國　立　交　通　大　學

資訊科學與工程研究所

碩士論文

電視新聞語音檢索之研究

The Study of Spoken Document Retrieval on TV news

研究生　：　蔡富評

指導教授: 傅心家　教授

中華民國九十五年七月

電視新聞語音檢索之研究

The Study of Spoken Document Retrieval on TV news

研 究 生：蔡富評　　Student：Fu-Ping Tsai

指導教授：傅心家 教授　　Advisor：Prof. Hsin-Chia Fu

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis Submitted to Institute of Computer Science and Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer and Information Science

July 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年七月

電視新聞語音檢索之研究

研究生:蔡富評　　指導教授: 傅心家 教授

國立交通大學資訊科學與工程研究所

摘要

語音資訊檢索主要是研究如何對大量的多媒體資訊(如廣播新聞),利用語音辨識技術,以自動的方式對於其內含的語音資訊建立起全文索引與檢索的機制。本篇論文主旨在針對台灣廣播新聞,在建立語音檢索的機制之前,需要針對電視新聞節目建立起自動新聞分析的系統,以偵測出新聞節目中主播的位置並切割新聞故事的問題作探討研究。近來許多新聞節目中主播音段常有明顯的背景音樂,為了正確的偵測出沒有背景音樂的主播音段,論文中提出結合 BIC 語者分段與分群以及語者識別的技術來偵測新聞中沒有背景音樂的主播音段。我們以台灣有線東森新聞台的新聞節目進行主播偵測的實驗,驗證所提的方法能正確偵測出沒有背景音樂的主播音段,論文最後更進一步實作語音音節辨識並且成功建立起以音節為索引特徵之電視新聞語音檢索系統。

The Study of Spoken Document Retrieval on TV news

Student: Fu-Ping Tsai Advisor: Prof. Hsin-Chia Fu

Institute of Computer Science and Engineering

National Chiao Tung University

Abstract

This thesis mainly describes broadcast news retrieval system for Mandarin Chinese. First, we need to construct automatically news analysis system to detect anchor segments in news program. Recently, we observed some anchor segments that have background music in many news programs. In order to correctly detect anchor segments without background music, we propose a method based on technologies such as BIC-Segmentation, BIC-Clustering and GMM-based speaker identification for TV news anchor detection. The experiment corpus is collected from daily news on ETT news program and the experiment result is good. Moreover, we integrate the proposed method and implement syllable-level indexing feature news spoken document retrieval system on TV news successfully.

誌謝

謝謝傅老師在我研究所兩年的生涯給予我的指導和照顧，並幫助我的論文找到研究方向，並學習到做研究方法與態度，才得以完成此篇論文。同時，感謝實驗室博士後研究以及博士班學長，永煜、柏伸、政龍、岳宏、士賢，還有學弟玉善，平常在生活上及學業上的建議與指教，還有感謝兩位同學建榮、政邦，兩年來同甘共苦，一起修課、玩樂、做研究，互相加油打氣。特別感謝士賢學長在論文上的極大幫助，讓我認識語音方面的知識也幫助我解決困難並修改論文，讓論文更爲完美。感謝大學同學以及朋友在生活上的鼓勵。最後，感謝爸爸、媽媽、妹妹一直在背後支持我，給我無憂無慮的生活，讓我可以專注在學業上，才得以順利完成學業。

目錄

表目錄

圖目錄

第 1 章

前言

1.1 研究動機

隨著電腦科學日漸發達與網際網路的興盛，電腦早已融入每個人的生活當中，隨著電腦網路越來越普及，網路上各式各樣的資源，如：文字文件、影像、視訊、音訊等資源，其中文字文件的搜尋引擎到目前為止已有不錯的成果(如:http://www.google.com)，幫助使用者在約數億的網頁中找到想要的資源，因此資訊檢索無疑是漫遊多媒體資訊時代的引擎，這種自動索引及檢索各種資訊的技術能協助人們在大量的多媒體資料庫中尋找想要的資訊。

由於語音辨識的研究已有數十年的歷史，用來做語音辨識最著名的聲學模型即是隱藏式馬可夫模型(Hidden Markov Model, HMM)，語音辨識率也已經有不錯的成果，因此近年陸續有不少研究是以此連續語音辨識技術為基礎對大量的音訊資源作索引與語音文件資訊檢索的相關技術，以期能以此技術為基礎發展出一套便捷的語音資訊檢索系統。

1.2 研究目標

本論文將針對台灣有線電視台的晚間新聞時段進行語音資訊檢索的相關問題進行研究，由於我們觀察到絕大部分的晚間新聞在主播播報新聞的同時，都會加上明顯的背景音樂(只有少數主播音段沒有背景音樂)，有背景音樂的新聞主播片段將會影響到對於主播音段進行音節辨識的音節正確率，進而影響後端欲實作的以音節為基礎之電視新聞語音檢索系統的正確率，為了克服上述問題，研究希望設計了一套自動新聞分析系統，能偵測出沒有背景音樂的新聞主播音段，並確認其在新聞節目中的位置，更進一步只針對沒有背景音樂的主播音段進行音節辨認，以作為後端實作新聞語音檢索的對象。

1.3 研究方向

首先，我們觀察到電視新聞語料不外乎就是主播音段(audio segment)、主播含背景音樂音段、外景男性音段、外景女性音段、廣告男性音段、廣告女性音段，這六大類語料所組成，論文中將利用以高斯混合模型(Gaussian Mixture Models)為基礎的語者識別(Speaker Identification)技術 【1】，配合適當的將新聞語料分類與收集，來達成將沒有背景音樂的主播語料與含有背景音樂的主播語料分離。

由於幾乎所有有關語者識別相關研究皆是採用高斯混合模型來代表語者的語音特性分布，且已經有不錯的成果，但是之前的那些作法對於以 EM 演算法來訓練語者的高斯混合模型時，該使用多少高斯元件(Gaussian Components)來代表訓練語料的特徵分布，到目前為止並未有理論上的方法去事先估計，都是以實驗的方式來決定高斯混合模型中高斯元件的個數，因此本論文將利用貝氏資訊法則(Bayesian Information Criterion, BIC) 【2】，以自我成長學習的方式來訓練語者的高斯混合模型 【3】，以自動決定並找出最適合代表該訓練語料的高斯混合元件個數。依上述方法訓練新聞中六大類語者的高斯混合模型，再以具有最

大機率的準則將每個新聞音段作分類，以期更準確的將沒有背景音樂的主播音段與含有背景音樂的主播音段分離。

1.4 章節介紹

在以下章節中，第二章首先介紹模型選擇與貝氏資訊法則(BIC)以及 BIC 如何應用於語者分段與語者音段分群的技術，緊接者介紹以高斯混合模型為基礎的語者識別技術;第三章則是本論文提出如何在主播音段有背景音樂的新聞環境下，找出純主播音段(沒有背景音樂的主播音段)並切割新聞故事;第四章是以第三章所提出的新聞沒有背景音樂的主偵測的方法的實驗結果，以驗證其可行性與效能;第五章是整合一些語音方面的相關技術，包括:以第三章所提之電視新聞主播偵測及新聞故事切割、沒有背景音樂的主播音段的音節辨識並實作電視新聞語音檢所系統;第六章則是結論及對未來的展望。

第 2 章

BIC 在語者分段及語者分群與語者識別之相關研究

2.1 模型選擇與貝氏資訊法則(BIC)

貝氏資訊法則(Bayesian Information Criterion,BIC)是一種模型選擇的法則，最先由 G. Schwarz 【2】所提出，所謂模型選擇指的是給定一資料集 $X=\{x_1,x_2,...,x_N\}$ 和一個機率模型候選集 $M=\{M_1,M_2,...,M_k\}$，我們要從模型候選集中選擇最適合的機率模型來表示 X 的資料分佈，則 BIC 定義為:

$$BIC\ (M_i) = \log\ P(X \mid M_i) - \frac{1}{2} d_i \log\ N \qquad (2.1)$$

其中 $P(X \mid M_i)$ 為資料 X 對模型 M_i 的最大相似度(maximum likelihood)，N 為資料總數，d_i 是模型 M_i 的參數個數，參數個數越多，表示模型複雜度越高，因此方程式(2.1)的第二項為對模型複雜度的懲罰(penalty)，期望能得到一個較簡單且又能精確的機率模型來估計資料的分佈情況，因此對於較複雜的模型施以較高的懲罰，以避免選擇到一個過於複雜的模型。

因此，當計算出來的 BIC 值(式(2.1))越大，就代表此組資料 X 的分佈越適合以此模型來表示，所以 BIC 法則告訴我們要選擇具有 BIC 值最大的模型。

2.2 BIC 語者分段與分群

近年來對於語者分段與分群 【13】的相關技術大概可歸為三大類:

