

## 本期要目

- 壹. 學術活動預告-漢語辭彙語義學研討會 第二頁  
貳. 專文-語音辨識簡介 (王逸如) 第三~七頁  
參. 專文-口語對話系統簡介-淺談跨領域合作 (廖元甫) 第八~二十八頁

### 2005 International Conferences List

#### **ACL-43**

43rd Annual Meeting of the Association for Computational Linguistics  
Conference date: June 25 - 30, 2005  
Location: University of Michigan, Ann Arbor, USA  
<http://www.aclweb.org/acl2005/>

#### **ACM SIGIR-28**

The 28th Annual International ACM SIGIR Conference  
Conference date :August 15 - 19, 2005  
Location: Pestana Bahia Hotel in Salvador, Brazil,  
Deadline for submissions: January 28, 2005  
<http://www.dcc.ufmg.br/eventos/sigir2005/>

#### **DISS-4**

Disfluency in Spontaneous Speech.  
Conference date :September 10-12, 2005  
Location: Aix-en-Provence, France,  
Deadline for submissions: April 8, 2005  
<http://www.up.univ-mrs.fr/delic/Diss05/index.html>

#### **Eurospeech INTERSPEECH - 9**

Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology  
Conference date :September 4-8, 2005

Location: Lisbon, Portugal  
Deadline full paper submission: April 8, 2005  
<http://www.interspeech2005.org/>

#### **IJCNLP-2**

Second International Joint Conference on Natural Language Processing  
Conference date :October 11-13, 2005  
Location: Jeju Island, Republic of Korea  
Deadline for submissions: April 15, 2005  
<http://www.afnlp.org/IJCNLP05/>

#### **MT SUMMIT-10**

The tenth Machine Translation Summit  
Conference date :September 12-16, 2005  
Location: Phuket, Thailand  
<http://www.aamt.info>  
[http://www.aamt.info/mtsummit10/1st\\_anaunce.pdf](http://www.aamt.info/mtsummit10/1st_anaunce.pdf)

#### **PACLIC-19**

The 19<sup>th</sup> Pacific Asia Conference on Language, Information and Computation  
Conference date :December 1-3, 2005  
Location: Centre for Academic Activities, Academia Sinica, Taipei  
Deadline for abstract submission: July 15, 2005  
<http://pacific.sinica.edu.tw> (will soon be available)



北京大學計算語言學研究所



第六屆漢語辭彙語義學研討會  
2005年4月20日-24日，中國，廈門  
徵文通知 (CLSW6)  
會議網址：<http://clt.xmu.edu.cn/clsw6>

**主辦單位：**

廈門大學

金美 (廈門大學中文系, 副教授)

蔡駿 (廈門大學, 副教授)

**協辦單位：**

中國中文資訊學會

北京大學計算語言學研究所

新加坡中文與東方語言資訊處理學會

**會議徵文範圍：**

第六屆漢語辭彙語義學研討會的目的是彙集各相關領域的學者，探討漢語辭彙語義學各個層面的問題。本次研討會所涉及的主題包括漢語辭彙語義學的理論、方法、計算及其應用，具體包括但不限於以下所列的研究領域：

- 漢語辭彙語義學的基本理論，例如：義項、義位、義原、概念分類體系、語義特徵、語義網、語義網路、現代漢語辭彙的詞量與分級等
- 漢語辭彙語義的表示、計算和推理機制
- 漢語辭彙語義基礎資源的建造，包括 CCD、知網、SUMO 以及 WordNet 相關內容
- 義項標注的理論、技術、方法、規範和資源
- 漢語辭彙語義學在資訊提取、資訊檢索、問答系統、機器翻譯、辭典編纂等領域的應用
- 漢語辭彙語義學的新方法，包括機器學習、演化計算、神經網路等方面的研究

**大會主席：**

孫世剛 (廈門大學, 副校長、教授)

俞士汶 (北京大學, 教授)

**學術指導委員會主席：**

鄭錦全 (臺灣中研院, 院士)

**學術指導委員會成員：**

俞士汶 (北京大學, 教授)

馮志偉 (教育部語言文字應用研究所, 研究員)

董振東 (中國科學院, 研究員)

鄒嘉彥 (香港城市大學, 教授)

李堂秋 (廈門大學, 教授)

周昌樂 (廈門大學, 教授)

李如龍 (廈門大學, 教授)

賴金錠 (新加坡中文與東方語言資訊處理學會, 教授)

