第2章　文獻回顧

2.1　現階段英文語音辨識研究內容

語音辨識系統的發展，從 1952 年美國貝爾實驗室發展的小詞彙獨立數字辨識，之後隨著演算法、電腦速度的進步，由數字單詞演進到任意口語連續語句的大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)系統，其中大詞彙連續語音辨識系統，從 1990 年初開始發展至今已有 15 年以上的歷史，目前較著名國外發展語音辨識器的主要學術單位、科技公司與機構如表 2-1 所示：

表 2-1　國外發展語音辨識器之學術單位、科技公司與機構

1	美國麻薩諸塞州 BBN 科技公司
2	美國 IBM 華生(T.J. Watson)研究中心
3	英國劍橋大學電機系(Cambridge University Engineering Dept.)
4	美國卡內基美隆大學電腦科學學校 (Carnegie Mellon University - School of Computer Science)
5	美國麻薩諸塞州 Dragon Systems 科技公司
6	法國 LIMSI-CNRS (Centre National de la Recherche Scientifique)機構
7	美國加州 SRI 國際機構之語音科技和研究實驗室 (Speech Technology and Research Laboratory)
8	美國 AT&T 實驗室
9	美國密西西比州 MsState – ISIP 學術機構
10	美國微軟(Microsoft)科技公司

語音辨識器，可能因為訓練語料、測試語料、詞典的不同而導致不同的語音辨識率，故美國國家標準技術局(National Institute of Standards and Technology,

NIST) [NIST 2007]自 1996 年起，提供標準的對話電話語料(Conversational Telephone Speech, CTS)，並每年舉辦語者辨識評比(Speaker Recognition Evaluation, SRE)競賽。NIST SRE 是目前國際公認的語者辨識評比基準(Benchmark)。評比項目包含有：

1. 單一語者偵測(Single-Speaker Detection)：判斷一段語音是否為假設語者(Hypothesized Speaker)的語音，相當於語者驗證(Speaker Verification)。

2. 雙語者偵測(Two-Speaker Detection)：從一段二人對話中判別假設語者是否在其中。

3. 語者分段(Speaker Segmentation)：經由找出一段語音中各語者的聲音段落，進而將這些聲音段落依據語者分群。

4. 語者追蹤(Speaker Tracking)：將一段語音中屬於某一假設語者的段落一一標出。

從 2002 年 3 月開始，美國國際電腦科學組織(International Computer Science Institution, ICSI)的語音研究團隊著手進行美國國防部先進研究計畫機構(DARPA)委託的 EARS(Effective Affordable Reusable Speech-to-text Program)計畫[EARS]，計畫目的為在接下來五年內，能讓多語辨識器的辨識率達到 10%以下的詞錯誤率(Word Error Rate, WER)，而主要的辨識語料為廣播新聞(Broadcast News, BN)與人類對話(Conversational Speech)語料。EARS 計畫主要包含有兩個子計畫，分別為大量轉寫文字(Rich Transcription)與創新方法(Novel Approaches)之研究，其中大量轉寫文字之研究主要為設計適當的評比語料，供辨識器研究者做測試，例如 RT03、RT04 等語音評比語料。

語音辨識之訓練語料的來源有很多，美國語言資料協會(Linguistic Data

Consortium, LDC)[LDC]為國際非營利組織之語言相關的教育研究及科技發展機構，提供有關於 Switchboard、Switchboard Cellular 及 Callhome 等語音語料。在 EARS 計畫中，就有幾千小時的語音資料來自於 LDC，這些語料被稱為費雪集合 (Fisher Collection)。

本論文將於 2.1.1 至 2.1.3 擇要簡介其中 BBN、IBM、與劍橋大學三家著名研究機構目前在英文大詞彙連續語音辨識器之發展現狀。

2.1.1 美國 BBN 科技公司

美國 BBN 科技公司於 2005 年，與法國 LIMSI 機構共同發表的「2004 BBN/LIMSI 英文對話電話語料辨識系統」(2004 BBN/LIMSI English Conversational Telephone Speech Recognition System)[Nguyen *et al.* 2005; Prasad *et al.* 2005; Matsoukas *et al.* 2002; Colthurst *et al.* 2000]，此系統需 20 倍時間(Real-time)，評比語料為 RT04，詞錯誤率(WER)為 13.5%。

1. 語料來源：

   訓練聲學模型的聲音語料為 2,300 小時的費雪集合中 Switchboard I and II、CallHome、Cellular 語料；而語言模型的文字語料為詞彙量 27M 同聲學模型的對話電話語料、260.3M 的廣播新聞文字語料、115.9M 的 CNN 文字語料與 525M 經由美國華盛頓大學蒐集的網路文字語料。

