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：吳宗憲
主　編：王新民
執行編輯：黃　琪
第二十二卷第二期
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

2011 語音訊號處理研討會

語音訊號處理研討會是中華民國計算語言學學會定期舉辦的學術交流盛會，今年度的會議將以前瞻的語音技術及語音應用(Advanced Speech Processing Technology and Application) 爲主題，邀請知名的學者專家介紹此領域最新的研發成果，以供學術界和產業界分享。

所邀請之演講者包括：成功大學吳宗憲教授、中山大學陳志堅教授、聯合大學黃豐隆教授、元智大學洪維廷教授、與工研院張森嘉博士。演講的內容包括：多語言之語音合成技術、語句無關之語者辨識技術、HCRF 聲學模型和語音辨識 IC、與結合 Google Map 之語音處理的應用等主題。

除了以上演講外，本次會議同時舉辦相關系統展示，對於語音相關工作者有極爲豐富之內容，會議議程請參閱本刊第二頁。

- 報名方式：本次會議一律採線上報名，欲報名者，請進入大會網頁點選報名，輸入報名者資料，系統將自動產生出繳費單供列印。
- 報名費用：
  一般人士：會員 500 元，非會員 600 元；學生：會員 300 元，非會員 500 元。
- 線上報名截止日 6/20(星期一)，現場報名加收 100 元。

會議網址：http://speech9.csie.ntust.edu.tw/2011Speech/

學術演講與短期課程公告

爲服務國內語音處理相關領域研究學者、學生及業界朋友，中研院資訊所與中華民國計算語言學學會(ACLCLP)特別邀請李錦輝教授於暑假期間返台講授 Digital Speech Processing and Applications 短期課程，課程內容及報名，請參閱課程網頁：http://slm.iis.sinica.edu.tw/speech_ss2011/ss2011.html，課程簡介請參閱本刊第三頁。

本課程將結合 Dr. Li Deng 的 ISCA Distinguished Lectures 及 ACLCLP 一年一度的語音訊號處理研討會，於 6 月最後一週舉辦，詳細活動日程如下：

1. Short Course on Digital Speech Processing and Applications - 李錦輝教授
   - 6/28 (TUE), 地點：台北中研院資訊所
   - 6/30 (THU), 地點：台北中研院資訊所
2. 2011 語音訊號處理研討會
   6/29 (WED), 地點：台北台灣科技大學
3. ISCA Distinguished Lectures - Dr. Li Deng (Microsoft Research)
   6/27 (MON), 地點：台北中研院資訊所
   14:00-15:30 Deep Learning: An Overview
   15:30-17:00 Feature-Domain, Model-Domain, and Hybrid Approaches to Noise-Robust Speech Recognition
   (ISCA Distinguished Lectures 演講活動免報名，演講摘要及講者簡介請參閱本刊第四~五頁)

2011 語音訊號處理研討會
Advanced Speech Processing Technology and Application

會議時間：2011 年 6 月 29 日(星期三) 9:10~16:30
會議地點：台北市基隆路四段 43 號(台灣科技大學)
主辦單位：台灣科技大學資訊工程系、中華民國計算語言學學會

會議議程:

時　　間	講　　題	主　講　人	主　持　人
09:10-09:30	**報　　到**		
09:30-09:40	**開幕致詞**	陳俊良 副院長 (台灣科技大學電資學院) 吳宗憲 理事長 (中華民國計算語言學學會)	
09:40-10:40	**Towards Multilingual Text-to-Speech Synthesis: From Monolingual to Polyglot Speech Synthesis**	吳宗憲 教授 (成功大學)	王小川 教授 (清華大學)
10:40-11:00	**Coffee Break**		
11:10-12:00	**Text-Independent Speaker Identification**	陳志堅 教授 (中山大學)	陳信宏 教授 (交通大學)
12:00-13:10	午　　餐		
13:10-14:10	**隱藏式條件隨機域聲學模型之研究與應用**	洪維廷 教授 (元智大學)	余明興 教授 (中興大學)
14:10-15:10	**整合語音處理技術與 Google Map 地理資訊之研究與應用**	黃豐隆 教授 (聯合大學)	蔡偉和 教授 (台北科大)
15:10-15:30	**Coffee Break**		
15:30-16:30	**強健式語音辨識晶片技術 (Robust Automatic Speech Recognition IC)**	張森嘉 博士 (工研院)	賴玟杏 教授 (高雄第一科大)
	閉　　幕		

會議聯絡人：(台灣科技大學)

古鴻炎 教授 guhy@csie.ntust.edu.tw

林伯慎 教授 bslin@cs.ntust.edu.tw

報名事宜：(中華民國計算語言學學會)

黃琪 小姐 aclclp@hp.iis.sinica.edu.tw

電話：02-27883799 分機 1502

2011 Short Course on Digital Speech Processing and Applications

http://slm.iis.sinica.edu.tw/speech_ss2011/ss2011.html
June 28 (TUE) and June 30 (THU), 2011
Institute of Information Science, Academia Sinica, Taipei

Prof. Chin-Hui Lee
School of ECE, Georgia Tech
Atlanta, GA 30332-0250, USA
chl@ece.gatech.edu

Course Description:
Speech is the most natural means of communication among humans. It also plays a critical role in enhancing human-machine communication. In this course, we attempt to cover all fundamental aspects of digital speech processing, including both theoretical and practical topics, starting with the acoustics of speech sounds, followed by speech analysis and parameter extraction, speech modeling, theory of linear prediction and hidden Markov models. Finally speech applications, including speech coding, synthesis, recognition and verification, will also be introduced. The linkage to acoustics and language processing will also be discussed, including topics on language modeling and microphone arrays. MATLAB demos will be used in class for illustration. Some homework exercises will also be provided for after-class learning.