(1) 以模型為基礎(GMM-based)【5】: 這種方式需要依靠事先訓練好的模型來區分 出不同的語者，但實際上並不可能取得各種的語者模型。

(2) 以距離為基礎(Metric-based)【6】: 利用滑動視窗(sliding window)，每次計算相鄰視窗間的距離，距離的量測方式可採用 KL2，但需要由訓練資料與經驗來決定門檻值(threshold)，最後選擇在門檻值之上的最高點，因此對於實際測試資料上，效果並不穩定。

(3) 以模型選擇為基礎(Model-Selection-based): 由 Chen 於 1998 年提出【4】，有別於傳統的語者分段方式，他利用貝氏資訊法則(BIC)的模型選擇理論來檢驗音段中是否有語者交換點(change point)，其優點是不需要事先知道各種語者的模型也無需訓練語者的模型。

若將模型選擇的問題應用到語者分段與分群上，則從多個模型的選擇簡化成兩個模型的選擇，分別介紹如下：

2.2.1 BIC 語者分段

給定 $X = \{x_1, x_2, ..., x_N\}$ 代表維度為 d 的特徵向量(MFCC)，假設此音段最多只有一個語者交換點，則可對每一個 x_b 做偵測，偵測其是否為語者交換點，Chen 用模型選擇的方法來檢定，下面兩個模型候選以偵測 x_b 是否為交換點:

M_0: 假設此音段沒有語者交換點， $x_1, x_2, ..., x_N \sim N(\mu, \Sigma)$ ，為一高斯分配

M_1: 假設 x_b 為語者交換點， $x_1, x_2, ..., x_b \sim N(\mu_1, \Sigma_1)$ 為一高斯分配;

$x_{b+1}, x_2, ..., x_N \sim N(\mu_2, \Sigma_2)$ 為另一高斯分配

然後以 BIC 來做選擇，定義 ΔBIC 為:

$$\Delta BIC_b = BIC(M_1) - BIC(M_0)$$
$$= \frac{1}{2}(N\log|\sum| - b\log|\sum_1| - (N-b)\log|\sum_2|) - \frac{1}{2}\lambda(d + \frac{1}{2}d(d+1))\log N \quad (2.2)$$

根據BIC法則告訴我們, 若$\Delta BIC_b>0$, 則我們應該選擇M_1的假設，且若b為語者交換點，必有$\Delta BIC_b>0$，所以最後選擇具有最大ΔBIC值且大於零的點為語者交換點，否則此音段沒有語者交換點。

2.2.2 BIC 音段分群

通常於一段音訊中相同語者經常會出現數次(如:在新聞節目中,新聞主播的音段會出現多次且分散在不連續的時間中)，當適當的將音訊分段之後，我們希望可以自動的將相同語者的音段集合起來。

假設S = {S_1,S_2,…,S_k}是音段的集合，Chen用階層式的分群法來對音段集合S作分群，一開始每一音段自成為一群(或一個節點)，每次考慮最近的(常以ΔBIC或KL2 距離來衡量兩音段的距離)兩音段S_i和S_j，接著考慮兩種候選模型:

$M_0 : x_1^i, x_2^i, ..., x_{n_i}^i, x_1^j, x_2^j, ..., x_{n_j}^j \sim N(\mu, \sum)$，此兩段為一高斯分配

$M_1 : x_1^i, x_2^i, ..., x_{n_i}^i \sim N(\mu_i, \sum_i);$
$x_1^j, x_2^j, ..., x_{n_j}^j \sim N(\mu_j, \sum_j)$，每一音段分別為一高斯分配

根據 BIC，計算其ΔBIC 如下式:

$$\Delta BIC = BIC(M_1) - BIC(M_0)$$
$$= \frac{1}{2}((n_i + n_j)\log|\sum| - n_i\log|\sum_i| - n_j\log|\sum_j|) - \frac{1}{2}\lambda(d + \frac{1}{2}d(d+1))\log(n_i + n_j) \quad (2.3)$$

根據BIC法則，若ΔBIC<0，則合併此兩個節點(即S_i, S_j為同一語者的音段)，並更新合併後節點的距離，若ΔBIC不小於零，則此兩音段屬於不同語者，則分群完成。

2.3 以高斯混合模型為基礎的語者識別

所謂語者識別(Speaker Identification)是從一群已知的語者中識別出與一個未知輸入的聲音最相似的語者，如圖 2-1 表示模型訓練的步驟，首先將 N 位語者的訓練語料經由特徵參數的擷取(Feature extraction)後得到特徵參數向量，再經由語者模型訓練的方法來訓練每位語者的模型，可代表每位語者的語音特性，語者模型訓練的方法主要有向量量化(Vector Quantization-VQ)、徑向基底函數(Radial basis function-RBF)和高斯混合模型(Gaussian Mixture Models-GMM)等方法。

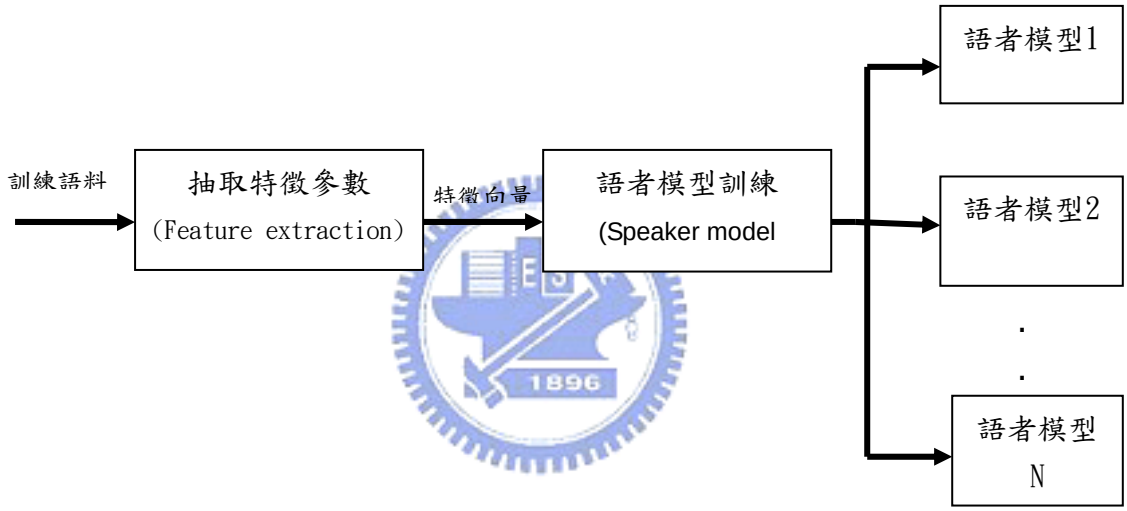

圖 2-1 語者模型訓練流程圖

將 N 位語者語音資料訓練成 N 個語者模型之後，接著進行語者識別的步驟，如圖 2-2 所示，同樣將一測試語料經過特徵參數擷取之後得到特徵向量，接著將語音的特徵向量丟進分類器(classifier)中進行分類，然後從已訓練好的語者模型中找出與輸入最相近的語者身份，即為識別出的語者。

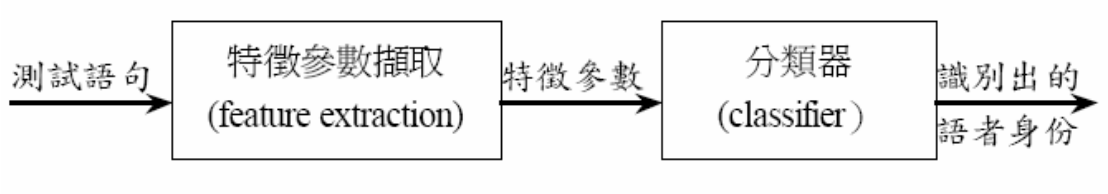

圖 2-2 語者識別的步驟

此外，根據訓練與測試語句可將語者識別分為文本相關的語者識別(Text-dependent Speaker Identification)和文本獨立的語者識別(Text-independent Speaker Identification)，前者的訓練和測試語料必須是特定的文字內容，而後者可以接受不受限制的文字發音，在測試時也不要求測試者所唸的文字內容。本論文主要探討文本獨立的語者識別，而前人的研究中以高斯混合模型為基礎的文本獨立的語者識別已有不錯的成果【1】，以下將簡單介紹以高斯混合模型為基礎的語者識別技術：

高斯混合模型(GMM)

高斯混合模型的機率密度函數(probability density function)如下:

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i b_i(x) \qquad (2.4)$$

其中第 i 個高斯元件(Gaussian component)的高斯函數為

$$b_i(x) = \frac{1}{(2\pi)^{D/2} \mid \sum_i \mid^{1/2}} \exp\{-0.5(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)\} \qquad (2.5)$$

- x 為 D 維資料點
- M 為高斯元件個數
- μ_i 和 $\sum_i$ 分別為第 i 個高斯元件的平均向量(mean vector)與共變異數矩陣(covariance matrix)