2. 前端語音特徵擷取：

   利用 Vocal Track Length Normalization [Molau *et al.* 2001]，解決聲腔長度因人而異的變異性，此為語者正規化(Speaker Normalization)的前端處理技術，目的為調整測試語音的線性頻率尺度，符合原本訓練語料之頻率特性。而此系統利用感知線性預測技術(Perceptual Linear Prediction

Coefficients, PLPC)擷取出 14 維倒頻譜係數、並加入第 15 能量維、音框重疊為 10ms，且用倒頻譜平均消去法(Cepstral Mean Subtraction, CMS)增強語音特性並減少噪音干擾，最後各取一階、二階與三階導數相加而成共 60 維度的語音特徵向量。

3. 聲學模型訓練：

系統之聲學模型分別訓練語者獨立(Speaker-Independent)與語者調適(Speaker-Adaptive)之模型，而此兩種模型都會利用最小音素錯誤(Minimum Phone Error, MPE)[Povey 2004]訓練法做模型訓練。

A. 最大相似度-語者獨立模型訓練(Maximun Likelihood- Speaker Independent model, ML-SI model)：

將前端擷取之 60 維特徵向量，利用異質性線性鑑別分析法(Heteroscedastic Linear Discriminant Analysis, HLDA) [Kumar 1997] [Kumar and Andreou 1998]作投影、降維成 46 維特徵向量，之後利用期望最大(Expectation Maximization, EM) [Dempster *et al.* 1977]演算法來訓練三個語者獨立三連音素聲學模型，分別是：

I. 連結狀態混合(State Tied Mixture, STM)模型。

II. 分群連結狀態混合(State Clustered Tied Mixture, SCTM)模型。

III. 詞間分群連結狀態混合(Cross-word SCTM)模型。

B. 最大相似度-異質性線性鑑別分析-語者調適模型訓練(Maximun Likelihood- Speaker-dependent HLDA Transforms, ML-HLDA-SAT)：

同樣將原本前端擷取之 60 維特徵向量，利用異質性線性鑑別分析

法降維成 46 維特徵向量，並利用條件式最大化相似度性線迴歸(Constrained Maximum Likelihood Linear Regression, CMLLR) [Gunawardana & Byrne 2001]來估測鑑別式線性轉換矩陣，之後利用最大化相似度(Maximum Likelihood)法訓練語者調適模型。

4. 語言模型訓練：

此系統利用各種來源的文字語料，並且利用改良式 Witten-Bell 語言模型[Witten *et al.* 1991]平滑化方法，以解決機率值為 0 的情況。

5. 解碼步驟：

系統辨識器核心為名 Byblos 的辨識器[Colthurst *et al.* 2000]，此辨識器須經過多階段(Multi-pass)的辨識過程，利用多階段不同種類的模型所產生的可能結果，以產生最大可能出現的語音訊號對應詞序列。解碼過程有：

A. 語者獨立(Speaker-Independent)解碼階段：

I. 快速對應(Fast Match)：利用向前(Forward)演算法對 STM 三連音素聲學模型、與詞二連語言模型做統計運算。

II. 利用向後(Backward)演算法對 SCTM 詞內(Within-word)五連音素(Quinphone)聲學模型、詞三連語言模型做統計運算，以找出 *N* 條最有可能的出現詞序列(*N*-best Hypotheses)。

III. 利用 Cross-word SCTM 聲學模型、詞四連(Fourgram)語言模型，對此 *N* 條可能出現的詞序列重算分數(Re-scoring)和重新排列(Re-rank)，最後出現機率最大的(Top 1)即為語者獨立階段的辨識結果。

B. 語者調適(Speaker-Adaptive)解碼階段：

利用語者獨立解碼階段最後產生的假設語句，訓練語者調適聲學模型，再重複相同解碼過程，找出新的假設語句。

C. 最後解碼階段利用大量的統計迴歸組合(Regression Classes) [Leggetter & Woodland 1995; Gales & Woodland 1996]使用對語者調適階段產生的假設語句再來調適聲學模型，產生最後辨識結果。

BBN 結合 LIMSI 的辨識模型，將不同的獨立、調適聲學模型與不同連結的 *N* 連語言模型做結合，最後利用 ROVER 方法[Fiscus 1997]，對每種模型產生最大假設語句中的部份小段(Segmentation)詞句，依據不同權重值做挑選(Voting)，產生組合成一條機率最大的詞序列。