Course Outline:
- Speech Communications and Acoustics of Speech Sounds (2.5 hour)
- Digital Speech Processing: Time and Frequency Domains (2.5 hours)
- Modeling of Speech: Linear Prediction and Speech Parametrization (2.5 hours)
- Speech Applications: Coding, Synthesis, Recognition and Verification (2.5 hours)

Intended Audience:
This short course is intended for researchers, engineers and professionals who are starting speech-related work and interested in more basic knowledge in digital speech processing, or those who are already involved in speech technology development and would like to learn more fundamentals. The course is designed with a broad coverage of all areas related to digital speech processing with linkages to language and acoustics.

Textbook:
(1) L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, Prentice Hall, 2010.

Supplement:
(2) C. Manning & I. Shutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

Readings:
(3) C. Cherry, *On Human Communications*, MIT Press, 1968.
(4) D. G. Stork (ed.), *HAL's Legacy*, MIT Press, 1997.

Speech tool:
Download wavesurfer from: http://www.speech.kth.se/wavesurfer/download.html

ISCA Distinguished Lectures - Dr. Li Deng (Microsoft Research)

地點：台北市中研院資訊所 106 演講廳

時間：2011-06-27 (星期一)下午

14:00~15:30 Deep Learning: An Overview

Abstract:

Today, signal processing research has a significantly widened scope compared with just a few years ago, and machine learning has been an important technical area of the IEEE signal processing society. Since 2006, deep learning-a new area of machine learning research-has emerged, impacting a wide range of signal and information processing work within the traditional and the new, widened scopes. Various workshops, such as the 2011 ICML Workshop on Learning Architectures, Representations, and Optimization for Speech and Visual Information Processing, the 2009 ICML Workshop on Learning Feature Hierarchies, the 2008 NIPS Deep Learning Workshop, the 2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, as well as an upcoming special issue on Deep Learning for Speech and Language Processing in IEEE Transactions on Audio, Speech, and Language Processing (2011) have been devoted exclusively to deep learning and its applications to various classical signal processing areas. We have also seen the government sponsor research on deep learning (e.g., the DARPA deep learning program).

The purpose of this tutorial is to introduce the readers to the emerging technologies enabled by deep learning and to review the research work conducted in this area since the birth of deep learning in 2006 that is of direct relevance to signal processing. Future research directions will be discussed that may attract interests of and require efforts from more signal processing researchers, students, and practitioners in this emerging area for advancing signal and information processing technology and applications.

15:30~17:00 Feature-Domain, Model-Domain, and Hybrid Approaches to Noise-Robust Speech Recognition

Abstract:

Noise robustness has long been an active area of research that captures significant interest from speech recognition researchers and developers. In this lecture, we use the Bayesian framework as a common thread to connect, analyze, and categorize a number of popular approaches to noise robust speech recognition pursued in the recent past. The topics covered in this lecture include: 1) Bayesian decision rules with unreliable features and unreliable model parameters; 2) Principled ways of computing feature uncertainty using structured speech distortion models; 3) Use of phase factor in an advanced speech distortion model for feature compensation; 4) A novel perspective on model compensation as a special implementation of the general Bayesian predictive classification rule capitalizing on model parameter uncertainty; 5) Taxonomy of noise compensation techniques using two distinct axes: feature vs. model domain and structured vs. unstructured transformation; and 6) Noise adaptive training as a hybrid feature-model compensation framework and its various forms of extension.

Li Deng received the Bachelor degree from the University of Science and Technology of China (with the Guo Mo-Ruo Award), and received the Ph.D. degree from the University of Wisconsin-Madison (with the Jerzy E. Rose Award). He joined Dept. Electrical and Computer Engineering, University of Waterloo, Ontario, Canada in 1989 as an Assistant Professor, where he became a Full Professor with tenure in 1996. From 1992 to 1993, he conducted sabbatical research at Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Mass, and from 1997-1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA as a Senior Researcher, where he is currently a Principal Researcher. Since 2000, he has also been an Affiliate Professor in the Department of Electrical Engineering at University of Washington, Seattle, teaching the graduate course of Computer Speech Processing. His current (and past) research activities include automatic speech and speaker recognition, spoken language identification and understanding, speech-to-speech translation, machine translation, language modeling, statistical methods and machine learning, neural information processing, deep-structured learning, machine intelligence, audio and acoustic signal processing, statistical signal processing and digital communication, human speech production and perception, acoustic phonetics, auditory speech processing, auditory physiology and modeling, noise robust speech processing, speech synthesis and enhancement, multimedia signal processing, and multimodal human-computer interaction. In these areas, he has published over 300 refereed papers in leading journals and conferences, 3 books, 15 book chapters, and has given keynotes, tutorials, and lectures worldwide. He is elected by ISCA (International Speech Communication Association) as its Distinguished Lecturer 2010-2011. He has been granted over 40 US or international patents in acoustics/audio, speech/language technology, and other fields of signal processing. He received awards/honors bestowed by IEEE, ISCA, ASA, Microsoft, and other organizations.

He is a Fellow of the Acoustical Society of America, and a Fellow of the IEEE. He serves on the Board of Governors of the IEEE Signal Processing Society (2008-2010), and as Editor-in-Chief for the IEEE Signal Processing Magazine (2009-2011), which ranks consistently among the top journals with the highest citation impact. According to the Thomson Reuters Journal Citation Report, released June 2010, the SPM has ranked first among all IEEE publications (125 in total) and among all publications within the Electrical and Electronics Engineering Category (245 in total) in terms of its impact factor.

