行政院國家科學委員會專題研究計畫　成果報告

視訊網頁技術及架構之研究(I)

計畫類別：個別型計畫
計畫編號：NSC91-2213-E-009-115-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學資訊工程學系

計畫主持人：傅心家

計畫參與人員：曾政龍　吳信憲　蘇俊銘　孫聖育

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中　華　民　國 92 年 10 月 22 日

行政院國家科學委員會專題研究計畫 成果報告

視訊網頁技術及架構之研究(一)

The Study of Video Web Technologies and Architecture (I)

計畫編號：NSC 91－2213－E－009－115
執行期間： 91 年 8 月 1 日至 92 年 7 月 31 日
計畫主持人： 傅心家
計畫參與人員：曾政龍 吳信憲 蘇俊銘 孫聖育
執行單位：國立交通大學資訊工程學系

一、 中文摘要

隨全球資訊網的普及，多媒體內容也大量成長，但當面對資料量龐大的視訊(Video)內容時，仍缺乏有效且能自動化處理的技術。對於有心提供視訊資料的非電腦專業人士而言，此類網站與資料庫的建立技術門檻過高，使得視訊資訊流通以及龐大視訊資料處理上一直處於滯礙難行的困境。本計劃旨在提供一整合環境架構，可讓使用者不需要相關背景知識即可快速建立並管理一個能提供查詢功能與事件分析追蹤的數位視訊網站。在目前第一期計劃中，先完成語者交替偵測，語者辨識，故事切割等技術。並規劃網站介面，讓使用者可以輕易使用這些技術而無須了解其技術背景。

關鍵詞：

語者交替偵測，語者辨識，故事切割，資料庫，多媒體。

Abstract

With the rapid proliferation of World Wide Web application, that stimulates the growth of multimedia content. However, there still lacks a fully automatic and efficient methodology to handle a huge amount of video data. Besides that, for the unprofessional people who want to provide video content, these multimedia-processing technologies stop them to build a web site to present those data. The whole project is intent on providing an integrated architecture to help users to manage and build a web site with search mechanism easy. In the first phrase of this project we have developed these technologies such as multi-speaker segmentation, speaker identification and story segmentation. We also arrange the web interface that

makes it easy to use these technologies and the user do not to know the technical background.

Keyword:
Multi-speaker segmentation, speaker identification, story segmentation and multimedia.

二、 緣由與目的

由於網際網路與多媒體的蓬勃發展, 若擬以非數位化之視訊資料建立一可提供搜尋的數位化網站, 相關技術對於非電腦專業人士而言門檻過高, 因此本計劃旨在提供一個整合相關技術的網站, 而且提供簡易的使用者介面, 可讓視訊資料自動分類整理, 便於後端建立資料庫以及提供搜尋機制。

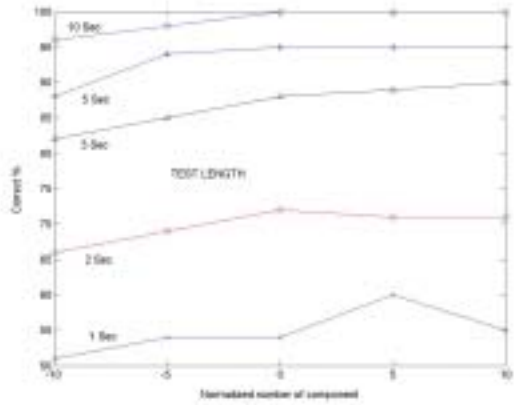
本計劃主持人以及交大資工系智慧型多媒體實驗室同仁已於四年前著手進行 Web 電視新聞資料庫完成的雛形實作, 且已有可供使用的系統上線運作。加之近年參與國家數位典藏等計劃, 對於建立前述之網站已累積相當經驗及技術。鑒於對於非電腦專業人士而言, 架設網站以及建立後端可供搜尋之資料庫可能過於繁雜不易。而面對大量的視訊資料以人工方式處理勢必耗費巨大的時間與金錢, 加之根據資策會調查指出對於全球資訊網的利用中, 以資料檢索及收集佔了七成以上比例, 故而在多媒體資料處理上透過網頁方式來展現並提供搜尋機制, 以及處理流程自動化已是勢在必行。因此本計劃擬將視訊資料處理之相關技術整合, 並將流程自動化, 使得使用者不需了解多媒體處理技術即可建立一數位化網站。