6. 系統架構圖：

如圖 2-1 所示，2004 BBN/LIMSI 英文對話電話語料辨識系統混合比較 BBN、LIMSI 不同解碼階段(B1、B2、B3、B4、L1、L2、L3)產生的不同假設語句，利用 ROVER(圖 2-1 中的 R1 及 R2)挑選出機率最大的詞序列，其中各階段設定如下：

A. B1：使用感知線性預測特徵(PLP)擷取、MPE 的三連音素 STM 聲學模型(365K 個高斯混合分布)與代入詞二連語言模型。

B. B2：使用感知線性預測特徵擷取、MPE 的五連(Quinphone)音素 SCTM 聲學模型(843K 個高斯混合分布)與代入詞三連語言模型。

C. B3：使用感知線性預測特徵擷取、MPE 的詞間五連(Crossword Quinphone)音素 Crossword SCTM 聲學模型(855K 個高斯混合分布)與代入詞四連語言模型。

D. B4：使用梅爾倒頻譜係數特徵擷取、MPE 的詞間五連音素 Crossword SCTM 聲學模型(708K 個高斯混合分布)與代入詞四連語言模型。

E. L1：使用兩個性別相依(Gender-dependent)聲學模型(48 個單連音素、包含 30K 個狀態連結(Tied States)，其中每個狀態有 32 個高斯混合分布) ，並使用詞三連與詞四連插補過的語言模型(Interpolated Language Model)。

F. L2：將單連音素減少成為 38 個，28K 個內文音素共有 30K 個連結狀態 (Tied States)，並代入詞三連與詞四連插補過的語言模型。

G. L3：使用最大事後機率演算法調適性別相依聲學模型(43K 個內文相依(Context-dependent)音素，共有 31K 個連結狀態)。

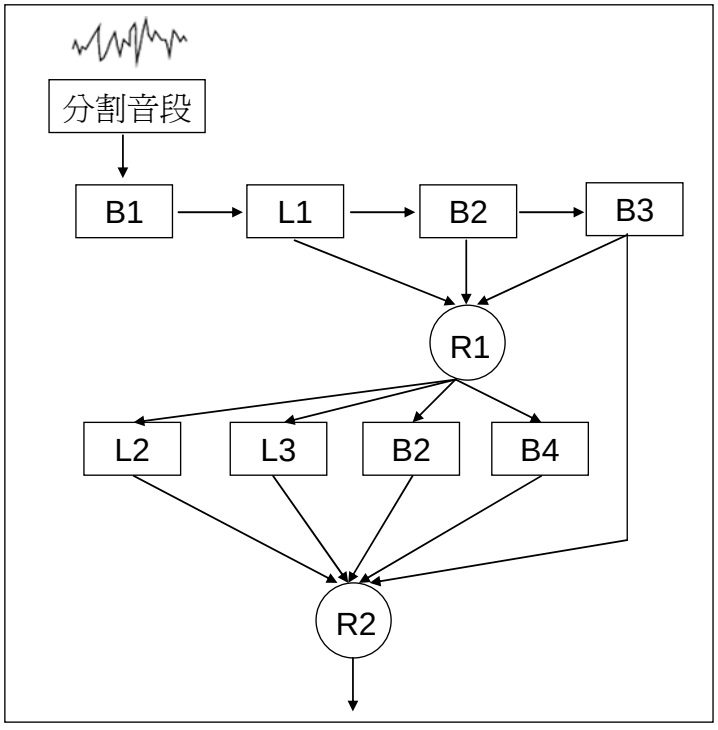

圖 2-1　2004 BBN/LIMSI 英文對話電話語料辨識系統架構圖

2.1.2 美國 IBM 華生研究中心

美國 IBM 華生研究中心曾於 2005 年發表的「IBM 2004 英文對話電話語料辨識系統」(IBM 2004 Conversational Telephony System for Rich Transcription)[Soltau *et al.* 2005]，此系統需 10 倍時間，評比語音語料為 RT04，詞錯誤率為 15.2%。

此系統特別之處是結合特徵最小音素錯誤 (Feature Minimum Phone Error, fMPE)[Povey *et al.* 2005]特徵擷取法，與最小音素錯誤(MPE)聲學模型訓練方法。實驗結果如表 2-2 所示，由表中數據發現，結合 fMPE 與 MPE 法，比單用 ML、fMPE 或 MPE 方法可得較低的詞錯誤率。