劉美君 (臺灣交通大學語言文化研究所, 教授)

**論文提交：**

論文可以用漢語、英語兩種語言書寫，不超過 6 頁 (包括圖表與參考文獻)，採用 PDF 或 Word 2000 以上 RTF 格式電子文檔投稿。論文詳細格式要求見

<http://clt.xmu.edu.cn/clsw6> 或

<http://machinetranslation.org/clsw6>

**議程委員會共同主席：**

蘇新春 (廈門大學, 教授)

黃居仁 (臺灣中研院, 教授)

**議程委員會成員：**

孫茂松 (清華大學, 教授)

史曉東 (廈門大學, 教授)

陳平 (澳大利亞昆士蘭大學, 教授)

Hilary Chappell (澳大利亞拉托貝大學, 教授)

姬東鴻 (新加坡資訊通訊研究所, 研究員)

靳光瑾 (教育部語言文字應用研究所, 研究員)

潘海華 (香港城市大學, 副教授)

王惠 (新加坡國立大學, 博士)

吳雲芳 (北京大學, 博士)

**論文評審：**

論文至少由兩位元程式委員會委員進行評審，根據原創性、重要性、相關性與表達清晰性進行選擇。被錄用的論文將在本次會議的論文集上發表，同時優秀論文將推薦給一些學報級刊物 (《廈門大學學報》，《語言文字應用》，《中文資訊學報》等) 上發表。會議將邀請國內外著名學者作大會報告。

**聯繫方式：**

蘇新春 教授

廈門大學中文系 (郵編：361005)

電話：0592-2184081、0592-8745786

傳真：0592-2181479

Email：[suxch@jingxian.xmu.edu.cn](mailto:suxch@jingxian.xmu.edu.cn)

**組織委員會主席：**

李紹滋 (廈門大學, 教授)

**組織委員會成員：**

史曉東 (廈門大學, 教授)

楊強 (廈門大學資訊科學與技術學院, 辦公室主任)

鄭藹兒 (香港城市大學, 博士)

陳群秀 (中文資訊學會, 副教授)

**重要日期：**

論文提交日期：2005年2月20日

錄用通知日期：2005年3月20日

提交論文修改稿日期：2005年3月31日

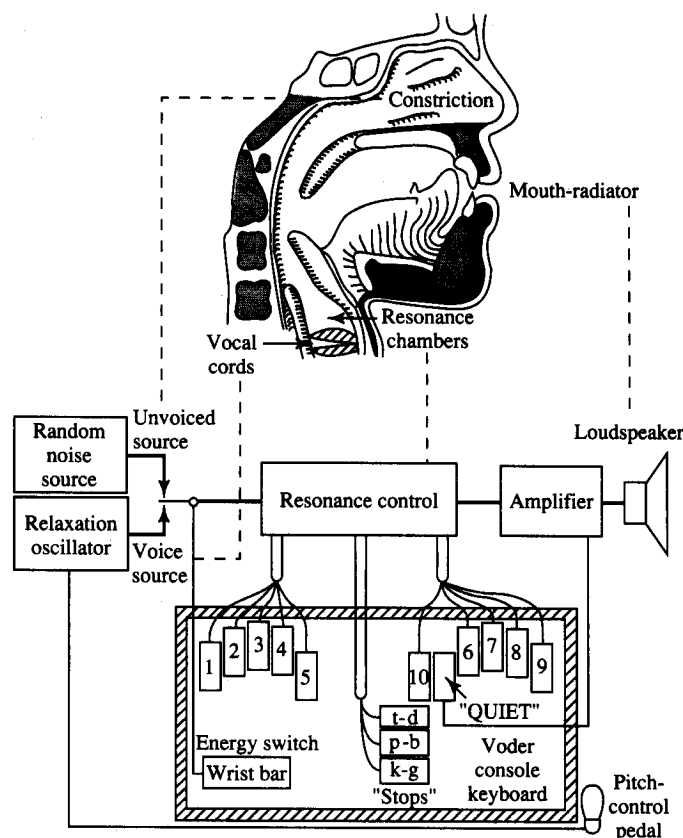
# 語音辨認簡介

王逸如

交通大學 電信工程系

1977 年導演喬治盧卡斯的『星際大戰』是好萊塢電影的一個新的里程碑，也許有人認為其主要的賣點在於其視覺科技勝過當時所有電影拍攝手法而名聞一時，也有人認為其故事內容過於膚淺而覺得不值得一看。然而，不管各方看法觀點如何，『星際大戰』這一系列的電影已經儼然已經形成了一種文化。它不但為人們開啓未來科技的一扇天窗，電影中 R2D2 與 Threepio (C3PO) 人們對語音技術的認知更是深深的刻印於人們的腦海中。Threepio 已成為人們對語音辨認技術能為人們提供之服務最刻板的印象。