本計劃分成三個部分，分別是語音、影像處理、作業平台移轉等三部份。本期計劃主要處理語音部分，其旨在建立語者交替偵測 (Multi-speaker segmentation) 、語者辨識 (Speaker identification) 、故事切割 (Story segmentation) 。藉由發展這些技術並配合網頁介面的使用，讓使用者可以在任何地方任何時間，只要能存取網路便可透過網頁傳入欲分析的資料，毋需任何專業背景，便可得到自動處理的結果。

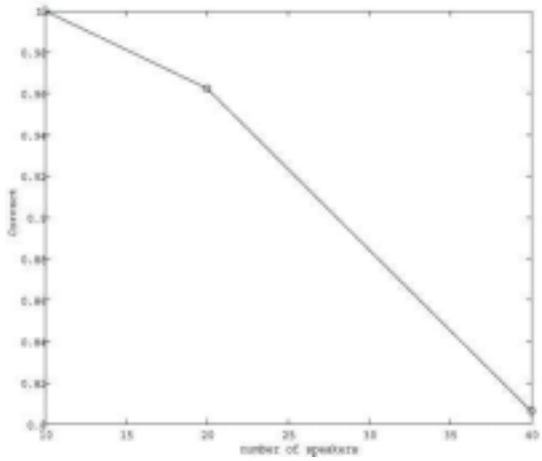
三、 結果與討論

在語者辨識上, 主要用於辨識說話者的身分。目前的做法是先取得說話者的語音採樣, 再以 MFCC(Mel-frequency cepstral coefficient)作為語者特徵。根據這些資料點在特徵空間(Feature space) 的分佈, 以 Gaussian Mixture Model 來建立語者模型, 並佐以 EM 演算法使 GMM 模型可以更接近語者特徵的分布。雖然以 GMM 模型佐以 EM 演算法訓練的方式可以有效地表現個別語者特徵, 但對於要以多少個 Gaussian components 來構成語者特徵模型, 卻一直缺乏良好的方式來決定。本實驗室利用 BIC(Bayesian information Criterion) 統計, 讓語者模型內的個數可以依照不同語者的特性而有所不同, 例如以性別而言, 女

性的個數常較男性為多。並且根據實驗可以發現，以 BIC 方式所找出的個數，均是比較好的解，過多或過少均會造成辨識率的下降。透過 BIC 統計來尋求較佳的個數方式如下：我們先以一個 Gaussian component 當作語者特徵的資料分布模型，並以 EM 演算法調整模型參數。然後對每一個 Gaussian component 計算 BIC 數值，並找出數值最大的群聚(Cluster)，檢查該群聚是否需要分成兩個新的群聚。若需要分成新的群聚，則以兩個 Gaussian components 當作群聚模型個數，最後，再以新的群聚個數重新以 GMM 模型適應並佐以 EM 演算法調整，並重複 BIC 測試，一直重複步驟到沒有任何群聚需要分離。而根據近半年來的研究，儘管是在較差的視訊品質上，仍可較過去五成至六成的辨識率提高到八成以上(86% 89%，在語者增加時辨識率會稍降)的精準度。相關結果請見下列圖示。



上圖表示以此方法找出的 Gaussian component 個數視為零(Normalization)，縱軸則是以 20 個語者辨識的正確率。從圖上可以看出，一旦測試語料長度超過三秒之後，就算增加 Gaussian components 也不能增加辨識率，可見以此法所找出的個數可相當適合資料的分布。



上圖則表示利用此法所訓練出來的語者模型，在語者增加到四十人之後辨識率仍有八成以上。

語者交替偵測可用於辨識在場景中有多少個說話者，或是對談當中說話者交替的情況。交替偵測主要是把語音做採樣後，把聲音資料以五秒或十秒為一個單

位時間做切割，以此單位時間內的資料，均抽取 MFCC 作為語者特徵，再以 Bayesian Information Criterion 的方式來統計在此時間區段內語音資料點是否可分布成兩群，若可分布為兩群則表示在此單位時間內的聲音並非同一語者所發出，故根據 BIC 統計的方式便可找出在時間區段內不同說話者的起始點，最後在輔以群聚的方法(Clustering)檢查不同段的語音資料是否可以群聚，如此便可完成說話者交替的辨認。經由語者交替偵測的結果可輔助故事切割的分析以及動態增加新的語者特徵模型。