表 2-2　RT03 評比語料的詞錯誤率

	ML	MPE	fMPE	fMPE+MPE
詞錯誤率	22.1	20.6	20.2	19.2

1. 語料來源：

   聲學模型的語音訓練語料為共 2,100 小時的 Fisher parts1-7、Switchboard-1、BBN/CTRAN Switchboard-2、Switchboard Cellular、Callhome English 語料。語言模型的文字訓練語料來源有很多，有 SWB (LDC transcripts of Switchboard-1, Switchboard Cellular and Callhome English) 、 BBN (BBN/CTRAN transcripts of Switchboard-2)、BN (廣播新聞語料)、FSH(Fisher Collection Parts1-7)、UW191(由華盛頓大學蒐集的 191M “Switchboard-like”網路語料)、UW175(由華盛頓大學蒐集舊版的 175M “Fisher-like”語料)、UW525(由華盛頓大學蒐集新版的 525M “Fisher-like”語料)。

2. 前端語音特徵擷取與聲學模型訓練：

此系統的聲學模型分三類如下：

A. 語者獨立-對角化共變異矩陣-感知線性預測技術(Speaker Independent- Diagonal Covarience- PLP, SI-DC-PLP)：

語音訊號特徵擷取利用感知線性預測(PLP)技術、線性鑑別分析(Linear Discriminant Analysis, LDA)加上最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)[Gopinath 1998][Saon *et al.* 2000]與倒頻譜正規化法(Cepstral Mean and Variance Normalization,CMVN) [Viikki *et al.*1998]做正規化。語者獨立模型共有 150K 個 40 維的對角化共變異高斯混合分布與利用 MPE 法訓練的 8K 個英文五連音素狀態。

B. 語者調適-共變異矩陣-fMPE (Speaker Adaptation- Full Covarience-fMPE, SA-FC-fMPE)：

語音訊號特徵利用 VTLN 及感知線性預測技術、fMPE 與 LDA 配合 MLLT 及倒頻譜正規化法做擷取。語者調適模型共有 143K 個 39 維高斯混合分布與利用最大交互資訊法(Maximum Mixture Information, MMI)和語音特徵最大化相似度線性迴歸(fMLLR)法訓練的 7.5K 個五連音素狀態。

C. 語者調適-對角化共變異矩陣-fMPE+MPE (Speaker Adaptation-Diagonal Covarience- fMPE+MPE, SA-DC-fMPE+MPE)：

語音訊號特徵利用 VTLN 及感知線性預測技術、fMPE 與 LDA 配合 MLLT 及倒頻譜正規化法做擷取。語者調適模型共有 849K 個 39 維對角化高斯混合分布，與利用 fMPE 和 MPE 及 fMLLR 法和訓練的 22K 個七連(Septaphone)音素狀態。

3. 語言模型訓練：

此系統利用各種來源的語言資料，在實作時會依據與語音相關的程度不同權重(Weight)值做詞四連語言模型插補調適，並且利用改良式 Kneser-Ney 語言模型平滑化方法[Ney *et al*. 1994]，以解決語言模型機率值為 0 的情況。