- w_i 為高斯元件(component)的權重(weight)，且 $\sum_{i=1}^{M} w_i = 1$
- $\lambda = \{w_i, \mu_i, \sum_i\}\ i = 1,...,M$ 為高斯混合模型的參數集

對於語者識別而言，每一位語者以一高斯混合模型來表示其語音特徵的分布情形。

訓練語者的高斯混合模型

所謂訓練語者的高斯混合模型即是去估計 GMM 的參數集 λ 以符合訓練語料，最常用的方法是最大概似估計(maximum-likelihood estimation, ML)，給定訓練語料並抽取特徵參數向量集 $X = \{x_1, x_2, ..., x_N\}$，ML 就是要去估計 GMM 的參數，使得 X 的 likelihood $p(X \mid \lambda) = \prod_{i=1}^{N} p(x_i \mid \lambda)$ 有最大值。然而目前尚無法用分析的方式直接求得 $p(X \mid \lambda)$ 的最大值。一個最常用的方法為 EM(expectation-maximization)演算法【14】。給定模型的參數初始值，以迭代的方式去調整 GMM 的參數，保證每次調整過後的新參數 $\bar{\lambda}$ 滿足 $p(X \mid \bar{\lambda}) \geqq p(X \mid \lambda)$，接著以新參數當作新模型的初始值進行下一次的迭代，直到收斂至事先指定的門檻值(threshold)為止，EM 演算法對 GMM 的參數調整方式如下。

對於第 i 個高斯元件而言:

$$p(i \mid x_t, \lambda) = \frac{w_i b_i(x_t)}{\sum_{k=1}^{M} w_k b_k(x_t)} \quad (2.6)$$

Mixture Weights:

$$\bar{p}_i = \frac{1}{N} \sum_{t=1}^{N} p(i \mid x_t, \lambda) \quad (2.7)$$

Mean vector:

$$\bar{\mu}_i = \frac{\sum_{t=1}^{N} p(i \mid x_t, \lambda) x_t}{\sum_{t=1}^{N} p(i \mid x_t, \lambda)} \quad (2.8)$$

Covariance matrix (diagonal):

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^{N} p(i \mid x_t, \lambda) x_t^2}{\sum_{t=1}^{N} p(i \mid x_t, \lambda)} - \bar{\mu}_i^2 \quad (2.9)$$

語者識別

若我們有 S 個語者，其對應的高斯混合模型為 $\lambda_1, \lambda_2, ..., \lambda_S$，對於輸入測試的語料特徵向量集 X，要從 S 中找出一位與 X 最相似的語者(即那一位語者的 GMM 最適合來描述 X)，一般以具有最大事後(posteriori)機率的準則來決定:

$$\begin{aligned} \hat{S} &= \arg\max_{1 \le k \le S} \Pr(\lambda_k \mid X) \\ &= \arg\max_{1 \le k \le S} \frac{p(\lambda_k) p(X \mid \lambda_k)}{p(X)} \end{aligned} \quad (2.10)$$

假設每一個語者的事前($p(\lambda_k)$)機率都相同，且對於每一位語者 $p(X)$ 亦相同，因此式子(2.10)可簡化為 $\hat{S} = \arg\max_{1 \le k \le S} p(X \mid \lambda_k)$，於實際運算時通常會取對數，故:

$$\begin{aligned} \hat{S} &= \arg\max_{1 \le k \le S} \log p(X \mid \lambda_k) \\ &= \arg\max_{1 \le k \le S} \sum_{t=1}^{N} \log p(x_t \mid \lambda_k) \end{aligned} \quad (2.11)$$

第 3 章

在主播有背景音樂的新聞環境下作新聞主播的偵測與新聞故事的切割

本章敘述在有些主播有背景音樂的電視新聞節目的環境下，說明如何將每天錄製的新聞做分析，以達成自動切割新聞故事(圖 3.1 為電視新聞結構示意圖)，於本章 3.1 節首先介紹前人【3】非監督式的自動切割新聞的作法，但是若新聞節目當中，有某些新聞主播片段有無法預測的背景音樂，則之前的方法將會受到有背景音樂的主播片段影響以致無法正確切割出新聞故事，因此本論文 3.2 節將針對此點提出一個監督式的解決方式，以語者識別為基礎將新聞語料做分類，以偵測出沒有背景音樂的主播新聞故事，因為我們將於後端實作主播音節辨識與新聞主播語音檢索系統(將在第五章介紹)，而若主播音段有背景音樂會大大降低主播音節辨識的正確率與語音檢索的效能，故我們捨棄有背景音樂的主播新聞故事。

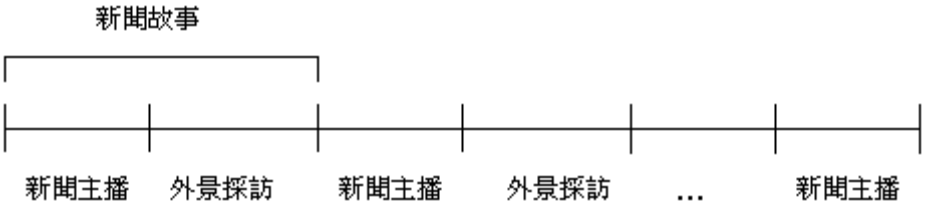

圖 3.1 電視新聞節目的結構

3.1 非監督式的電視新聞故事切割

本節說明論文【3】所提之電視新聞主播偵測方法，一般來說，電視新聞的結構如同圖 3.1，而且新聞主播的音段數目通常是新聞節目中最多的，所以根據以上的特性可以偵測出新聞節目中的主播，進而做新聞故事的切割，作法主要分為二個步驟:

步驟一: 以 BIC 作新聞語者分段。

步驟二: 由步驟一所分割出來的新聞音段作 BIC 音段分群，並根據新聞節目中主播音段為最多的原則，認為段數最多最大群的為新聞主播的音段群。

最後將主播音段群內的音段以時間前後做排序，因此每段主播的開始時間到下一段主播的開始時間之間即為一段新聞故事，如此即完成新聞故事的切割。

其中步驟一、二的演算法如下:

以 BIC 作新聞語者分段(步驟一)

新聞節目屬於多個語者交換點的偵測，圖 3.2 說明了它的作法，以 10 秒為一音窗(audio window)，用 BIC 偵測音窗內是否有語者交換點，若此音窗無語者交換點則以重疊 5 秒的方式偵測下一個音窗，一旦偵測到交換點，則從此交換點重新開始再取下一個音窗繼續偵測，如此重複直到找完整段新聞為止。

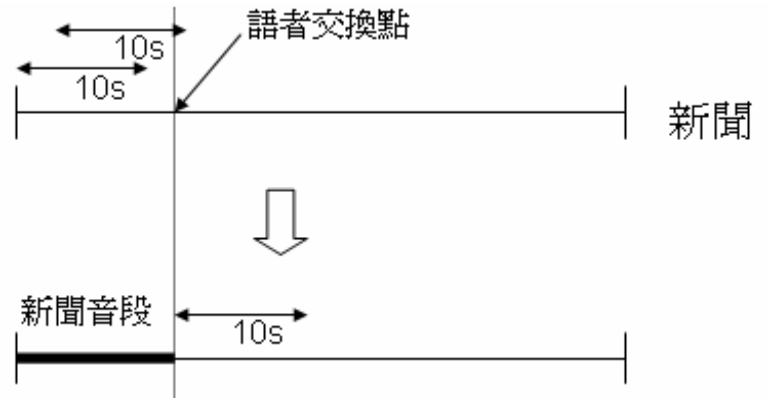

圖 3.2:電視新聞的多個語者交換點偵測

以 BIC 作新聞音段的分群(步驟二)

當偵測出新聞音段中所有的語者交換點之後,即可將此音段分割成多個沒有語者交換點的音段,這些音段當中新聞主播音段會出現多次且散落在新聞節目當中，所以在做完語者分段之後可做分群，將相同語者的音段集合起來，假設 $S=\{S_1,S_2,...,S_n\}$ 是新聞音段的集合，以階層式的分群法來對 S 作分群，演算法如下:

1. 初始每一群節點 C_i 只包含一個新聞音段 S_i
2. 計算任兩節點的距離(以 ΔBIC 為距離衡量，方程式 2.3)
3. 選擇最近的兩節點(ΔBIC 最小)