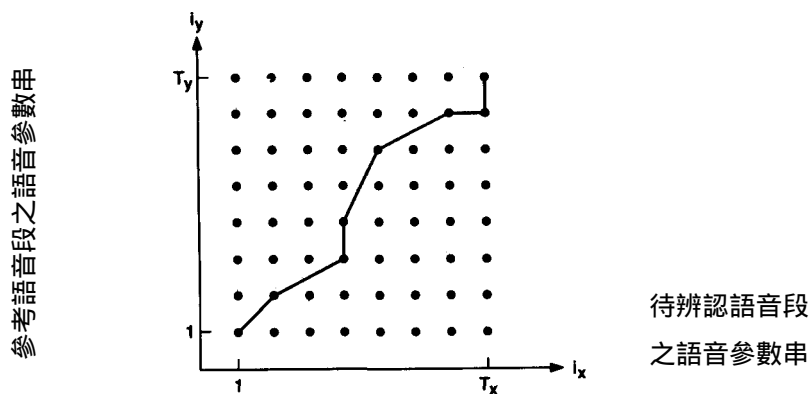
在這篇文章中，希望能對計算機科學或語音學背景的人簡單介紹語音辨認學者這數十年來所做的一些工作。讓我們來回顧一下語音辨認這個問題在歷史上的演進，從 1791 年 Wolfgang Ritter von Kempelen 提出第一個『說話機器』的架構來模擬人類發音機制，到 1940 年製作出使用他提出的模型來模擬人類發音的機器 - Voder 問世，科學家早已發展 C3PO 的工作。1955 年 PCM 語音編碼被提出，到 1962 年第一條數位傳輸系統開始使用，1960 年代不但數位電子計算機來臨，數位語音信號時代也同步開始。Atal 及 Markel 等人在 1970 年前後提出了 APC(Adaptive Prediction Coding)、LPC(Linear Prediction Coding)語音編碼技術，而線性預估(LP)中的參數也可用於描述語音的特性。就在同一時間，MIT 的林肯實驗室也在電子計算機上製作出 FFT(Fast Fourier Transform)，於是可以將短時間的語音信號轉換成頻譜。



圖一、Voder 架構圖。

於是人們就可以將短時間的語音信號轉換成 LPC 參數、頻譜參數、聲音波導管(wave-tube)參數等，就是語音特徵參數抽取。於是我們就可以將語音辨認轉換成語音參數比對的問題，語音辨認的研究於是展開了。

我們雖然可以假設語音信號在短時間內(我們稱之為音框；frame)其特性不變，可以使用上述的數學模式(如：LPC、聲音頻譜參數、…)去描述它的特性，一個音框的長度會比一個音素的長度還小，一般會取 10 – 30 msec。然而語音信號的特性會隨著時間改變，所以我們必須以一串參數(a sequence of acoustic feature)來描述一段語音信號(如一個音節；syllable 或是一個詞；word 的語音信號)；如此語音辨認就變成兩個時序參數的比對問題了。只不過一音段參數串與參考語音之參數串比對時，兩段語音其長度不見得會相同；正如我們唸同一個 word 兩次，其信號長度不會相同。更複雜的是，其中每一個音素拉長或縮短的比例也會不同。於是我們遇到一個問題是我們需要將待辨認音段中之每一音框與參考音段中最相近(相同音素)的音框做對齊(time alignment)後，再將相對應(對齊)的音框做參數比對(計算距離或相似度)，就可累加出兩段語音信號間的距離了。所幸這個尋找最佳音框對齊的問題早已有 dynamic programming 的方法可以解決，在語音辨認問題中我們加了些音框對齊的限制條件後，發展出了 dynamic time warping (DTW)技術。若一輸入語音參數串，同時對許多參考語音做比對，找出最佳參考語音—也就是辨認答案，這樣就成為語音辨認了。於是在 1960 年代工程師們使用 DTW 及語音參數比對的方法，開始了語音辨認的研究。如何尋找更好的語音參數及對更複雜的問題做最佳且有效率的比對是當時所探討的問題。而 Viterbi search algorithm 則是常用來尋找最佳路徑(音框對齊的方式)問題的演算法。在 1960 年代使用 dynamic time warping 辨認技術所探討的辨認問題多是獨立詞(isolated word)或多個連接詞彙(connected words)的問題而已。