4. 解碼步驟：

A. 對語音訊號做前端切刻(Segmentation)，區分出語音段(Speech)與非語音段(Non-speech) 。

B. 將語音訊號經由第一階段 SI.DC.PLP 所使用的語者獨立聲學模型做解碼。

C. 利用 VTLN 技術對與音訊號做語者正規化，將辨識結果再次經由 SA.FC.fMPE 調適模型做解碼。

D. 將辨識結果經由 SA.DC.fMPE+MPE 辨識模型，產生最後詞圖(Lattice) 。

E. 再經由語言模型分數做重新評分(Rescoring)，產生模糊網路(Confusion Network) [Mangu *et al*. 2000]，產生事後機率最大詞序列。

5. 系統架構圖：

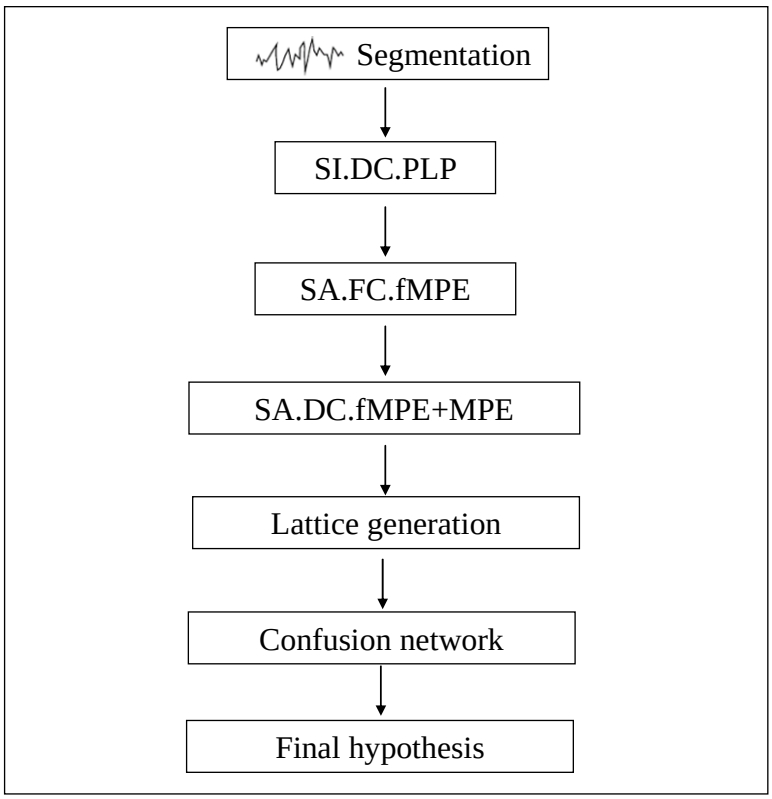

圖 2-2　IBM 2004 英文對話電話語料辨識系統架構圖

2.1.3　英國劍橋大學

英國劍橋大學曾於 2003 年發表的「2003 CU-HTK 英文對話電話語料辨識系統」(2003 CU-HTK Conversational Telephone Speech Transcription System)[Evermann *et al*. 2003]，此系統需 190 倍時間，評比語料為 RT03，詞錯誤率為 20.7%。而在 2004 年加入上達數千小時的訓練語料，發表新版的「2004 CU-HTK 英文對話電話語料辨識系統」(2004 CU-HTK Conversational Telephone Speech Transcription System)[Evermann *et al*. 2004]，系統時間為 10 倍時間，評比語料為 RT03，詞錯誤率為 17%。

1.　語料來源：

2003 CU-HTK 所用聲學模型語音訓練語料有 363 個小時的 Switchboard I、

Call Home English、Switchboard Cellular 語料，而語言模型文字訓練語料為 427M 詞彙量的廣播新聞文字語料、對話文字語料與從網路蒐集的 62M 詞彙量文字語料。

2004 CU-HTK 用到更多訓練語料，聲學模型部份為 2,180 小時、語言模型是 1,044M 詞彙量的 Fisher data。

2. 前端語音特徵擷取：

前端語音訊號利用 HLDA 做降維，並加入 VTLN 做語者正規化及倒頻譜正規化(CMVN)。

3. 聲學模型訓練：

利用最大相似度與 MPE 三連音素聲學模型，其中有 6K 個連結狀態，每個狀態中有 28 個高斯混合分布。

4. 語言模型訓練：

此系統利用各種來源的語言資料，在實作時會依據與語音相關的程度不同權重(Weight)值做詞四連語言模型插補調適，並且利用 Kneser-Ney 與 Good-Turing 語言模型平滑化方法，來找出沒有出現詞的語言模型機率。