   若 ΔBIC < 0

   則合併此兩節點並回到演算法步驟 2

   否則(即 ΔBIC>=0) 分群完成
4. 最後選擇最大群(段數最多)為新聞主播音段群

近來觀察到一般電視新聞節目主播在播報新聞的同時,有些新聞主播音段並

非純粹只有主播的聲音，通常會伴隨有明顯的背景音樂，造成沒有背景音樂的主播音段與含有背景音樂的主播音段的語音特性明顯不同，破壞了上述方法中最大群為新聞主播群的假設，故若以上述非監督式的電視新聞主播偵測方法偵測將無法正確切割新聞故事；再者，若要繼續針對新聞主播音段進行語音的音節辨認，具有背景音樂的主播音段勢必會大大降低音節辨認的正確率，因此本論文下一節將以監督式的方式針對固定時段的新聞以語者識別為基礎來做電視新聞主播音段(不具背景音樂的主播音段)的偵測，以利進一步實作電視新聞語音檢索系統(在第五章介紹)。

3.2 以語者識別為基礎的電視新聞主播偵測

在第二章中介紹了如何用高斯混合模型來做語者識別，在本節中，我們要介紹在新聞主播有背景音樂的環境下以 GMM 為基礎的語者識別方法來電視新聞語料的分類，以擷取新聞中沒有背景音樂的主播音段並切割新聞故事。

3.2.1 新聞語者的分類與訓練新聞語者的高斯混合模型

在電視新聞節目中，除了新聞主播之外，還有外景記者和外景中不特定人士的語料和廣告語料等，對於語者的高斯混合模型而言，我們將新聞語料細分為六大類—沒有背景音樂的主播語料、含有背景音樂的主播語料、外景男記者包含外景不特定的男性語料、外景女記者包含外景不特定的女性語料、廣告中男性語料、廣告中女性語料，接著以 GMM 為基礎的語者識別方法來做新聞語料的分類，為這六大類新聞語者的語料訓練其 GMM，首先需要收集六大類語者的新聞訓練語料，收集方式如下:

1. 沒有背景音樂的主播語料: 收集事先選定的時段的新聞主播的語料約四分鐘，不具任何背景音樂。
2. 含有背景音樂的主播語料: 收集該時段的新聞主播含有背景音樂的語料約四分鐘，此四分鐘的語料須盡量包含新聞中出現過在主播音段的背景音樂。
3. 外景男性聲音語料: 收集外景新聞的語料約四分鐘，其中包括外景男記者的聲音，與外景之中非記者的男性聲音，且大部分語料需伴隨著明顯的外在環境的吵雜聲。
4. 外景女性聲音語料: 收集外景新聞的語料約四分鐘，其中包括外景女記者的聲音，與外景之中非記者的女性聲音，且大不分語料需伴隨著明顯的外在環境的吵雜聲。
5. 廣告中男性語料: 收集了約四分鐘廣告中男性的語料，其中絕大部分語料有各種不同的背景音樂或聲音特效。
6. 廣告中女性語料: 收集了約四分鐘廣告中女性的語料，其中絕大部分語料有各種不同的背景音樂或聲音特效。

當收集完新聞語者的訓練語料之後，接著進行訓練六類新聞語者的 GMM，然而若純粹以傳統的 EM(Expectation-Maximization)演算法的方式來學習新聞語者的高斯混合模型會有下列幾項缺點:

1. EM 演算法必須事先設定好六類新聞語者 GMM 的高斯元件個數(components)，也必須先給定模型參數集的初始值，因此容易發生預測過多或過少高斯元件的情形，造成過度表示或不足以表示語者的語音特徵分佈。
2. 模型參數集的初始值對 EM 演算法的結果有決定性的影響，初始值給的不好，可能使得 EM 演算法最終只求得局部最大值(local maximum)。
3. 高斯元件的共變異數矩陣(covariance matrix)可能是奇異的(singular)，因而無法求得共變異數矩陣的反矩陣。

因此本論文避免共變異數矩陣的奇異發生，採用對角共變異數矩陣，且對於如何決定新聞語者的高斯混合模型的高斯元件，採用論文【3】所提的 “以 BIC 為基礎的自我成長學習法” 來學習新聞語者的高斯混合模型，可自動決定六大類新聞訓練語料的高斯元件個數，演算法介紹如下:

為方便描述演算法，先定義一些名詞與符號:

- $X = \{x_1, x_2, ..., x_N\}$:欲訓練新聞語者的資料集。
- GMM_k :有 K 個高斯元件的高斯混合模型，

  其模型參數集以 $\theta_i = \{w_i, \mu_i, \sum_i\}\, i = 1, ..., K$ 表示。
- $BIC(GMM_i, D)$:表示資料集 D 在候選模型 GMM_i 之下的 BIC 值，BIC 值若越大表示此組資料越適合以此模型來表示。
- 當候選模型只有兩個時，定義了 ΔBIC 為 $\Delta BIC_{21}(D) = BIC(GMM_2, D) - BIC(GMM_1, D)$，若 ΔBIC 大過某個值(growing-confidence, 信心度)，根據 BIC 法則，選擇 GMM_2 作為 D 的機率模型是比選擇 GMM_1 來的適合的。

以 “BIC 為基礎自我成長學習法” 訓練新聞語者的高斯混合模型步驟如下:

1. 演算法從一個高斯元件開始成長，初始化，以 GMM_1 來代表訓練資料 X 的機率模型，令 K 為目前高斯元件的個數，其高斯混合模型的參數集為 $\theta = \{w_i, \mu_i, \sum_i\}\, i = 1, ..., K$，此時 K=1。

2. 分群(clustering): 對每筆訓練語料 x_i，$p(\theta_k \mid x_i)$ 表示 x_i 由第 k 個高斯元件產生的機率，計算其和每個高斯元件產生的機率，找出具有機率最大的高斯混合元件 t，並將 x_i 標記為第 t 群(每一個高斯元件表示一群)，故可將所有

訓練資料分成 K 群(K 表示成長到目前為止高斯元件的個數)，令每一群的資料集為 $D_i\ i=1,2,...,K$ 。

3. 長出一個高斯元件(grows one component): 由步驟 2 將資料 X 分成 K 群，而對於每一群，以 BIC 來做選擇，看其是用一個高斯元件 GMM_1 來表示比較好?還是用兩個高斯元件 GMM_2 來表示比較好?接下來先以 EM 演算法來計算資料集 D_i 分別在候選模型 GMM_1 和 GMM_2 下的最大相似度(likelihood)以求得每一群的 ΔBIC 值為 $\Delta BIC_{21}(D_i)\ i=1,...,K$ ，從 BIC 的觀點來看，選擇 $\Delta BIC_{21}(D_i)$ 最大且大於 growing-confidence 的群來長出一個高斯元件，即以 GMM_2 取代原本對應的高斯元件，因此:

令 $\max \Delta BIC = \max_i \{\Delta BIC_{21}(D_i)\}$ (i.e.,找出最大的 ΔBIC 值)

$j = \arg\max_i \{\Delta BIC_{21}(D_i)\}$ (i.e.,找出哪一群擁有最大 ΔBIC 值)

且 $\bar{\theta}$ 為以 EM 演算法計算資料 D_j 在 GMM_2 下的所得到的模型估計參數

若 max ΔBIC > growing-confidence

則表示第 j 個高斯元件以兩個高斯元件來表示比較好，

$\theta = \theta \setminus \{w_j, \mu_j, \sum_j\}, \theta = \theta \cup \bar{\theta},\ K = K+1$ (長出一個高斯元件)

否則若 max ΔBIC < growing-confidence，表示每一群都認為其用一個高斯元件來表示比較好，因此訓練語者的 GMM 完成，演算法結束。

4. Global EM: 對新聞語者的訓練資料 X 實行 EM 演算法，高斯元件個數為步驟 3 成長過後的 K ，模型參數集採用步驟 3 長出一個高斯元件後的 θ 為 EM 演算法的初始值，以 EM 訓練完成後回到步驟 2。

3.2.2 應用語者識別於新聞語料的分類與新聞主播的偵測

依上一節所述我們可將電視新聞語料分成六大類並且分別訓練六類新聞語者的高斯混合模型(GMM)，六類語者分別為沒有背景音樂的主播、主播含有背景音樂、外景男記者、外景女記者、廣告男音、廣告女音且其對應的 GMM 參數集分別為 $\theta_1,\theta_2,\theta_3,\theta_4,\theta_5,\theta_6$，則我們可將 3.1 節介紹的非監督式的電視新聞故事切割方法修改為三個步驟:

1. 以 BIC 做新聞語者的分段，可得到多個沒有語者交換點的新聞音段，假設 $S=\{S_1,S_2,...,S_N\}$ 為分段後的新聞音段的集合。
2. 將新聞音段集合 S 中的每一新聞音段分別丟入 GMM 語者識別器做新聞語者的分類，因此可將新聞語料分成六類，以圖 3.3 來表示 GMM 新聞語者識別器。
3. 將被分類到新聞中沒有背景音樂的主播群內的所有新聞音段作 BIC 語者音段分群，因為以 GMM 為基礎的語者識別的正確率不可能達到百分之百完全正確，但只會有極少部分非真正沒有背景音樂的主播語料被分類到新聞中沒有背景音樂的主播群中，因此我們利用 BIC 語者音段分群，藉此將極少部分的非新聞純主播語料再分離開來，此時再分完群之後，我們就可認定最大群即為真正的新聞中沒有背景音樂的主播音段群，完成新聞主播(不具背景音樂)偵測。