圖二、Dynamic time warping 示意圖。

使用 DTW 方法做語音辨認事實上會有許多限制，在待辨認語音是連續語音、非特性語者或待辨認詞彙數加大時語音辨認系統的複雜度會大幅提高。所幸在 1960-70 年代在 CMU 及 IBM 兩的研究單位都提出了隱藏的馬可夫模型(Hidden Markov Model；HMM)方法。在此只能約略描述一下 HMM 的原理，HMM 模型是用一個隱藏的馬可夫鍊(Markov chain)來模擬語音特性隨時間改變的特性，在馬可夫鍊中落入每一個狀態的音框將具有相同的統計特性。圖三是一個三個狀態的隱藏的馬可夫模型，每個狀態間可以狀態轉換機率(state transition probability)來描述狀態轉移的特性；而每個狀態中語音信號的特性則以信號觀察機率函式(observation

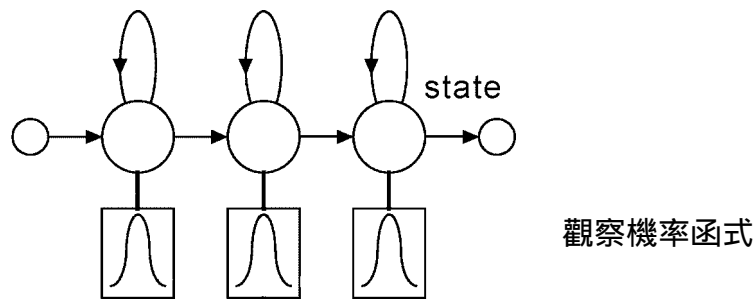
probability)來描述落入此狀態之資料的機率分佈。語音辨認問題將會變成尋找最佳狀態序列使得每輸入語音參數序列的機率最高。也就是

$$\underset{\text{All possible state sequences}}{\text{MAX}} \text{Pro}(\text{state sequence}|\text{feature sequence}) ;$$

或者再簡化成

$$\underset{\text{All possible state sequences}}{\text{MAX}} \text{Pro}(\text{feature sequence}|\text{state sequence}) .$$

我們就可以找到一串輸入語音由某個 HMM 模式所產生之機率，同時可以找到最佳狀態序列；事實上最佳狀態序列可以用來描述輸入語音哪些音框具有相同特性；也就可以將語音信號切割成數段，每一段中的信號有相同特性(可以視為是屬於相同音素)，這樣就可用來做語音的切割(segmentation)工作。



圖三、HMM 示意圖。

對每一辨認單元(如 syllable)我們都建立一個 HMM 模型；若一輸入語音參數串，同時對這些 HMM 模式求取最佳機率，在找出機率最高的模型，就可以做語音辨認工作了。也就是

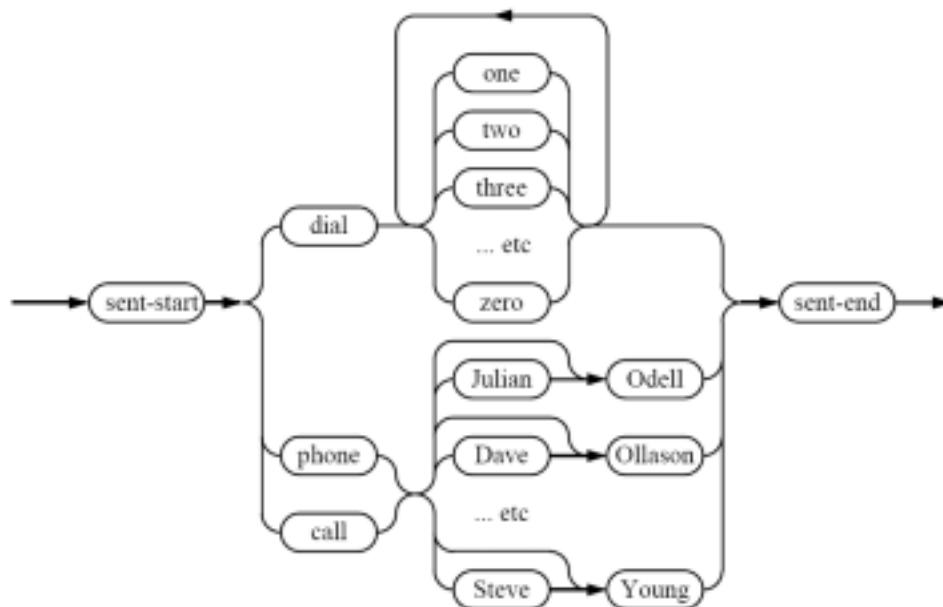
$$\underset{\text{All models}}{\text{MAX}} \underset{\text{All possible state sequences}}{\text{MAX}} \text{Pro}(\text{feature sequence}|\text{state sequence,model}) .$$