5. 解碼步驟：

此系統需三個主要階段的解碼過程。

A. 第一階段(Pass 1,P1)使用 MPE 三連音素聲學模型、前端語音訊號利用 HLDA 做降維、及使用詞四連語言模型分數產生一條詞序列。

B. 將此詞序列利用 VTLN 找出語者正規化與倒頻譜正規化找出語音特徵向量。

C. 第二階段(Pass 2,P2)使用同第一階段的 MPE 三連音素聲學模型、VTLN 與 HLDA 前端特徵向量與語言模型產生較小的詞圖(Lattices)。

D. 第三階段(Pass 3,P3)使用詞圖最大相似度線性迴歸法(Lattice MLLR)[Uebel *et al*.2001]做調適，並使用詞四連插補法語言調適模型產生新詞圖。

E. 將第三階段產生的詞圖於 P4.1~P4.n(圖 2-3)與 P5.1~P5.n(圖 2-3)各階段，使用三連音素、五連音素聲學模型與語言模型重新評分，尋找機率最大(Top 1)詞序列。

F. 最後利用模糊網路結合(Confusion Network Combination, CNC) [Mangu *et al*. 2000]技術找出最終機率最大的辨識詞序列。此多重解碼步驟讓辨識結果更準確。

6. 系統架構圖：

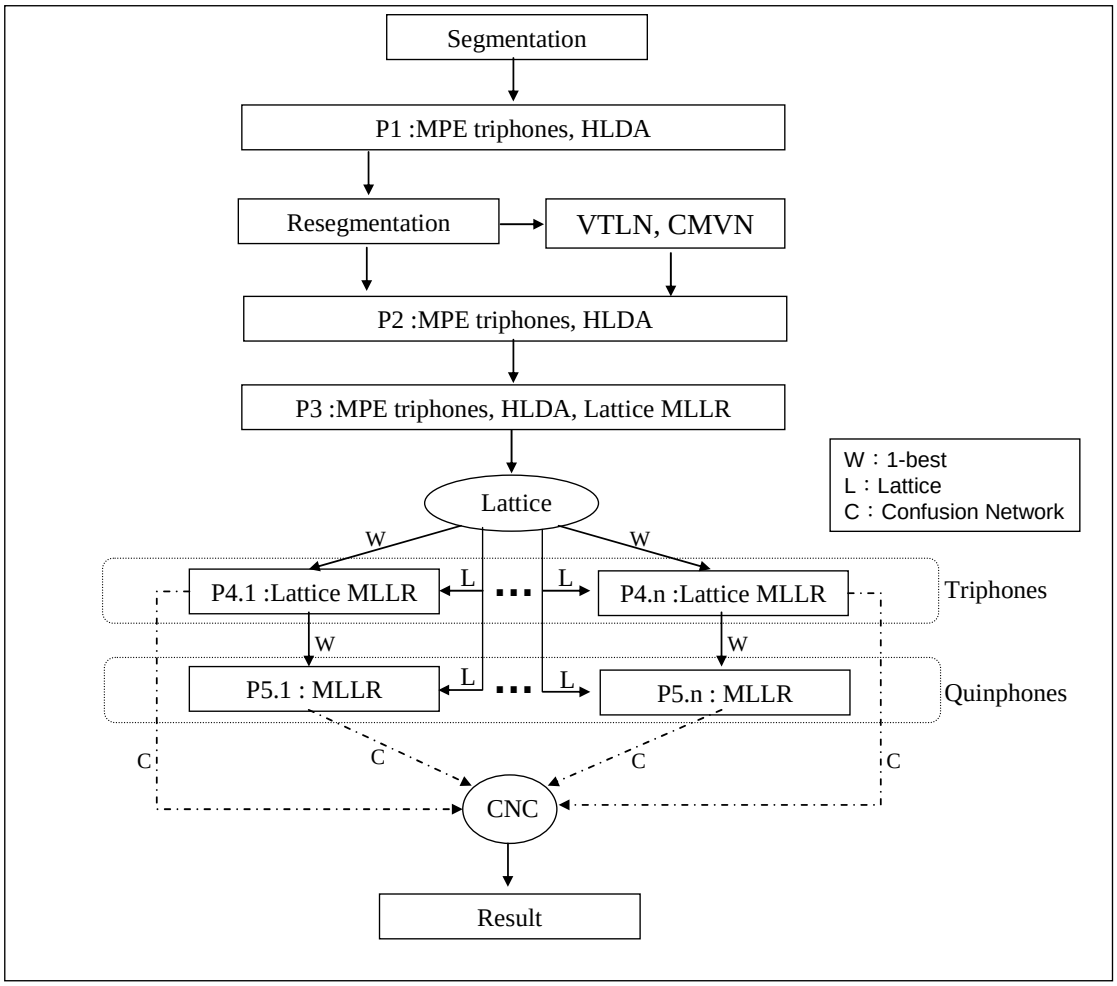

圖 2-3　2003 CU-HTK 英文對話電話語料辨識系統架構圖

2.1.4　綜合討論

此節綜合討論國外三家機構現階段大詞彙連續語音辨識器之內容特色，如下表 2-3 所示。

表 2-3　國外三家現階段大詞彙連續語音辨識器之內容特色

	BBN	IBM	CU
系統名稱	2004 BBN/LIMSI 英文對話電話語料辨識系統	IBM 2004 英文對話電話語料辨識系統	2004 CU-HTK 英文對話電話語料辨識系統
執行時間	20RT	10RT	10RT
評比語料	RT04	RT04	RT03
詞錯誤率	13.5%	15.2%	17%
聲學語料	2,300(時)	2,100(時)	2,180(時)
前端特徵	VTLN PLP + CMS HLDA+MLLT	VTLN PLP + CMVN +LDA fMPE + LDA+MLLT	VTLN HLDA+ CMVN
聲學模型	1. ML-SI (+HLDA) I. STM II. SCTM III. Cross-word SCTM 2. ML-HLDA-SAT (+MLLT)	1.SI.DC.PLP 2.SA.FC.fMPE 3.SA.DC.fMPE+MPE	MPE + Triphone Quinphone
語言模型	Witten-Bell+ Interpolated LM	Kneser-Ney + Interpolated LM	Kneser-Ney + Good-Turing+ Interpolated LM
解碼步驟	1. ML-SI : I.Triphone + Bigram II.Within-word Quinphone + Trigram III.Cross-word Quinphone+Fourgram 2. ML-HLDA-SAT 3. Regression Classes	1. SI.DC.PLP: Quinphone + Fourgram 2. SA.FC.fMPE: Quinphone + Fourgram 3. SA.DC.fMPE+MPE: Septaphone + Fourgram	1.Triphone+Fourgram 2.Quinphone+Fourgram 3.Lattice MLLR