最後將主播音段群內的音段以時間前後做排序，因此每段主播的開始時間到下一段主播的開始時間之間當作一段新聞故事，如此即完成新聞故事的切割。

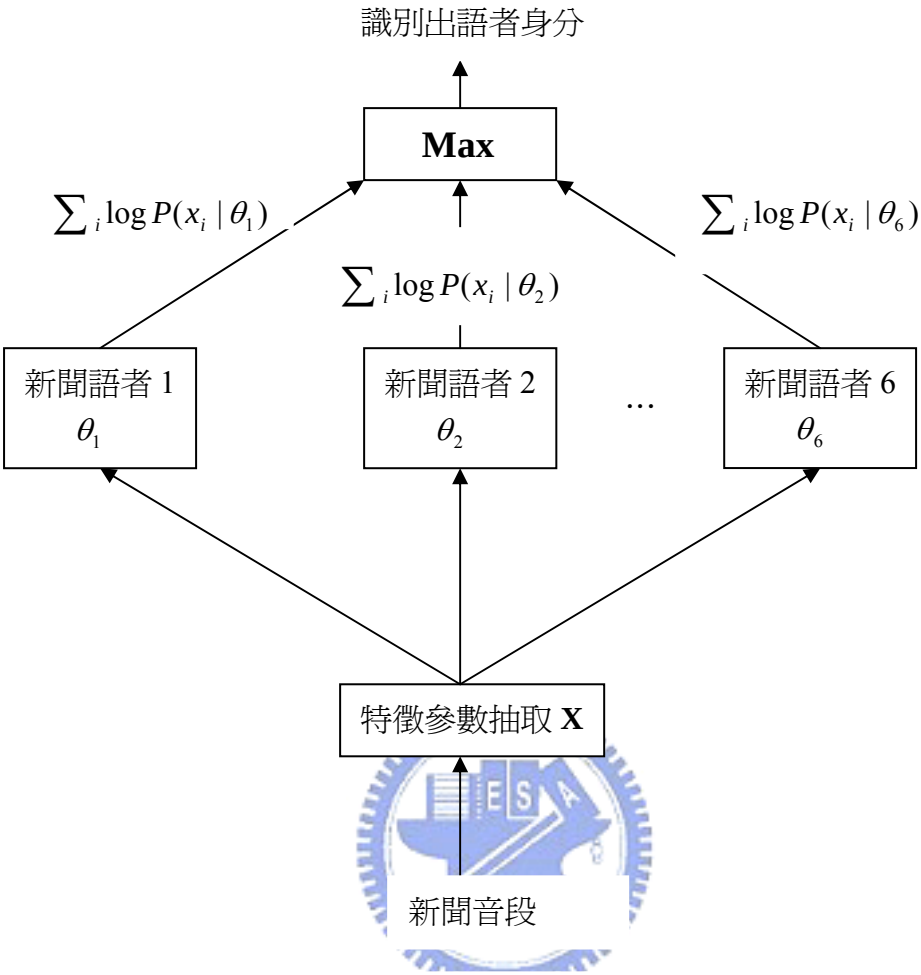

圖 3.3 GMM 新聞語者識別器

第 4 章

實驗結果

在這章中對於第三章所提之在有些主播音段有背景音樂的新聞環境下，作電視新聞純主播(沒有背景音樂的主播片段)偵測的方法，加以實作，並設計實驗以評估此方法的效能。

4.1 實驗環境及資料來源

對於實驗的平台，在硬體方面使用了以 Intel Pentium-4 2.4Ghz 的時脈速率中央處理器的個人電腦，搭配有 1Gigabytes 主記憶體，並接有電視影像擷取卡，作業系統為 Microsoft Windows XP 專業版。

實驗對象為有線電視東森新聞台，且選擇大部分主播音段部份都有明顯的背景音樂，只有少部分主播音段為沒有背景音樂的主播語音的新聞時段，並以電視影像擷取卡將電視新聞直接錄成影像檔(.asf)和音訊檔(.wav) ，音訊取樣頻率(sample rate)為 44.1k，每個 sample 16bits 大小，且實驗中的語音參數抽取為 mfcc，維度皆為 24 維，用於電視新聞主播偵測。

4.2 實驗方式

本論文實驗流程分兩部份:訓練電視新聞語者的高斯混合模型及計算以第三章所提的以高斯混合模型為基礎的語者識別技術用於電視新聞主播偵測的方式

偵測沒有背景音樂的主播音段的正確率，分別敘述如下:

1. 訓練六大類新聞語者的高斯混合模型:

   如第三章所提，將新聞語料分成六大類，分別收集這六類語料以作為訓練新聞語者的高斯混合模型，分別為:沒有背景音樂的主播音段、主播含有背景音樂、外景男音、外景女音、廣告男音、廣告女音，實驗中將收集四個時段的新聞來評估此方法的可行性，即有四個沒有背景音樂的主播語者的 GMM，和其對應的含有背景音樂的四個 GMM，再加上外景兩個 GMM 和廣告的兩個 GMM，分別以 “BIC 為基礎的自我學習成長的方法”來學習語者的 GMM，訓練語料的長度約為三到四分鐘，流程如圖 4.1，分別得到四個沒有背景音樂的主播(四位主播分別為盧秀芳、王佳婉、趙心屏、馬千惠)的 GMM 其高斯元件個數分別為 68、81、78、84，其對應的含有背景音樂的的四個 GMM 其高斯元件個數分別為 72、86、83、61，而外景男音與外景女音 GMM 的高斯元件個數分別為 133、121，廣告男音與廣告女音 GMM 的高斯元件個數分別為 112、139。

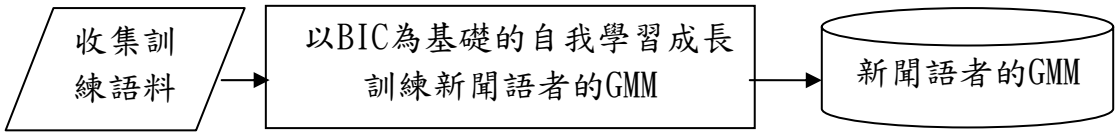

圖 4.1: 訓練新聞語者 GMM 流程

2. 計算以語者識別為基礎的新聞沒有背景音樂的主播音段偵測的正確率:

   在完成訓練新聞語者的高斯混合模型之後，我們將分別測試這四個時段的主播新聞語料，每次測試為一個主播的新聞時段，長度為一個小時，實驗流程如圖 4.2:首先錄製一個小時的新聞，接著以 BIC 語者分段將新聞分成多段沒有語者交換點的新聞語者音段，再將每一段的新聞音段分別丟入 GMM 新聞語者識別器作新聞分類，可將所有新聞音段分成六類，同時需要人工標記所有

新聞音段有那些屬於純主播音段(沒有背景音樂的音段)，接著把 GMM 語者識別的結果被分類到純主播類的作比較，分別計算 precision 及 recall，來評估其優劣，其中 precision 代表程式所找到的純主播群中的音段個數(分母)，當中有幾個和人工標出的純主播音段相符(分子)；recall 代表人工標出的所有純主播音段中(分母)，被程式找到純主播音段的個數(分子)。