CALL FOR PAPERS

2011 International Conference on Speech Database and Assessments (Oriental COCOSDA 2011)

October 26-28, 2011 – National Chiao Tung University, Hsinchu, Taiwan

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

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.

Papers are invited on substantial, original and unpublished research on all aspects of speech databases, assessments and speech I/O, including, but not limited to:

Topics

- Speech databases and text corpora
- Assessment of speech input and output technologies
- Phonetic/phonological systems for Oriental languages
- Romanization of Non-Roman Characters
- Segmentation and labeling
- Speech Prosody and Labeling
- Speech processing models and systems
- Multilingual speech corpora
- Special topics on speech databases and assessments
- Standardization
- Any other relevant topics

Official Language & Publication

- The official language of O-COCOSDA2011 is English.
- The proceedings of O-COCOSDA2011 will consist of two volumes.
- The papers accepted for oral presentation will be published in a volume that will be included in IEEE Xplore and indexed by Ei Compendex.
- The papers accepted for poster presentation will be published in a companion volume.

Important Dates

Full Paper Submission	July, 1, 2011
Notification of acceptance of paper	Aug. 5, 2011
Final Manuscript	Aug. 26, 2011

現階段鑑別式聲學模型訓練研究之簡介

羅永典、陳柏琳
國立台灣師範大學資訊工程系
{ytlo, berlin}@ntnu.edu.tw

一、前言

以最大化相似度估測(Maximum Likelihood Estimation, MLE)[1]來訓練聲學模型(Acoustic Model)，在過去數十年廣爲語音辨識領域所採用；它主要是考量如何能從訓練語料中獲得統計資訊，以讓聲學模型可以代表訓練語料(換句話說，使聲學模型產生對應的訓練語料之相似度最大)。但此種訓練方法並沒有考慮語音辨識時聲學模型彼此間的關係，在調整聲學模型參數之後，雖可使相關的語音特徵落在某一個聲學模型的相似度變大，卻也可能同時讓非相關的語音特徵落在此聲學模型的相似度更大，造成辨識上的混淆。因此，近來有不少研究針對此項缺點，提出鑑別式訓練(Discriminative Training)[2][3]法則來加以改進。使用鑑別式訓練法則在進行聲學模型訓練時，不僅考慮到訓練語句的正確(或參考)轉寫(Correct or Reference Transcription)，同時也考慮到由語音辨識器對語句進行辨識後產生的與正確轉寫不同的候選詞序列(Candidate Word Sequences)，以增進訓練後聲學模型的鑑別能力。

長久以來，鑑別式訓練爲語音辨識系統中聲學模型估測與調整的重要一環，相關研究與延伸族繁不及備載，以下列三種訓練法則較具代表性：(一)最小化分類錯誤法則(Minimum Classification Error, MCE)[4][5]：考慮到訓練語句的正確轉寫與不正確轉寫(或其它候選詞序列)的分離程度；(二)最大交互資訊法則(Maximum Mutual Information, MMI)[6][7][8]：以最大化訓練語句與其對應詞正確轉寫的交互資訊；(三)最小化音素錯誤法則(Minimum Phone Error, MPE)[9][10]：目的爲最小化語音辨識器輸出(亦即所有候選詞序列)的期望音素錯誤率。這三種以不同思維出發的法則，它們背後涵意皆是描述訓練語句的正確轉寫與此語句其它候選詞序列之間的關係；而日後許多陸續被提出的鑑別式聲學模型訓練方法也都是架構於這樣的關係上。

通常在執行鑑別式聲學模型訓練時，每一句訓練語句所對應的詞圖(Word Lattice)扮演的角色不僅爲所有候選詞序列(可能包括正確轉寫)之近似假設空間(Hypothesis Space)，更是提供聲學模型參數估測過程中鑑別資訊的主要依據。近年來，許多學者對於鑑別資訊的使用有深入的研究與探討，因而有了所謂的資料選取(Data Selection)概念。像是在機器學習(Machine Learning)中，支撐向量機(Support Vector Machines, SVM)的邊際資訊概念在分類問題上有著非常成功的成效；「邊際資訊」一詞所闡述的概念是決策邊際、訓練資料分布、以及重要訓練資料選取等對一般分類問題所產生影響。這種概念近年來被推廣到語音辨識的聲學模型訓練中使用，本文將新近應用邊際資訊概念之鑑別式聲學模型訓練進一步區分爲「邊際估測法則」與「強化混淆權重」兩大類，其中「邊