上面的 HMM 辨認問題事實上還是可以像使用 DTW 時一樣視為一個尋找最佳路徑(狀態序列)的問題。

在 HMM 語音辨認器中比較困難的問題是如何由一些訓練語料中求取辨認模型中的參數，因為對每筆訓練語音有一個資訊是我們不知道的，那就是狀態序列，於是這個問題就變成了 incomplete data 的參數估計問題。事實上，這個問題在統計學門已有解決的方法，我們可用 EM (Expectation-maximization) algorithm 來估計 HMM 模型中所需的所有參數－狀態轉換機率及各狀態的信號觀察機率分佈。

如果我們再進一步考慮更複雜的問題，以就是將音節或詞的辨認問題延伸到一句話的辨認，我們可以將一句話中音節與音節(或詞與詞)連接的規則(或機率)一併考量，我們常將這些連接規則稱為語言模式(language model)。圖四是一個 HTK (Hidden Markov Model Toolkit；一個研究 HMM 語音辨認常使用的工具軟體)使用手冊中的例子，利用一個 word graph 來描述特定語句的語言模式。如果我們知道待辨認語句滿足下列語法結構(syntax)，語音辨認問題就變成了如何在一個 word graph 中尋找一條最佳比對路徑的 search problem 了。而 word graph 中

的每一個詞又可用一個 sub-graph (HMM 聲學模式)來置換。於是圖四就變成一個具有數百個節點(node)的 graph 了。所以對 computer science 背景的人，事實上可以將語音辨認視為一個 search problem。今日我們在語音辨認中所用的國語語音辨認模型(包含聲學及語言模型)常常都是數十萬乃至百萬個節點，已經超乎 computer science 的人的想像了。



圖四、語音辨認器之語法結構圖範例。

從 1970 年代起許多使用 HMM 語音辨認方法的研究紛紛展開，例如：使用不同形式的信號觀察機率函式、不同 HMM 參數估計的方法等。所應用的範圍也從獨立詞、連接詞彙的問題進展到連續語音的辨認。由於辨認效能的提升，研究學者也將研究對象拓展至各種複雜的語音環境。一個個越來越複雜問題一一被提出來研究。語音辨認所處理的問題從特定語者、語者相關、不特定語者到語者調適；從連續數字串、voice command、朗誦語料(read speech)、廣播語料到自發性口語語料(spontaneous speech)。所面對的環境也越來越接近真實環境，從錄音室環境、實驗室環境(信號雜訊比約為 30dB)到信號雜訊更低的真實環境及廣播環境。你可以參考 HTK 網頁<sup>1</sup>中有近年來 DARPA、NIST 所做的一些語音辨認技術評比中所設定的工作及各單位所做系統之辨認效能可以略窺今日語音辨認技術可達到的效能。

語音辨認問題已經一一被解決了嗎？今日的語音辨認技術是否已可應用於一般人生活之中了呢？而業界已將語音技術用於哪些應用中或看好哪些語音辨認應用呢？現今的語音技術一般大眾對語音技術的應用充滿了高度的期望與想像。然而，在現實世界中，語音及自然語言處理技術離此期望目標仍有一大段距離。目前的語音技術應用，根據操作平台的不同，大致可以概分為三類：以資訊查詢為主的伺服器端應用、行動手持裝置上的語音命令以及應用於玩具上的聲控晶片。