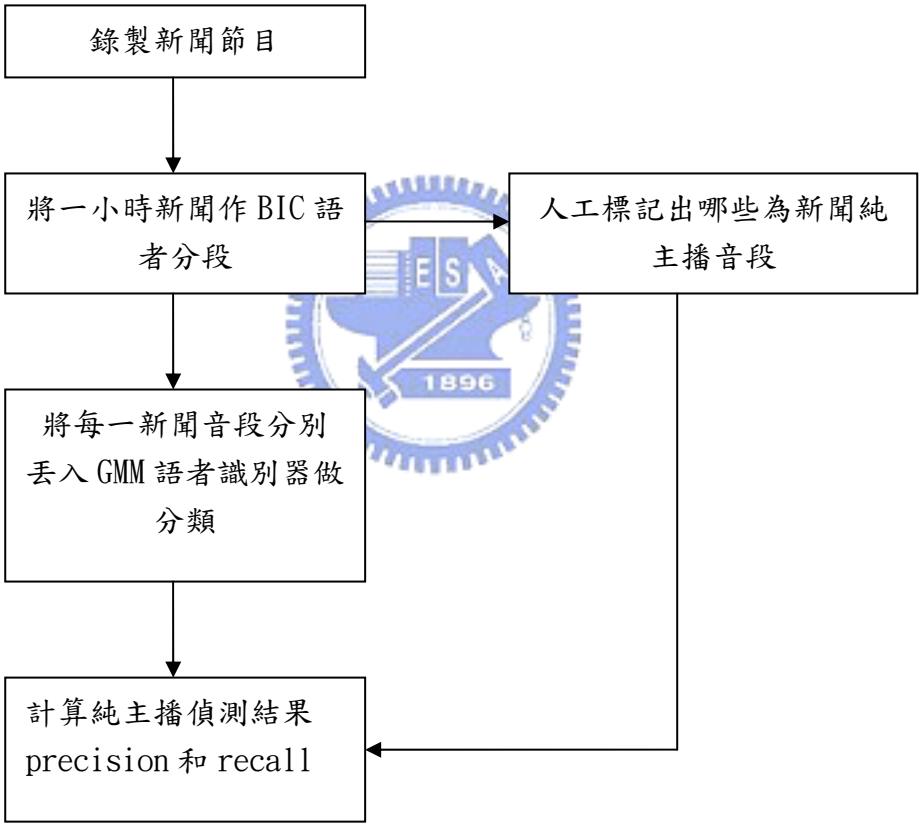

圖 4.2: 以 GMM 語者識別來偵測純主播實驗流程

4.3 實驗數據與結果

在本節中將列出以 GMM 語者識別為基礎的新聞沒有背景音樂的主播偵測的實驗結果，以驗

證我們所提出的在有些主播音段有背景音樂的電視新聞的環境下，仍然能偵測出純主播的音段(沒有背景音樂的音段），以確認其在新聞節目中的位置。

實驗一:

如上節所述，我們選擇四個不同時段的四個主播為實驗對象，每次實驗為完整一各小時的新聞節目，每一主播時段分別用五天(即每天一個小時，共五天)的新聞作為測試語料，表 4-1 分別列出 precision 和 recall 的實驗結果。

表 4-1 以 GMM 語者識別來偵測新聞純主播的四個時段五天的實驗結果

GMM68-盧秀芳	Precision	Recall
1	17/17(100%)	17/18(94%)
2	8/8(100%)	8/9(89%)
3	16/16(100%)	16/18(89%)
4	4/4(100%)	4/7(58%)
5	7/8(88%)	7/7(100%)
平均	97.6	86
GMM81-王佳婉	Precision	Recall
1	7/7(100%)	7/8(88%)
2	13/13(100%)	13/15(87%)
3	3/4(75%)	3/3(100%)
4	4/4(100%)	4/5(80%)
5	10/10(100%)	10/12(83%)
平均	95	87.6

GMM78-趙心屏	Precision	Recall
1	17/17(100%)	17/19(89%)
2	17/18(94%)	17/20(85%)
3	16/16(100%)	16/19(84%)
4	12/12(100%)	12/12(100%)
5	15/15(100%)	15/17(88%)
平均	98.8	89.2
GMM84-馬千惠	Precision	Recall
1	19/20(95%)	19/19(100%)
2	25/27(93%)	25/27(93%)
3	20/20(100%)	20/22(91%)
4	19/19(100%)	19/19(100%)
5	16/16(100%)	16/18(89%)
平均	97.6	94.6

實驗一討論:

由實驗數據我們可以發現，precision 大都非常的高，表示純主播 GMM 接受條件越嚴格，較不容易有錯誤出現(所謂錯誤是說非純主播音段被認為是純主播音段的情況)，但相對的 recall 就較低，即比較可能出現比較多的漏失(純主播音段被分類到其他語者)，整體來說，新聞純主播群仍然包含了絕大部分測試新聞節目中的新聞純主播語料，因此以語者識別為基礎的新聞純主播偵測確實可以切割出新聞節目中沒有背景音樂的新聞故事，更以利於後端實作語音音節辨識與新聞主播語音檢索系統。

實驗二:

實驗二選擇新聞時段為盧秀芳主播為實驗對象,將新聞六大類語料分別以固定 16、32、64、96 個高斯元件的方式以 EM 演算法訓練新聞六大類語者的高斯混合模型，並分別與原本以 “BIC 為基礎的自我學習成長的方法” 來訓練新聞六大類語者的方式做比較(由實驗一盧秀芳其高斯元件為 GMM68)，同樣選擇五天五個小時的新聞語料做測試，表 4-2 分別列出其純主播的 precision 與 recall 的結果。

表 4-2: 比較固定高斯元件個數的方式來訓練語者 GMM 與以 “BIC 為基礎自我學習成長的方法” 來訓練語者 GMM，以比較兩種方式的優劣

GMM16	Precision	Recall	GMM96	Precision	Recall
1	12/12	12/13	1	13/13	13/13
2	9/12	9/10	2	10/11	10/10
3	15/19	15/15	3	15/15	15/15
4	9/10	9/9	4	9/9	9/9
5	9/10	9/9	5	9/10	9/9
平均	86.8	96.4	平均	96.2	100%
GMM 32	Precision	Recall	GMM68	Precision	Recall
1	13/14	13/13	1	10/10	10/13
2	10/10	10/10	2	7/7	7/10
3	15/15	15/15	3	13/13	13/15
4	9/12	9/9	4	8/8	8/9
5	8/9	8/9	5	7/7	7/9
平均	91	97.8	平均	100%	80.2
GMM 64	Precision	Recall			
1	12/13	12/13			
2	9/10	9/10			
3	15/15	15/15			
4	7/7	7/9			
5	8/9	8/9			
平均	94.2	89.8			

實驗二討論:

由實驗二結果顯示以自動決定高斯元件個數的方式比其他以固定元件方式訓練的高斯元件方式，其結果雖然 precision 比其他高，但整體效果似乎沒有明顯差距，其可能原因:1. 測試資料不夠多，2. 實驗列表只列出純主播類的 precision 與 recall，或許其他類其結果會明顯比以固定元件訓練方式來的好。

第 5 章

系統應用:電視新聞語音檢索系統

在本章中將介紹應用本論文所提之電視新聞純主播偵測方法偵測出沒有背景音樂的新聞主播音段，進一步針對主播音段加以實作語音音節辨識，成為新聞主播語音檢索文件，最後實作出電視新聞語音檢索系統。因此本章第一節先介紹語音音節辨識與語者調適的實作方法，接著第二節介紹以音節(syllables)為索引特徵(indexing terms)的資訊檢索模型(information retrieval model)的實作方式，最後第三節為整個新聞語音檢索系統的架構以及末節的檢索效能評估。

5.1 語音辨識與語者調適之實作

目前在語音辨識的部份較為普遍而辨識效果較好之語音辨識核心大多使用隱藏式馬可夫模型(Hidden Markov Model, HMM) 【7】【8】，因此本論文在語音辨識與語者調適實作方面，使用了劍橋大學工程系(Cambridge University Engineering Department)所發展的 Hidden Markov Model Toolkit(HTK)第 3.2.1 版的發展工具來建立我們系統所需的語音音節辨識的功能 【9】。

在語音特徵參數方面使用了 12 維的梅爾倒頻譜參數(Mel-Frequency Cepstral Coefficients, MFCC)，加上對數能量參數，共 13 維參數，再計算此

13 維參數其一階差量和二階差量共計 39 維的語音特徵參數，而在訓練語料方面，使用了 TCC300 國語語音資料庫 【10】，TCC300 為國立台灣大學，國立交通大學，國立成功大學各自之語音資料庫所集合而成，屬於麥克風朗讀語音，共三百人的語音資料，當中兩百六十個語者當作訓練語料，四十個語者為測試語料，以 HTK TOOL 訓練出語者無關(speaker independent)的隱藏式馬可夫模型(聲母與韻母共一百五十一個中文語音基本單位的 HMM)，用來做音節辨識。

辨識過程分為兩階段(使用 HTK recognizer)，對於連續語音，先辨識出整段音節結果與各音節在此段語音中出現的時間，再以此時間資訊，對特徵參數依各音節作分段(utterance segment)，分成各單音節的特徵參數，進入第二階段的辨識，對各單音節作辨識，輸出機率值前五大的候選音節結果(成為 syllable lattice)，本論文將每一主播音段經過此兩階段的音節辨識所得到的 syllable lattice 當作欲被檢索的新聞主播語音文件，表 5-1 列出 TCC300 測試語料的音節正確率(Correct)與精確率(Accuracy)，公式說明如下:

$$Correct = \frac{H}{N} \times 100\%$$

$$Accuracy = \frac{H - I}{N} \times 100\%$$

其中 N 為測試語料文稿中所有音節的數量，H 為辨識結果中正確的音節數量，I 為插入型錯誤的數量(Insertion error)。

表 5-1: 以 TCC300 中 260 人所訓練的 HMM(稱為原始 HMM)，並以 TCC300 的測試語料測試其音節辨識率

原始 HMM	Correct(%)	Accuracy(%)
TCC300 測試語料 40 人	74.45	69.3

若拿電視新聞當中的主播語料(沒有背景音樂)當作測試語料，其結果如表 5-2:

表 5-2: 以 TCC300 中 260 人所訓練的 HMM，並以電視新聞主播的語音當測試語料測試其音節辨識率

原始 HMM	Correct(%)	Accuracy(%)
電視新聞純主播語料(10 分)	6.17	2.79

由表 5-2 觀察得知，因為訓練語料和測試語料間聲學特性的不匹配，使得正確率都會較低落，因此，根本無法用於辨識新聞主播的語音，更進一步影響語音檢索。

在新聞訓練語料難以大量取得的情況下，非得要以 TCC300 語音資料庫訓練而來的 HMM 作為我們辨識新聞主播的模型時，為了提升辨識率，讓我們得以用來辨識新聞主播語料，我們利用目前極為有效的兩種語者調適 (Speaker Adaptation)技術，分別為最大事後機率估測法 (Maximum a Posteriori, MAP) 和最大相似度線性迴歸法 (Maximum Likelihood Linear Regression, MLLR)，調適的目的即是希望藉由改變原本語者不特定(speaker indenpendent)的聲學模型參數，使模型所代表的語音特性能和測試語料的語者能匹配，成為該語者的語者特定(speaker dependent)模型，藉此提高辨識率。

因此我們收集了約四十分鐘的新聞純主播語料(東森晚間新聞主播盧秀芳)，作為 HMM 的調適語料，以上述調適技術調適原始由 TCC300 所訓練的語者不特定模型，使其成為語者特定模型(盧秀芳)，再另外收集盧秀芳的語料作測試，測試結果如表 5-3。

表 5-3: 進行語者調適後的模型之音節辨識率

語者調適後的 HMM	Correct(%)	Accuracy(%)
主播盧秀芳測試語料(10 分)	64.3	59.5

由上述實驗得知音節辨識率已有大幅度的提升，因此我們的系統將以此調適後的

模型來辨識新聞純主播音段，進一步供後端實作新聞語音文件檢索之用。另外我們拿有背景音樂的主播音段以調適過後的模型來辨識，其音節辨識率如表 5-4:

表 5-4: 以調適後的模型來辨認含有背景音樂的主播音段的正確率

語者調適後的 HMM	Correct(%)	Accuracy(%)
主播盧秀芳(含有背景音樂)測試語料(10 分)	47.1	38.1

由此實驗可知道即使以調適後的模型來辨識有背景音樂的主播片段，其辨識率依然不理想，因此才會只針對沒有背景音樂的主播音段作音節辨識。

5.2 電視新聞語音檢索之實作

我們的語音檢索系統採用 Chen et al.(2002) 【11】和 【15】所提的以音節為基礎的索引特徵(syllable-based indexing feature)與常被廣泛使用的向量空間檢索模型(Vector space retrieval model)，證明以音節為索引特徵用於中文語音檢索上，其效能比以字(character)與詞(word)都還要來的更好。

音節索引特徵(syllable-level indexing terms)

檢索系統是對每一個新聞主播語音文件(上節說明每一主播音段經過語音辨識後可得到 syllable lattice)各自抽出重疊音節 N 連索引(overlapping syllable N-grams,N=1~3)與間隔 N 音節的重疊音節對索引(overlapping syllables pairs separated by N,N=1~3)，對於重疊音節 N 連索引主要能擷取到文件中詞組的資訊，而間隔 N 音節的重疊音節對索引能處理中文上一些用語的彈性(如:交通大學=交大)、以及語音辨識的錯誤(如插入型錯誤) ，表 5-5 舉例

說明此六種型態的音節索引。

表 5-5: 以音節串 $S_1S_2...S_{10}$ 為例，抽取六類音節索引項

重疊音節 N 連索引	例子
N=1	$(S_1)(S_2)\cdots(S_{10})$
N=2	$(S_1, S_2)(S_2, S_3)\cdots(S_9, S_{10})$
N=3	$(S_1, S_2, S_3)(S_2, S_3, S_4)\cdots(S_8, S_9, S_{10})$
間隔 N 音節的重疊音節對索引	例子
N=1	$(S_1, S_3)(S_2, S_4)\cdots(S_8, S_{10})$
N=2	$(S_1, S_4)(S_2, S_5)\cdots(S_7, S_{10})$
N=3	$(S_1, S_5)(S_2, S_6)\cdots(S_6, S_{10})$

資訊檢索模型

向量空間模型是目前最被廣泛用於資訊檢索的模型，向量空間模型將每一篇新聞主播檢索文件視為空間中的一個向量，向量的每一個維度代表某一型態的索引項(如(S_1, S_2))在文件中的統計資訊，對於每一篇新聞主播檢索文件(document d)建立上述六類的索引項的特徵向量，可表示成:

$$\vec{d}_j = (w_{j1}, w_{j2}, \ldots, w_{jt}, \ldots, w_{jM_j}) \quad j = 1,2,...,6 \qquad (5.1)$$

$\vec{d}_j$ 代表文件的 j-th 型態的索引特徵向量。

w_{jt} 代表索引項 t 在文件 d 中的分數。

M_j 代表 j-th 型態的索引特徵項的總數。

其中索引項 t 的分數 w_{jt} 為:

$$w_{jt} = [1 + \ln \sum_{i=1}^{n_t} c_t(i)] \cdot \ln(N / N_t) \qquad (5.2)$$

n_t 代表索引項 t 在文件 d 中出現的次數，若 n_t 為零則 w_{jt} 等於零。

$c_t(i)$ 代表索引項 t 在文件 d 中出現的 i-th 的聲學信心度量測，對於字(character)和詞(word)為基礎的索引項，$c_t(i)$ 設為 1。

$[1 + \ln \sum_{i=1}^{n_t} c_t(i)]$ 為索引項 t 的頻率(Term Frequency, TF)。

$\ln(N / N_t)$ 為反文件頻率(Inverse Document Frequency, IDF)，N 為所有文件總數(對我們系統來說，N 即為我們收集新聞主播的語音文件總數)，N_t 是所有文件中有出現索引項 t 的文件數目，當索引項 t 出現在越多文件當中(即 N_t 越大 IDF 就越小)，代表它越不重要。

對於每一新聞主播音段經過音節辨識每一段音節(utterance segment 0)輸出其聲學辨識機率(acoustic recognition scores)前五大的候選音節，成為 syllable lattice，由 utterance segment 0，定義候選音節 s 其聲學信心度量測 c(s)為:

$$c(s) = \frac{2}{1 + \exp(\alpha \times [\log p(O \mid s^*) - \log p(O \mid s)])} \qquad (5.3)$$

其中 $\log p(O \mid s^*)$ 與 $\log p(O \mid s)$ 分別代表 Top-1 音節 s^* 與候選音節 s 的聲學辨識機率，$c(s)$ 範圍為 0~1，於方程式 5.2 中的聲學信心度量測 $c_t(i)$ 為索引項 t 中每一音節的聲學信心度量測 $c(s)$ 的平均。

對於查詢句子(Query)也依照上述方式表示成六個向量($\vec{q}_j$, j=1~6)，而 j-th 型態的索引特徵其查詢 $\vec{q}_j$ 與文件 $\vec{d}_j$ 的相關程度以餘弦值來評估:

$$R_j(\vec{q}_j, \vec{d}_j) = (\vec{q}_j \bullet \vec{d}_j) / \| \vec{q}_j \| \cdot \| \vec{d}_j \| \qquad (5.4)$$

最後查詢與文件整體相關性(overall relevance measure)為所有類型的相關程度的權重和(weighted sum):

$$R(\vec{q}, \vec{d}) = \sum_j w_j R_j(\vec{q}_j, \vec{d}_j) \quad j=1, 2, \cdots, 6 \qquad (5.5)$$

其中每一類索引特徵的權重需要以經驗來決定。

5.3 整合:電視新聞語音檢索系統之架構

本論文新聞語音檢索系統主要分成兩大部分:自動新聞分析系統與新聞語音檢索系統。前者其系統平台為 AMD XP 2000+的中央處理器，搭配有 1Gigabytes 的主記憶體，作業系統為 Microsoft Windows 2000 Service Pack4，其架構圖如圖 5.1，主要為新聞前處理:新聞錄影、純主播偵測與新聞切割、純主播音段音節辨識、抽取六類索引特徵，並且將相關資料寫回 Sever(第二部份檢索系統)。

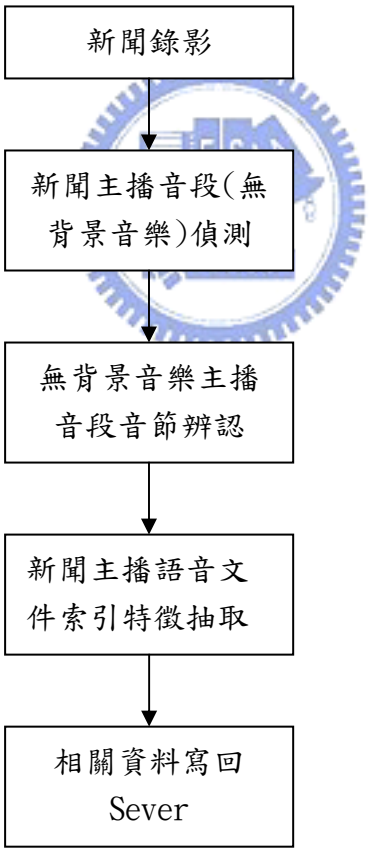

圖 5.1: 自動新聞分析系統架構(前處理)