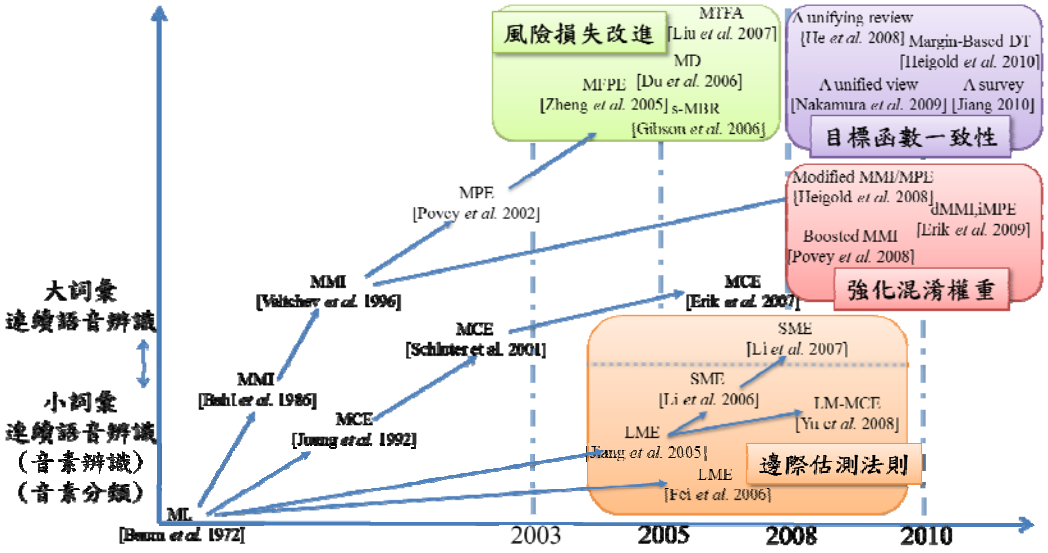

圖 1. 鑑別式聲學模型訓練相關研究之趨勢圖

際估測法則」[11][12][13][14][15][16][17]旨在邊際範圍區間內選取訓練資料；例如[18]說明如何在聲學模型訓練時，能像支撐向量機(SVM)一樣的有效使用邊際資訊。而單純地僅基於資料選取概念，卻同樣擁有邊際估測法則意涵的做法有[19][20]。此外，「強化混淆權重」[21][22][23]是在訓練目標函數(Objective Function)裡額外乘上一個權重函數以強化所選取的資料；透過這樣的方法來柔性地增強所選取資料的重要性，可視爲一種柔性邊際概念，同樣對於語音辨識率的提升有著正面的影響。鑑別式聲學模型訓練研究之演進如圖 1 所示，本文後續將著重於使用邊際資訊的兩大類鑑別式聲學模型訓練方法之回顧，亦即所謂的「邊際估測法則」與「強化混淆權重」。

由於邊際資訊該概念的應用可以架構在各種傳統鑑別式聲學模型訓練方法上，因此本文將先從目標函數一致性的觀點出發，來回顧前述三種代表性鑑別式聲學模型訓練法則(MCE、MMI、MPE)；然後，闡述當前「邊際資訊」概念應用於鑑別式聲學模型訓練方法的發展現況；最後是結論與對於鑑別式聲學模型訓練研究的未來展望。

二、鑑別式聲學模型訓練之一致性

語音解碼原則與鑑別式函數—在語音辨識的解碼原則上，大多數的作法是採用貝式決策定理(Bayesian Decision Theory)：即是在給定語句的語音特徵向量序列已知的情況下，要找出一候選詞序列的相似度(發生的機率)是最大的。利用貝式定理，我們可將此一事後機率展開，並且因爲語音特徵向量序列的事前機率不影響候選詞序列的排序，將候選詞序列在給定語音特徵向量序列的情況下之事後機率簡化成正比於聲學模型(Acoustic Model)與語言模型(Language Model)機率的乘積[24]，故稱之爲最大化事後機率(Maximum a Posteriori Probability, MAP)解碼方法。有鑑於語音辨識解碼原則爲評估候選詞序列是否可能爲較正確輸出的依據，自然地可將聲學模型與語言模型相似度(機率)的乘積，或是對數相似度(機率)的和，視作爲一種鑑別式函數(Discriminant Function)。

在聲學模型訓練時，鑑別式訓練設法將訓練語句的正確轉寫與基礎語音辨識器產生的其它候選詞序列在假設空間上的關係透過特定的目標函數來描述；在經由對此函數的最佳化過程，鑑別式訓練能因而讓訓練後的聲學模型更能分辨訓練語句的正確轉寫與其它候選詞序列的差異來達到增進辨識率之目的。一般來說，在執行鑑別式聲學模型訓練前，我們通常先使用最大化相似度估測(MLE)做爲聲學模型的基礎訓練法則以產生基礎語音辨識器，其目的是最大化聲學模型產生對應的訓練語料之相似度；而在爾後的聲學模型鑑別式訓練上，透過不同的思維來運用訓練語句本身因爲語音辨識所產生的混淆資訊，而這些資訊通常以正確轉寫與語音辨識器所產生的詞圖(內含許多候選詞序列)共同形成之所謂的假設空間(Hypothesis Space)。下面將先以三種最具代表性之鑑別式訓練法則(MCE、MMI、MPE)爲例，說明它們的聲學模型訓練目標函數的最終目的都是在於使用鑑別式函數描述正確轉寫與詞圖上其它候選詞序列在假設空間所形成的關係。同時，我們亦可以歸納出這些鑑別式訓練法則之間最主要差別在於不同層級(如語句層級、音素層級)，或者說是不同細緻程度的訓練資料選取方式[25]：

1. 最小化分類錯誤法則(Minimum Classification Error, MCE)[4][5]：以最小化分類錯誤爲基礎設計的語音辨識器(分類器)，其決策的法則可以透過分類錯誤評估(Misclassification Measure)函數來表示，分類錯誤評估函數代表的是語音辨識器對於其它非正確轉寫之候選詞序列產生的對數相似度減去對於正確轉寫產生的對數相似度。若分類錯誤評估函數輸出大於零則表示語音辨識(分類)錯誤；反之，則表示語音辨識(分類)正確。若聲學模型的訓練若能依循最小化分類錯誤評估函數的輸出來設計，則預期將會有較少的語音辨識錯誤。對於訓練語料而言，以 MCE 爲法則的鑑別式聲學模型訓練，其訓練目標在於最小化所有訓練語句的期望語句錯誤率(或最大化所有訓練語句的期望語句正確率)。