以個人電腦為操作平台的語音應用，由於硬體設備擁有較強大的運算能力與連接網路的

<sup>1</sup> The HTK Book ; <http://htk.eng.cam.ac.uk/>

功能，因此往往可以結合自然語言處理技術與資料庫搜尋引擎，建立針對特定領域的語音資訊查詢系統。目前這類應用主要是以電話(行動電話)做為遠端的語音輸入裝置，透過電腦電話整合(Computer Telephony Integration, CTI)介面與語音伺服器連結。其中以自動總機分機號碼語音查詢與轉接系統最為普遍，也是目前最成功的語音技術應用。此外，結合自然語言處理技術的語音對話系統也陸續出現，氣象及飛機航班資訊查詢為其應用的例子。隨著 W3C 對語音應用標準語言 VoiceXML 的大力推廣，以 VoiceXML 所開發的伺服器端語音應用也逐漸出現，目前在國內外都可見到以 VoiceXML 瀏覽器為基礎所建構的語音入口網站(Voice Portal)。

行動手持裝置，如行動電話與個人數位助理(PDA)，已逐漸成為一般民眾日常生活的一部分，其強調輕薄短小的產品特性使得操作介面的設計受到諸多的限制，利用語音做為輸出也因此成為人性化操作介面的設計重點之一。基於硬體資源的限制，早期行動手持裝置上的語音輸入以語者特定(Speaker Dependent)語音辨識為主，由於需要使用者進行冗長的語音標記訓練，因此此應用並未獲得一般使用者的喜愛。隨著行動手持裝置硬體能力的提昇，非特定語者(Speaker Independent)語音指令辨識以及泛用形文字轉語音(General Purpose Text-To-Speech)技術均已在行動手持裝置上出現，使得其語音操作介面的設計更具彈性。目前此類應用以聲控撥號及語音指令操作為主。

在消費性電子產品中，例如玩具，成本價格一直是一項重要的考量因素，因此其使用的晶片往往是低價位低運算能力的 DSP 或 8051 單晶片。在此限制下，雖然語音技術一直被看好可以為消費性電子產品帶來新的產品設計概念，但應用上卻一直無法有效的發揮。目前雖然已有特定語者，甚至是非特定語者的語音辨識技術應用於玩具上，但由於在遠距語音輸入以及環境噪音問題上仍無法有效解決，因此這類型的語音應用雖有其潛在市場及未來性，但短時間內可能仍無法有大量的產品出現在市場上。

由上面種種可以看出，語音辨認技術還有一段漫長的路要走，除非有一項其它劃時代的技術如：腦神經學家能將人類腦部信號完全解讀並連接至機器，否則語音仍是最自然的人機介面，人們仍在期盼 Threepio 能早日誕生。

## ↪ 口語對話系統簡介- 淺談跨領域合作 ↩

廖元甫

台北科技大學電子系

### 1. 簡介

口語對話系統使用語音處理相關技術，直接以語音輸出輸入方式，建立方便使用的人機介面，其目的在使電腦能理解人類的語言與語意，並以代理人身份與使用者應答，自動幫助使用者擷取資源或是達成某樣工作。例如圖一的口語汽車導航系統，可以讓駕駛人在行車中，以口語詢問交通資訊或操作汽車導航系統，系統也會用語音回應，避免駕駛人需以手操作傳統人機介面而發生危險。



圖一. 口語汽車導航系統使用情境

在國內外，口語對話系統相關的研究很多，也有許多已經實際上線使用中的系統，包括最早的飛利浦的火車時刻表查詢系統（Philips Automatic Train Timetable Information System, 1994）[1]，之後的麻省理工(MIT)[2]的Mercury飛航時刻表查詢系統(Air Travel Information System, ATIS, 1993 至今)，可讓使用者透過電話用語音查詢時刻表並直接定票。此外最近的貝爾實驗室（Bell Lab.）客戶服務中心自動轉接系統，可在電話中直接判別使用者所說的服務需求，並轉接到正確的部門[3]。此類系統的典型對話流程如圖二所示（以 ATIS 為例），系統會以一開放式問句（open question）開場，如“*How can I help you?*”，然後等使用者回應，通常也允許使用者以自發性口語（spontaneous speech）詢問，並可以容忍非正規文法形式（illegal form）的語句。



[Sys] Hi! Welcome to Pegasus. How can I help you?  
[Usr] Are there any flights from Boston, sorry, New York City to Boston arriving around one pm?  
[Sys] There are 4 flights from New York arriving in Boston at 1'clock pm local time. US airways flight 6812 from La Guardia, ..., and Delta flight 1830 from Kennedy are possible choices. Please select one.  
[Usr] Can you tell me if delta flight, hmm, 1830 will be on time?  
[Sys] Delta flight 1830 from Kennedy is expected in Boston at 12:14 pm. It is 26 minutes ahead of schedule. Is there something else I can do for you?

.....