在故事切割上，則須結合語者辨識和語者交替分析。以新聞影片為例，在新聞模式中，每則新聞常常是由主播簡介後便接著一段新聞影片。所以可利用語者交替偵測的結果便可得到說話最多次或時間最長者的語者，而此語者即可被認定為主播。主播一旦確定，根據前述規則便可把故事段落分出，而目前以電視新聞作分析的結果在切割故事段落上的誤差為一秒。其他影像資料如紀錄片式電影或是家庭錄影帶透過模式分析，以及場景變化、語者辨識和語者交替分析之後，便可將整段的影片根據場景，或是語者的交替(例如所有之前出現的語者都不在目前的片段中有聲音)和人像的辨識，將內容分割成不同的段落。

四、 計劃結果自評

本計劃針對語音研究上的難題：尋求語者模型的 Gaussian component 個數上，提出了有效率、且結果也令人滿意的方法，並實作於語者辨識上，而根據實驗的結果的確得到不錯的辨識率：隨著語者人數增加會有稍微地遞減，但辨識率均較過去的結果提昇 20%~30%。

配合此技術及相關的語音開發經驗，我們也建立一個網站可以讓使用者透過網頁介面上傳語音資料之後而得到語者交替偵測的分析。並且將此技術整合進過去本實驗室所開發之 Web 電視新聞自動分類檢索系統，對於語者辨識率，以及故事切割上，都提昇了相當的準確度。相關系統及網站均業已開放並上網，可供使用者查詢、瀏覽電視新聞。

五、 參考文獻

[1] http://www.informedia.cs.cmu.edu/

[2] Wactlar, H., Hauptmann, A., Smith, M., Pendyala, K., "Automated Video Segmentation for On-Demand Retrieval from Very Large Video Libraries," The 138th SMPTE (Society of Motion Picture and Television Engineers) Technical Conference and World Expo, Los Angeles, CA, 1995.

[3] Furht, Borivoje, "Multimedia systems and techniques", Kluwer Academic Publishers, 1996.

[4] Del Bimbo, Alberto. "Visual information retrieval," Morgan Kaaufmann Publishers, 1999.

[5] Wactlar, H., Christel, M., "Digital Video Archives: Managing through Metadata," Background paper for the Library of Congress National Information Infrastructure and Preservation Program, November 2001.

[6] 智慧型網際網路新聞視訊查閱系統的研發(The study of the Intelligent Web-Based News Video Search System), 計畫編號：NSC 89-2213-E-009-015 (I), NSC 90-2213-E-009-047 (II), 執行期限：89 年 8 月 1 日至 91 年 7 月 31 日, 主持人：傅心家, 交通大學資訊工程學系.

[7] C. Saraceno and R. Leonardi,“Audio as a support to scene change detection and characterization of video sequence,” proc. ICASSP, pp.2665-2668, 1997.

[8] H. J. Zhang, A. Kankanhalli and S. Smoliar, “Automatic partitioning of video,” Multimedia Systems, pp. 10-28, 1993.

[9] HongJiang Zhang, Yihong Gong, Smoliar, S.W. and Shuang Yeo Tan, “Automatic parsing of news video,” proceeding of IEEE international conference on Multimedia Computing and Systems, 1994.

[10] Wei Qi, Lie Gu, Hao Jiang, Xiang-Rong Chen and Hong-Jiang Zhang "Integrating Visual, Audio And Text Analysis For News Video",(Invited Paper), 7th IEEE Intn'l Conference on Image Processing (ICIP 2000), Vancouver, British Columbia, Canada,10-13 September 2000

[11] Uri Iurgel, Ralf Meermeier, Stefan Eickeler and Gerhard Rigoll. New Approaches to Audio- Visual Segmentation of TV News for Automatic Topic Retrieval. In IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City , Utah, May 2001

[12] Adrian E. Raftery, “Bayesian Model Selection in Social Research,” University of Washington Demography Center Working paper no. 94-12, September 1994.