2. 最大交互資訊法則(Maximum Mutual Information, MMI) [6][7][8]：以最大化交互資訊估測法爲準則的鑑別式聲學模型訓練，主要目的是最大化所有訓練語句的語音特徵向量序列與其對應正確轉寫的交互資訊(Mutual Information)。在假設所有訓練語句擁有相同機率分布值(Uniformly Distributed)情況下，經由若干數學推導，我們可以得知 MMI 其訓練目標其實在於最大化所有訓練語句的期望語句正確率。

3. 最小化音素錯誤訓練法則(Minimum Phone Error, MPE)[9][10]：相較於最大交互資訊法則著重在訓練語句之句層次(Sentence or String Level)正確率提升，最小化音素錯誤訓練法則著重於對訓練語句較細微層級，如音素(Phone)層級的正確率提升。MPE 在訓練時，把原本以 0-1 語句層級的損失函數(Loss Function)轉變成訓練語句候選詞序列之原始音素正確個數，可以定義爲在正確轉寫上所有音素個數減去產生插入(Insertion)、刪除(Deletion)、替換(Substitution)等錯誤個數[9]，MPE 訓練目標在於最大化語音辨識器對於訓練語料所有辨識輸出(也就是所有候選詞序列，同時包括了正確轉寫)的期望音素正確率。

值得注意是，MCE 的聲學模型訓練目標函數爲所有訓練語句之期望語句正確率的總和而 MMI 爲所有訓練語句之期望語句正確率的連乘積，它們訓練目標背後意義在某種程度

鑑別式訓練法則	目標函數意義	單一訓練語句損失函數
最小化分類錯誤訓練 (MCE)	最大化所有語句期望語句正確率總和	語句層級的零壹損失函數
最大化交互資訊訓練 (MMI)	最大化所有語句期望語句正確率連乘積	語句層級的零壹損失函數
最小化音素錯誤訓練 (MPE)	最大(最小)化所有語句期望音素正確(錯誤)率總和	音素層級的損失函數

圖 2. 三種主要鑑別式聲學訓練法則及其一致性之比較

上，皆代表最大化語音辨識器對於所有訓練語句之期望語句正確率。針對三種代表性鑑別式訓練法則的比較如圖 2 所示；在[25]已證明，經過適當地數學推導(與化簡)後，上述三種代表性鑑別式訓練法則可以得到相同型式的目標函數。目前已有許多研究針對這些訓練法則的一致性有相當深入的探討[25][26][27][28][29]，亦提出許多新穎的鑑別式聲學模型訓練法則。

三、使用邊際資訊於鑑別式聲學模型訓練

基於「邊際資訊」之資料選取方法目的是爲選取對於聲學模型訓練目標函數較具重要性的訓練資料，以期能夠提升辨識率；基於「邊際資訊」概念之鑑別式聲學模型訓練方法如圖 3 所示，可大致分爲「邊際估測法則」與「強化混淆權重」兩類，其中「邊際估測法則」代表性方法的演變爲由(1)最大邊際估測法則至(2)柔性邊際估測法則，而「強化混淆權重」的演進爲由(3)增進式最大交互資訊法則開始，而後得到了以(4)整合邊際錯誤爲基礎的統一觀點：

1. 最大邊際估測法則(Large-Margin Estimation, LME)

   啓蒙於最大邊際分類器[30]，在語音辨識議題中，最大邊際估測法則的主旨在於藉由調整隱藏式馬可夫模型的參數來最大化語音辨識器的邊際，使得在訓練集中屬於正邊際一邊的訓練語句，藉此能讓語音辨識器的一般化能力(Generalization Capability)可以獲得增進。分離邊際(Separation Margin)的通常定義爲是正確轉寫與最有可能(相似度最大)的候選詞序列的對數相似度之差，而所使用訓練語句之正確轉寫的相似度必須恆大於候選詞序列。故最大邊際估測法則旨在最大化訓練語句正確轉寫與候選詞序列最小的分離邊際，可視爲是一個約束型「最小最大」最佳化問題(Constrained Minimax Optimization Problem)。而過去從事此種以最大化邊際估測爲主題的研究，主要大都