圖二. 口語對話查詢系統的典型對話流程

若粗略分割口語對話系統所使用的技術內容，通常可分為語音信號處理（speech processing）與自然語言處理（natural language processing）兩個主要領域，前者由語音信號（speech signal），求取語音特徵參數（feature），建立語音單元模型（model），以比對（search）方式找出說話者所說的文字語句（text sentence），通常由電機（electronic engineering）或資訊工程（computer engineering）作訊號分析的人研究。後者則處理文字資料，常由語言學（linguists），與計算語言學（computational linguistics）等研究人員，經由文法（grammar），語意（semantic），語段（discourse）等分析，找出人、事、時、地、物等內容，以嘗試理解說話者欲傳遞之意向（intension）與目的（goal）。

但是口語對話系統要能夠成功好用，需要語音信號處理與自然語言處理兩領域的研究者共同合作，然而跨領域合作通常不大容易，最常見的問題是，雙方不瞭解彼此的語彙與觀點，以至雞同鴨講溝通困難，或是不清楚對方所能提供的處理結果，導致各模組輸出輸入定義不同，系統各模組間不知如何連接整合，此外亦可能過份高估對方技術的極限，以致己方忽略處理某些問題，卻滿懷希望問題能全由對方解決，等等。

因此在本文章中，首先將嘗試以系統觀點簡介紹國內外口語對話系統相關研究與應用，然後說明口語對話系統的一般架構，並著重在不同領域間的分工與整合問題，再以一簡單的應用實例，解釋如何利用一些工具程式庫，尤其是公開程式碼（open source code），建立一口語對話汽車導航系統。希望能拋磚引玉，激發更多語音信號處理與自然語言處理兩領域研究者的跨領域合作機會。

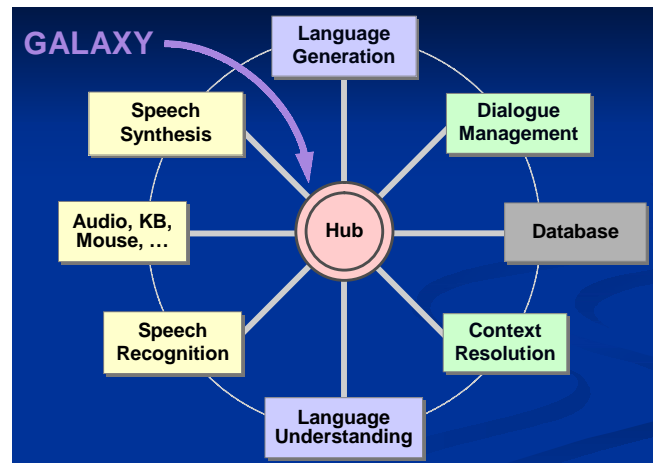
## 2. 國內外相關之研究與應用

常見的口語對話系統主要應用在資料查詢，人機介面，與跨語言對話翻譯上，包括一般電話環境，車用環境與嵌入式系統的口語對話系統，並被應用在個人電腦（PC），個人數位助理（PDA）與手機上。以下進一步介紹國內外各種口語對話系統相關的研究與應用。

## 2.1. 國外相關之研究

最早商業化的系統當屬 Philips 的火車時刻表查詢系統[1]與自動電話總機系統（Philips Automatic Directory Information System, PADIS) [4]，其可以口語詢問電話資訊轉接分機，其甚至可以使用自然數字（natural number），或以拼音方式拼寫人名。

此外最著名的例子當屬之前提及之航空運輸資訊服務系統，其為美國國防部高等研究計畫局（Defense Advanced Research Projects Agency, DARPA）資助的 Galaxy Communicator [5] 計畫的一部份，此計畫有多所大學參與，包括 MIT [2]，Carnegie Mellon University（CMU） [6]，University of Colorado at Boulder [7] 與 SRI [8]等等，DARPA 並在 2000 與 2001 年辦過實際評比[5]，最後釋放出其使用之 Galaxy Communicator Software Infrastructure (GCSI) 工具程式庫 [9]，如圖三，可供大眾自由使用，我們之後也會使用此工具程式庫。



圖三. Components of Conversational Systems and Galaxy Communicator Software Infrastructure (GCSI)

此外麻省理工另有一套 Jupiter 氣象查詢系統[2]，它主要是透過電話使用語音查詢世界各大城市氣象相關的資訊，包括氣象報告及回答氣溫、風速、降雨量、氣壓、溼度、日出時間等。此外它也可以具備多語言翻譯的能力，可接受多語言輸入，或將氣象資訊翻譯成其他不同語言輸出。

