倒頻譜(cepstrum)參數向量每一維度的平均值爲零，如果檢測出的平均值不爲零，則認定此不爲零的值是來自於雜訊，而將其扣除。此方法最先是用於處理通道的摺積性雜訊，但後來發現處理加成性雜訊仍十分有效用。
(2) 相對頻譜法：此方法是假設乾淨語音頻譜隨時間的變化速度是在某個範圍之間，換句話說，其調變頻譜是集中在某頻帶之內的，因此只要藉由濾波器將此頻帶中的調變頻譜保留下來，自然能夠消除不屬於語音的部分。

第二類方法，則是在不改變輸入語音的情況下，藉由調整語音模型，來『適應』輸入之具有雜訊的語音訊號，這類的方法諸如：

(1) 平行模型合併法：由於加成性雜訊在倒頻譜參數裡是與乾淨語音成非線性的關係，因此此法將原本乾淨語音的倒頻譜模型參數，轉換至線性頻域，使它們能夠與雜訊在線性頻域上的模型參數做線性相加，然後再轉回倒頻譜域，而得到雜訊環境下的聲學模型參數。

(2) 最大可能性線性回歸法：此方法是利用最大可能性的法則，求取一個轉換矩陣，試圖將原本聲學模型參數做線性轉換後，能夠給予輸入之帶有雜訊的信號特徵最大的機率，此法最早是用於語者調適，但同樣也適用於不同環境的調適。

以上所提之各種方法只是諸多強健性技術之一隅，由於雜訊環境之下語音辨認關係到語音辨認系統是否能夠成功地付諸應用，因此強健性的語音辨認技術仍是語音學界的一個相當重要的研究領域，每年在此主題上的期刊論文與會議論文亦不計其數。期許此強健性技術能日臻成熟，促使語音辨認系統能夠趨近於人耳聽覺系統之靈敏與智慧，帶給人們更大的便利。

參考文獻：

(1) Thomas F. Quatieri, “Discrete-Time Speech Signal Processing, Principles and Practice”, Prentice Hall Signal Processing Series

(2) Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, “Spoken Language Processing”, Prentice Hall

(3) K.C. Cole 著，邱宏義 譯, “物理與頭腦相遇的地方”, 天下文化

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。