後者為網頁式的電視新聞檢索伺服器(sever)，系統平台為 Inter Pentium4 3.0Ghz 的中央處理器，搭配有 1Gibabytes 主記憶體，作業系統為 Microsoft Windows 2000 Sever 版，其架構如圖 5.2，其流程及說明如下:使用者透過網頁輸入欲查詢的中文字串並送回網頁伺服器，透過網頁伺服器會呼叫檢索系統程式，將中文字串轉成音節並抽取六大類型的索引特徵向量，接著計算查詢與資料庫內的每一則新聞語音文件之間的相關程度，將相關性由高至低排序(ranking)，並將檢索結果傳回給使用者，使用者可透過點選來觀看查詢結果的新聞片段。

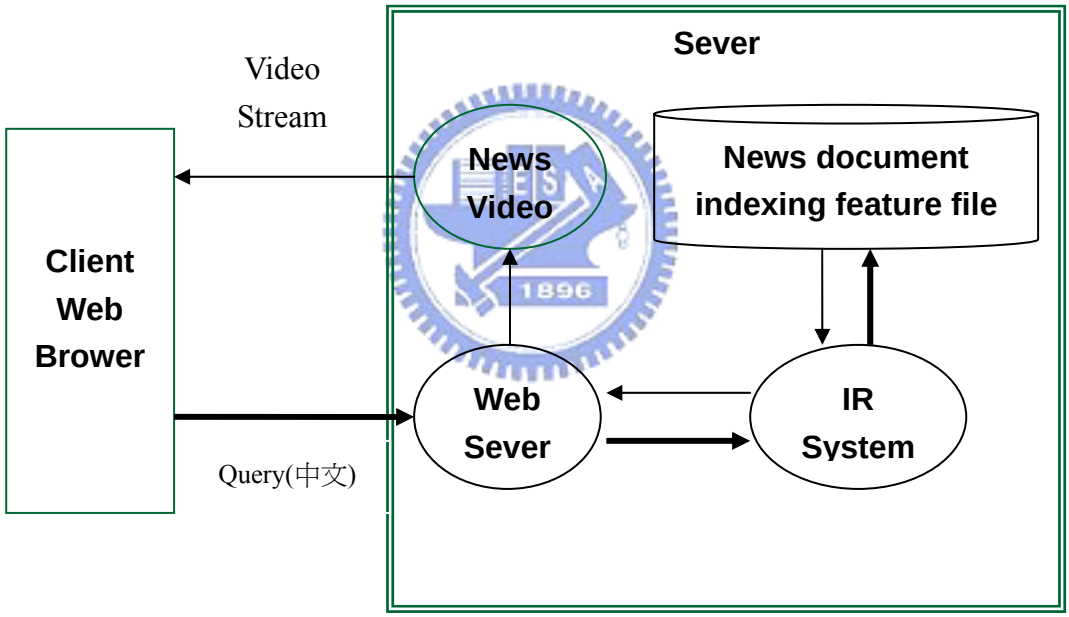

圖 5.2: 電視新聞語音檢索(sever)架構圖

5.4 語音檢索效能評估

檢索評估方法是考慮前 k 篇排序較大的檢索文件，以平均精確度(mean average precision, mAP)來評估，公式如下:

$$mAP_k = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{N_l}\sum_{s=1}^{N_l}\frac{s}{r_{l,s}}\right) \qquad (5.6)$$

L 為查詢輸入個數

N_l 為在前 k 篇檢索出的新聞主播文件中，與查詢(query)有關的新聞主播文件數

$r_{l,s}$ 為在前 k 篇檢索出的新聞主播文件中，與查詢有關的第 s 篇主播文件，從檢索排序過後數來的位置

我們目前收集了約五百段的新聞故事，而測試查詢字串共有 11 句：陳總統、趙建銘、李泰安、李雙全、二二六零、馬英九、罷免案、立委邱毅、世界盃足球賽、畢利斯颱風、上班上課，分別取 k=3,5,10，則檢索平均精確度為(mAP) 0.98、0.96、0.9。

第 6 章

結論及未來展望

6.1 結論

本論文針對前人所提的非監督式新聞主播音段偵測方法【3】作一改進，在主播音段有背景音樂的情況下，由於沒有背景音樂的主播音段與有背景音樂的主播音段其語音特徵分布已明顯不同，作 BIC 音段分群時純主播音段與有背景音樂的主播音段會被分類至不同群，而背景音樂的主播音段並非會分到同一類，因為其他外場或廣告的背景音樂可能與把主播有背景音樂的主播音段其聲音特徵分佈反而較相似，也因此前人假設最後分群後的最大群必為主播音段群已不成立，因此我們改以監督式的語者識別方式將新聞語者適當的分類以擷取新聞純主播音段，於實驗中也確時能有效的找出新聞中沒有背景音樂的主播音段。

而本論文也成功地將偵測新聞純主播的方法應用於新聞語音檢索系統之中，建立起能對新聞主播語音文件搜尋的功能，讓使用者輕易的查詢到想看的重點新聞。

6.2 未來展望

在本論文的研究與實驗和應用中，發現有數個主題是我們還可以繼續改進的重點，在此說明如下:

1. 由於一段語音當中，若存在有背景音樂勢必會影響到語音辨識的正確率，若能事先以訊號處理的技術將背景音樂的影響降低甚至消除，將會使新聞主播的偵測更加容易與準確。

2. 在語音辨識方面，由於在訓練隱藏式馬可夫模型以及語者調適時，都需要大規模甚至是需要設計過的語料資料，而本論文所用的語音資料庫本身與新聞語料並不符合，希望可以收集更大量有系統的新聞語料，訓練出更一般化語者不特定語音模型，再者可以再結合語言模型(Language model)，進一步將音節辨識出中文字(character)，將可提供更多資訊可用於資訊檢索。

3. 在語音資訊檢索方面，除了音節索引特徵更可結合字(character)與詞(word)的資訊，將可提供更多資訊以提升檢索效能。

參考文獻

【1】 D.A.Reynolds, and R.C.Rose, “Robust Text-idependent Speaker Identification Using Gaussian Mixture Speaker Models,” IEEE Trans. SAP, pp.72-83, Jan.1995

【2】 G.Schwarz, “Estimation the Dimension of a Model,” The Annals of Statistics, Vol.6 pp.461-464,1978

【3】 鄭士賢, “Model-based learning for Gaussian Mixture Model and its application on Speaker Identification,” 國立交通大學, 資訊工程研究所 碩士論文, 民國九十一年

【4】 S.Chen, P.Gopalakrishnan, “Speaker Environment and Channel Change Detection and Clustering via The Bayesian Information Criterion,” Proc. Broadcast News Trans.Under Workshop, pp. 127-132,Feb. 1998

【5】 R.Bakis, et al., “Transcription of Broadcast News Shows with the IBM large Vocabulary Speech Recognition system,” Proc. Of the Speech Recog. Workshop, pp. 67-72, 1997

【6】 M.Siegler, et al., “Automatic Segmentation Classification and Clustering of Broadcast News Audio,” Proc. Speech Recog. Workshop, pp. 97-99,1997

【7】 X.Huang,A.Acero,and H.W.Hon, “Spoken Language Processing-A Guide to Theory, Algorithm, and System Development,” Carnegie Mellon University 2001

【8】 Introduction of Hidden Markov Models
“http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html”

【9】 S.Young, et al, “The HTK Book 3.2.1,” Cambridge University Engineering Department 2001

【10】 中華民國計算機語言學會,TCC-300 國語語音資料庫, http:// rocling.iis.sinica.edu.tw/ROCLING

【11】 Chen.B., Wang, H.M.,and Lee,L.S.(2002) "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese," IEEE Transactions on Speech and Audio Processing, pp. 303-314

【12】 B.L. Chen, H.M. Wang,and L.S. Lee "A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents," ACM Transactions on Asian Language Information Processing, Vol. 3,No. 2,June 2004, Pages 128-145

【13】 Bowen Zhou,and John H.L. Hansen, "Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion," ICSLP2000 Inter. Conference on Spoken Language Processing,

【14】 Jeff. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," Technical Report ICSI-TR-97-021, International Computer Science Institute, University of Berkeley,1998

【15】 Hsin-min Wang,Shi-sian Cheng,and Yong-cheng Chen, "The SoVideo Mandarin Chinese Broadcast News Retrieval System," IJST-2002