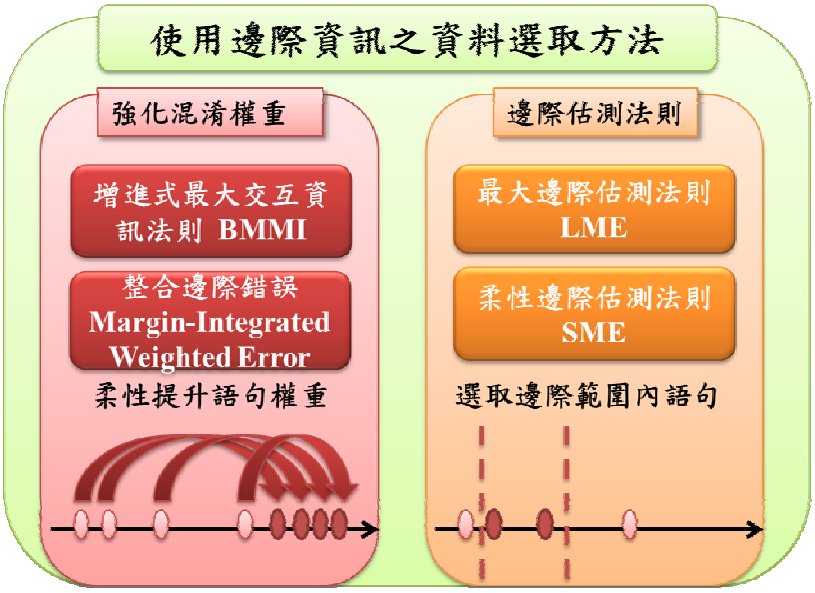

圖 3. 使用邊際資訊之資料選取方法於鑑別式聲學模型訓練的相關研究

在強調目標函式最佳化之改進。

最大邊際估測法則所強調是那些原本可以被正確辨認的訓練語句(或樣本)。而那些不滿足最大邊際估測法則限制條件(即不能正確辨認)的訓練語句將被最大邊際估測法則排除在訓練之外。但是，這樣一來所衍生的問題將是：首先，這些訓練語句對語音辨識器來說是關鍵的，應可以提供聲學模型訓練所需的鑑別資訊。其次，在理論上，經由最大邊際估測法則訓練後，訓練語句的分離邊際應變大，代表聲學模型更具有一般化能力。然而，實際上的語音辨識器多半是無法完全正確辨識所有訓練語句(也就是說辨識器對於訓練語料的辨識率尚未臻完美)，尤其是當最大邊際估測法則被應用於複雜的大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)任務上時。這樣一來，會造成存在於正邊際範圍內的樣本個數非常有限，使得於聲學模型調整後對整體辨識率的提升影響不大。於是，有所謂柔性最大邊際估測法則(Soft-Large-Margin Estimation, Soft-LME)被提出以改善此一問題[31]；它同時將所有辨識正確與錯誤的訓練語句全部納入考量，並使用半正定規劃法(Semi-definite Programming)[32]加以最佳化目標函式。對於邊際概念在區間內的資料選取影響的進一步討論，可參考[36]。

2. 柔性邊際估測法則(Soft-Margin Estimation, SME)

有鑑於前述最大邊際估測法則忽略訓練樣本所造成決定邊際附近資訊的遺失，一昧地最大化分類邊際可能使得所訓練出來的語音辨識器一般化的能力不足，因此有所謂的以柔性邊際爲訓練法則的做法被提出[28]。它的觀點是承襲於統計學習理論[30]，語音辨識器對測試語料的辨識錯誤，可視爲一種風險，已被證明會受限於一個上界，而此上界爲分類器之所謂的「經驗風險」(Empirical Risk)加上另一項「一般化量值」(Generalization or Regularization Term)。因此，欲加強語音辨識器在測試語料上的辨識

率，除了要最大化此語音辨識器的邊際外，更要同時降低其在訓練語料上的經驗風險。因此，柔性邊際估測法則便根據此精神來訓練聲學模型，期望能降低辨識器於測試語料的辨識錯誤。反觀傳統鑑別式訓練只專注於經驗風險的最小化，而忽略一般化量值；典型的最大邊際估測法則僅強調邊際的最大化，以期增加一般化能力，卻忽略了經驗風險的影響。

3. 增進式最大交互資訊法則 (Boosted MMI, BMMI)

同樣受到最大邊際概念[17][30]影響，有別於前兩者是以邊際估測作爲資料選取的範圍；增進式最大交互資訊法則以一個額外權重來選取邊際附近的資料來提升辨識率。對於每一句訓練語句而言，增進式最大交互資訊法則認爲當給定假設空間(如詞圖)上的候選詞序列與正確轉寫差別越大(錯誤越多)時，表示該候選詞序列發生的可能性越低；則嘗試將該候選詞序列相似度乘上一個額外權重來強調錯誤越多的候選詞序列。聲學模型經 BMMI 調整模型後，能讓正確轉寫與候選詞序列在假設空間拉開越多，使得邊際得以最大化；但可以預見的是，BMMI 容易造成過度訓練(Over Fitting)致使辨識率下降。

4. 以整合邊際錯誤(Margin-Integrated Weighted Error)爲基礎的統一觀點

此一觀點的提出是以論述鑑別式訓練演進之間共通性[26]爲出發，以鑑別式函數爲基礎、額外乘上了指數函數做爲增進式權重因子。其目的如同前項一要點所述是爲了強化訓練語句的假設空間(詞圖)上之混淆資訊，以帶有指數權重型態的鑑別式函數爲出發，根據不同準則所定義聲學模型訓練目標函數，並且引入了增進因子權重。經由[26]研究發現，發現透過了調整邊際參數可以達到不同的目標函數；同時亦發現了最小化音素錯誤(MPE)爲增進式最大化交互資訊(BMMI)目標函數的微分。我們可以把這樣的架構解讀爲整合邊際錯誤(Margin-Integrated Weighted Error)是一種在最小音素錯誤所形成的空間上，對於邊際的範圍不同而求得的錯誤估測。受到這層微分關係的啓發，日本學者定義了兩種在邊際空間上以微積分基本定理爲架構的聲學模型訓練目標函數，主要的目的還是著重於微(積)分區間內強調資料的不同來調整模型參數。第一種作法爲給定一段邊際段落區間，在此段落上求取最小化錯誤音素的積分故稱爲集成式最小音素錯誤訓練(Integrated MPE, *i*MPE)；第二種方法爲集成式最小音素錯誤訓練，一般化該段邊際區間，稱之爲微分式最大交互資訊訓練(Differentiated MMI, *d*MMI)。