觀察表 2-3 可得知，三家研究機構利用上千小時之聲學模型訓練語料，與豐富的大量語言模型訓練語料訓練模型，及前端語音特徵擷取技術、模型調適技術，增加語音辨識率。在辨識過程中為多階段(Multi-pass)之辨識，並分別訓練語者獨立

與語者調適之模型。三家機構之辨識器之詞錯誤率介於 13.5%至 17%之間。

2.2 聲學模型音素單位相似度測量

語言是亙古以來人類仰賴彼此溝通、了解最自然快速的重要工具，目前世界上有多達數千種的不同語言。本國人(Native)在學習非本國人(Non-Native)之語文時，可能因擁有本國人或第一語言之發音特性或習慣，故在學習非本國語言時會產生不同的發音腔調或變異(Variability)。如本國人說非本國語言，容易將某二個音混淆，則在本國人所說非本國語之聲學模型上，測量兩個音素於聲學模型中的音素相似度(Phoneme Similarity)，如果距離愈近代表相似程度愈高，可以合併來思考。

本節介紹在聲學模型尋找發音變異特性。主要有兩種方向，分別為資料導向方法(Data Driven Methods)與以知識為基準之方法(Knowledge Based Methods)。

2.2.1 資料導向方法

此方法為由上而下(Top-down)運算每個音素模型間的距離(Distance)，利用距離值當作相似度的比較，如果距離愈近代表相似程度愈高，可以合併來思考，以彌補語音訓練資料量的不足，運算此距離的運算方式有如下幾種[Le *et al.* 2006]：

1. HMM 距離 [Kohler *et al.* 1996]：

   如假設兩個模型 w_i 與 w_j ，與分別對應的語音特徵向量序列為 O_i 與 O_j ，則 HMM 距離算法為式(2-1)所示，即分別算出觀察值在模型中的 log 機率值(Log Likelihood)，再依對稱(Symmetric)取平均的方法運算其距離。

$$D(w_i, w_j) = \log p(O_i \mid w_i) - \log p(O_i \mid w_j)$$
$$D(w_j, w_i) = \log p(O_j \mid w_j) - \log p(O_j \mid w_i) \quad (2\text{-}1)$$
$$D(w_i, w_j) = \frac{1}{2}\left(D(w_i, w_j) + D(w_j, w_i)\right)$$

2. Kullback-Leibler 距離：

兩個機率密度函數 $p(o)$ 和 $q(o)$ 的相關熵值(Relative Entropy)可以式(2-2)所表示，用來表示兩個機率分布的差異程度，其中 o 代表語音特徵向量：

$$D(p \| q) = \int p(o) \log \frac{p(o)}{q(o)} do \quad (2\text{-}2)$$

3. Bhattacharyya 距離[Brian Mak *et al.*1996]：

如式(2-3)來測量兩個機率密度函數 $p(o)$ 和 $q(o)$ 的距離，其中 o 代表語音特徵向量：

$$D(p, q) = \int \sqrt{p(o)q(o)} do \quad (2\text{-}3)$$

4. Euclidean 距離[Sooful *et al.* 2001]：

如式(2-4)來測量兩個機率密度函數分布 i、j 的距離，其中 μ_i 、μ_j 與 σ_i 、σ_j 分別代表平均值與標準差，V 表向量串列的維度(Dimension) [Young *et al.* 2006]：

$$D(i, j) = \frac{1}{V} \sum_{k=1}^{V} \left[\frac{(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik}\sigma_{jk}} \right]^{1/2} \quad (2\text{-}4)$$

5. 模糊矩陣(Confusion Matrix) [Beyerlein *et al.* 1999] [Bayeh *et al.* 2004]：

建立模糊矩陣來表示兩個音素之間的模糊機率(Likelihood)。計算語料之正確文字標記與辨識結果之最小編輯距離(Levenshtein Distance) ，找出每個音素 M「取代」(Substitution)成 $N_1 ... N_k$ 的次數正規化值，以 A_{MN_i} 表示，其中 $i = 1...k$