這兩個目標函式本質上非常類似，兩者的差別僅在於後者多針對該段邊際區間一般化處理。因此，集成式最小音素錯誤(*i*MPE)訓練較適合探討於區間較大的邊際資訊段落；微分式最大交互資訊訓練適合較小的邊際資訊區間。在這兩種以不同邊際資訊區間的方法上，集成式最小音素錯誤(*i*MPE)式其實就是一般式最大交互資訊，增進式最大化交互資訊(BMMI)也是其中的一個特例。而在與一般最小錯誤音素訓練的比較上，從過去試驗可以發現：某些議題上[21][22]，增進式爲基礎的最大交互資訊結果是優於傳統最小音素錯誤訓練。對於此種情況，可以解讀爲在最小音素錯誤形成空間上，有時只針對單一邊際頂點(最小化音素錯誤訓練)的資訊，有時考量一整段的邊際資訊(增進式的最大化交互資訊訓練)。而上述兩種方法(*d*MMI 與 *i*MPE)皆透過了權重

參數來對落在邊際附近訓練資料作掌握與重要性調整，以求得較好的聲學模型。

四、使用邊際資訊於鑑別式訓練之相關應用

啓蒙於機器學習的邊際資訊概念已成功應用於鑑別式聲學模型訓練發展上；同樣地，在其他領域上亦可以見到邊際資訊概念的運用，例如語言模型、手寫辨識等。在鑑別式語言模型訓練上，全域條件式對數線性模型(Global Conditional Log-Linear Model, GCLM)[34]目標函數類似於最大交互資訊法則，在自然語言處理中也有相同應用[35][36]。同樣使用邊際資訊於訓練中，有相似於鑑別式聲學模型訓練中的增進式最大交互資訊法則(Boosted MMI)；另外，權重式全域條件式對數線性模型(Weighted Global Conditional Log-Linear Model, WGCLM) [37]可視爲 GCLM 的延伸(如同 BMMI 之於 MMI)，利用調整權重來獲得邊際資訊來增進辨識率。最小化錯誤率訓練(Minimum Error Rate Training, MERT)[38][39]則與最小音素錯誤訓練 MPE 有相同的目標函數，來最小化當使用語言模型於語音辨識時所發生的錯誤。其它，如在手寫辨識，同樣也發現引入邊際資訊[40]改善辨識率。

五、結論與未來展望

從鑑別式訓練至邊際資訊的利用，不論是訓練資料的不同細微程度（語句層級至音素層級）運用或資料的選取，在近十年中一直都被熱烈的討論，並且被應用於許多領域中獲得很好的效果。然而，在眾多理論的基礎上，許多微小的細節差異都會造成結果好壞有天南地北之不同，這將是研究上的一大挑戰。以下整理出數個鑑別式聲學模型訓練未來可能的研究方向：(1)現階段有許多的參數仍然是經由實驗人工去調整，參數值的大小與實驗語料有密切的關係，故如何使用有效的自動方式決定參數將是一項重要議題；(2)龐大的運算與冗長的訓練時間一直爲鑑別式訓練所面臨的問題；(3)由於監督式聲學模型的建立需有大量含人工標記的訓練語料集，故如何利用非監督式(Unsupervised)或半監督式(Semi-supervised)的方式來訓練聲學模型，也是另一個研究重點。

六、參考文獻

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer, “A Maximum Likelihood Approach to Continuous Speech Recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 5, No.2, pp.179-190, 1983

[2] X. He and L. Deng, *Discriminative Learning for Speech Recognition: Theory and Practice*, Morgan and Claypool Publishers, 2008.

[3] J. Keshet and S. Bengio (eds.), *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*, Wiley, 2009.

[4] B. H. Juang, W. Chou, and C. H. Lee, “Minimum classification error rate methods for speech recognition,” *IEEE Transactions Speech and Audio Processing*, Vol. 5, No. 3, pp. 257–265, 1997.

[5] W. Chou, “Minimum classification error approach in pattern recognition,” in Pattern Recognition in Speech and Language Processing, W. Chou and B.-H.Juang, Eds. Boca Raton, FL: CRC, 2003, pp. 1–49.

[6] L. R. Bahl, P. F. Brown, P. V. de Souza, and L. R. Mercer, “Maximum mutual information estimation of hidden Markov model parameters for speech recognition,” *International Conference on Acoustics, Speech and Signal Processing*, pp. 49–52, 1986.

[7] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, “MMIE training of large vocabulary recognition systems,” *Speech Communication*, Vol. 22, No. 4, pp. 303–314, 1997.

[8] Y. Normandin, *Hidden Markov models, maximum mutual information estimation, and the speech recognition problem*, Ph.D. Dissertation, McGill University, Montreal, 1991.

[9] D. Povey and P. C. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 105–108, 2002.

[10] D. Povey, *Discriminative training for large vocabulary speech recognition*. Ph.D Dissertation, Peterhouse, University of Cambridge, July 2004.

[11] H. Jiang, X. Li and C.-J. Liu, “Large margin hidden Markov models for speech recognition,’’ *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1584–1595, Vol. 14, No. 5, September 2006.

[12] J. Li, M. Yuan, and C. -H. Lee, “Approximate test risk bound minimization through soft margin estimation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol.15, no. 8, 2007.

[13] J. Li, Z. Yan, C. -H. Lee, and R. -H. Wang, “A study on soft margin estimation for LVCSR,” *IEEE workshop on Automatic Speech Recognition and Understanding*, pp. 268–271, 2007.

[14] J. Li, *Soft margin estimation for automatic speech recognition*. Ph.D. Dissertation, Electrical and Computer Engineering, Georgia Institute of Technology, July 2008.

[15] F. Sha and L. K. Saul, “Large margin Gaussian mixture modeling for phonetic classification and recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 265–268, 2006.

[16] F. Sha and L. K. Saul, “Large margin hidden Markov models for automatic speech recognition,”*Advances in Neural Information Processing Systems 19*, pp. 1249–1256, B. Schölkopf, J.C. Platt, and T. Hofmann (eds.) Cambridge, MA, 2007. MIT Press.

[17] F. Sha, *Large margin training of acoustic models for speech recognition*. Ph.D. Dissertation, University of Pennsylvania. 2007.

[18] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney. “Modified MMI/MPE: A Direct Evaluation of the Margin in Speech Recognition,” *International Conference on Machine Learning*, pp. 384-391, 2008.

[19] D. Yu, L. Deng, X. He, and A. Acero, “Large-margin minimum classification error training: A theoretical risk minimization perspective,” *Computer Speech and Language*, Vol. 22, No. 4, pp. 415–429, 2008.

[20] S. -H. Liu, F. -H. Chu, S. -H. Lin, H. -S. Lee, and B. Chen, “Training data selection for improving discriminative training of acoustic models,” *IEEE workshop on Automatic Speech Recognition and Understanding*, pp. 284–289, 2007.

[21] G. Saon and D. Povey, “Penalty function maximization for large margin HMM training,” *Annual Conference of the International Speech Communication Association*, pp. 920–923, 2008.

[22] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, “Boosted MMI for model and feature space discriminative training,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4057–4060, 2008.

[23] E. McDermott, S. Watanabe, and A. Nakamura: “Margin-space integration of MPE loss via differencing of MMI functionals for generalized error-weighted discriminative training,” *Annual Conference of the International Speech Communication Association*, pp. 224–227, 2009.

[24] F. Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, 1999.

[25] X. He, L. Deng, and C. Wu, “Discriminative Learning in Sequential Pattern Recognition — A Unifying Review for Optimization-Oriented Speech Recognition,” *IEEE Signal Processing Magazine*, Vol. 25, No. 5, pp. 14–36, September, 2008.

[26] A. Nakamura, E. McDermott, S. Watanabe, and S. Katagiri, “A unified view for discriminative objective functions based on negative exponential of difference measure between strings,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1633–1636 , 2009.

[27] H. Jiang, “Discriminative training for automatic speech recognition: A survey,” *Computer and Speech, Language*, Vol. 24, No. 4, pp. 589–608, 2010.

[28] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, “Investigations on errorminimizing training criteria for discriminative training in automatic speechrecognition,” *Annual Conference of the International Speech Communication Association*, pp. 2133–2136, 2005.

[29] R. Schlüter, W. Macherey, B. Müller, and H. Ney, “Comparison of discriminative training criteria and optimization methods for speech recognition,” *Speech Communication*, Vol. 34, pp. 287–310, May 2001.

[30] V. Vapnik, *The nature of statistical learning theory.* New York: Springer-Verlag, 1995.

[31] H. Jiang and X. Li, “Incorporating training errors for large margin HMMs under semi-definite programming framework,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 629–632, 2007.

[32] X. Li and H. Jiang, “Solving Large Margin HMM Estimation via Semi-definite Programming,” *IEEE Trans. on Audio, Speech and Language Processing*, pp.2383-2392, Vol. 15, No. 8, November 2007.

[33] J. Li, M. Yuan, and C. -H. Lee, “Soft margin estimation of hidden Markov model parameters,” *Annual Conference of the International Speech Communication Association*, pp. 2422–2425, 2006.

[34] B. Roark, M. Saraclar, and M. Collins, “Discriminative n-gram language modeling,” *Computer Speech & Language*, Vol. 21, No. 2, pp. 373–392, 2007.

[35] A. Ratnaparkhi, S. Roukos, and R. T. Ward, “A maximum entropy model for parsing,” *International Conference on Spoken Language Processing*, pp. 803–806, 1994.

[36] M. Johnson, S. Geman, S. Canon, Z. Chi, and S. Riezler, “Estimators for stochastic “unification-based” grammars,” *Annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 535–541, 1999.

[37] T. Oba, T Hori, and A. Nakamura, “A comparative study on methods of weighted language model training for reranking LVCSR n-best hypotheses,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5126–5129, 2010.

[38] F. J. Och, “Minimum error rate training in statistical machine translation,” *Annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 160–167, 2003.

[39] K. Akio, O. Takahiro, H. Shinichi, S. Shoei, I. Toru, and T. Tohru, “Discriminative rescoring based on minimization of word errors for transcribing broadcast news,” *Annual Conference of the International Speech Communication Association*, pp. 1574–1577, 2008.

[40] P. Dreuw, G. Heigold, and H. Ney, “Confidence and Margin-Based MMI/MPE Discriminative Training for Offline Handwriting Recognition,” *International Journal on Document Analysis and Recognition*, accepted for publication.

[41] D. Yu, L. Deng, X. He, and A. Acero, “Large-margin minimum classificationerror training for large-scale speech recognition tasks,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1137–1140, 2007.

[42] G. Heigold, P. Dreuw, S. Hahn, R. Schlüter, and H. Ney.“Margin-Based Discriminative Training for String Recognition.” *IEEE Journal of Selected Topics in Signal Processing - Statistical Learning Methods for Speech and Language Processing*, volume 4, number 6, pages 917-925, Aachen, Germany, December 2010.

[43] E. McDermott, T. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, “Discriminative training for large vocabulary speech recognition using minimum classification error,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15. No. 1, pp. 203–223, 2007.

[44] W. Macherey. *Discriminative Training and Acoustic Modeling for Automatic Speech Recognition*. PhD Thesis, Aachen, Germany, March 2010.