且 $\sum_{i=1}^{k} A_{MN_i} = 1$。本論文欲利用此法，尋找語料辨識結果與正確解答之間的音素差異性，修改訓練聲學模型前的音素狀態連結規則，與將變異性加入於語音辨識階段，觀測語音辨識率。以圖 2-4 為例，代表統計辨識結果中，正確標記之單連音素「eh」易被取代為「ae」與「aa」的次數經正規化後分別為 0.3 與 0.7。

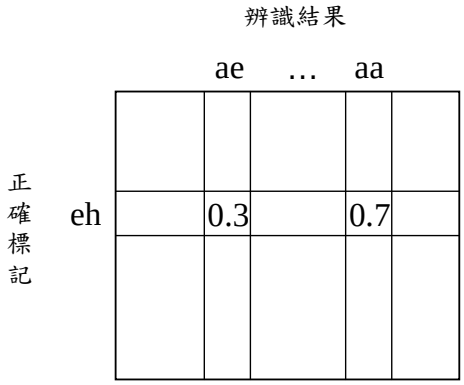

圖 2-4　模糊矩陣示意圖

2.2.2　以知識為基準之方法

傳統上以知識為基準的方法用來尋找原始語言(Source Language)對應目標語言(Target Language)的最佳音素[Beyerlein *et al*. 1999；Schultz *et al.* 2001]，但是並無比較兩個音素之間的相似度，在[Le *et al.* 2006]論文中提出一個新的以知識為基準方法來計算兩個音素之間的相似度，此法為一由下而上(Bottom-up)演算法，由兩個步驟組成：

1. 使用階層圖(Hierarchical Graph)做由上而下的分類：

   利用國際音素標準(International Phonetic Alphabet, IPA)所訂定的音素規則，如子音(Consonant)、母音(Vowel)分類，子音分類中又可分為破裂音

(Plosive)、雙唇音(Bilabial)，使用者自訂不同規則所占訓練資料量應有的比例，產生 k 層(Layer)分配，將每層訂定比例值，以 G_i 表示，代表使用者自訂第 i 層(i=0…k-1)的比例值，愈下層代表分配比例愈細緻，但內容所佔比重愈小，故 G 值隨之遞減。

2. 由下而上音素距離估測：

使用步驟 1 的定義，建立好規則階層圖型後，如果想找音素 s 與 t 之音素距離，首先觀察音素 s 與 t 是否有出現在階層的葉節點(Leaf Node)，如果有則從階層圖回朔、由下往上尋找兩音素最近的相交母節點(Parent Node)，觀察母節點所在的階層 G 值，當作兩個音素的距離，如式(2-6)：

$$d(s,t) = G_i \tag{2-6}$$

如圖 2-5 為例，某使用者自訂分層規則，將階層圖分五層，分別為層 0 至層 4，G_0 至 G_4 值分別為 0.9、0.45、0.25、0.1、0.0。此時欲找葉節點音素 p 與 m 之相似度，由下往上回朔觀察兩音素之相交母節點於層 2 之「Bilabial」(爆破音)，故兩音素之相似度設定為 0.25。

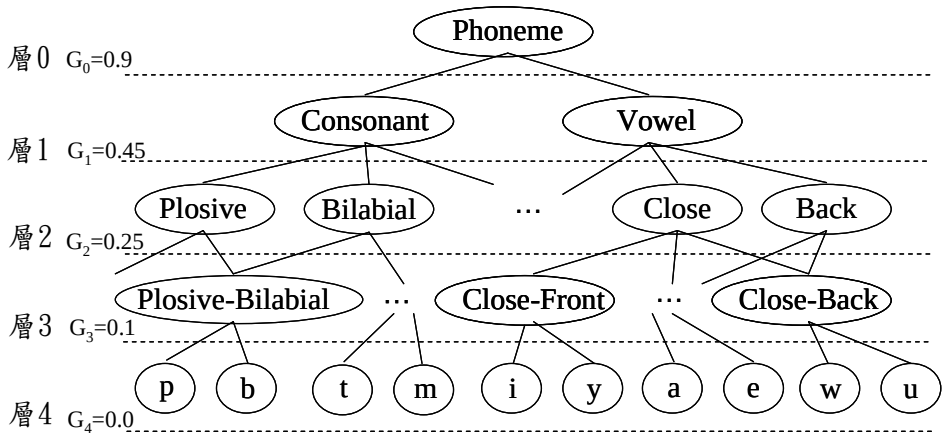

圖 2-5　音素相似度階層圖範例