相對於美國，在歐洲也有結合十數個大學，研究單位與公司的大型研究計畫，例如 1993~2000 年德國科專 Verbmobil [10] 計畫，及後繼之 SmartCom[11]計畫，致力於智慧型多模式（multi-modality）人機互動科技，目的在使德國在語言科技上取得領先地位，其發展出三種應用情境，包括 SmartCom-Mobile，SmartCom-Office/Home 與 SmartCom-Public，分別將多模式對話系統用在行動中，辦公室或家庭與公共場合使用，如行動中使用手機或 PDA 獲取資訊，如圖四；在家中則透過平板電腦，控制家電，如圖五；在辦公室，則使用嵌在桌面的液晶螢幕，如圖六；在公共場合，則設有如電話亭的資訊站，如圖七。



圖四. Design Study of a SmartKom-Mobile



圖五. Photo of Prof. Wahlster with SmartKom-Home

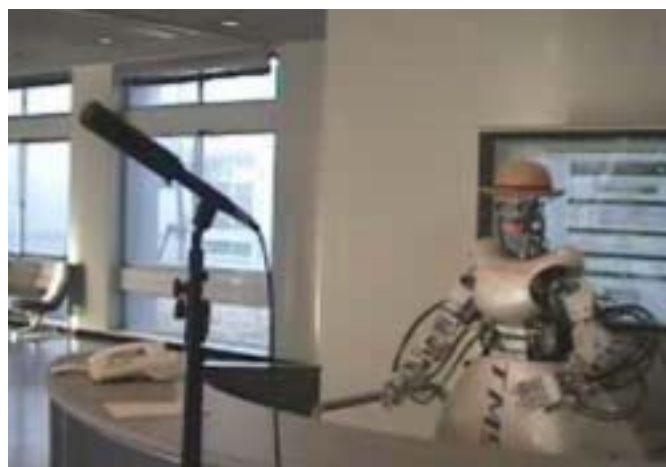


圖六. Design Study for a SmartKom-Home/Office



圖七. A Multimodal Communications Booth: SmartKom-Public

在日本方面，除口語對話系統外，更進一步發了不少力氣在製作機器人上，如奈良先端科學技術大學院大學 Shikano 教授發展之 Human Robot Communication Receptionist Robot (ASKA) [12]，如圖八，可代替接待員在服務台與人對話，提供校區指引，並以語音，手勢與表情回應。

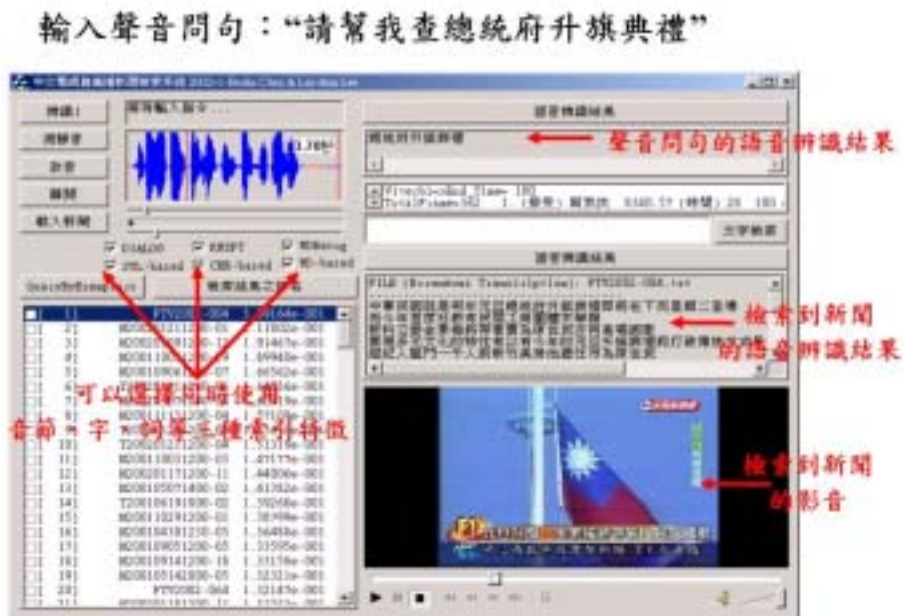


圖八. Human-Robot Dialog System (ASKA)

## 2.2. 國內相關之研究

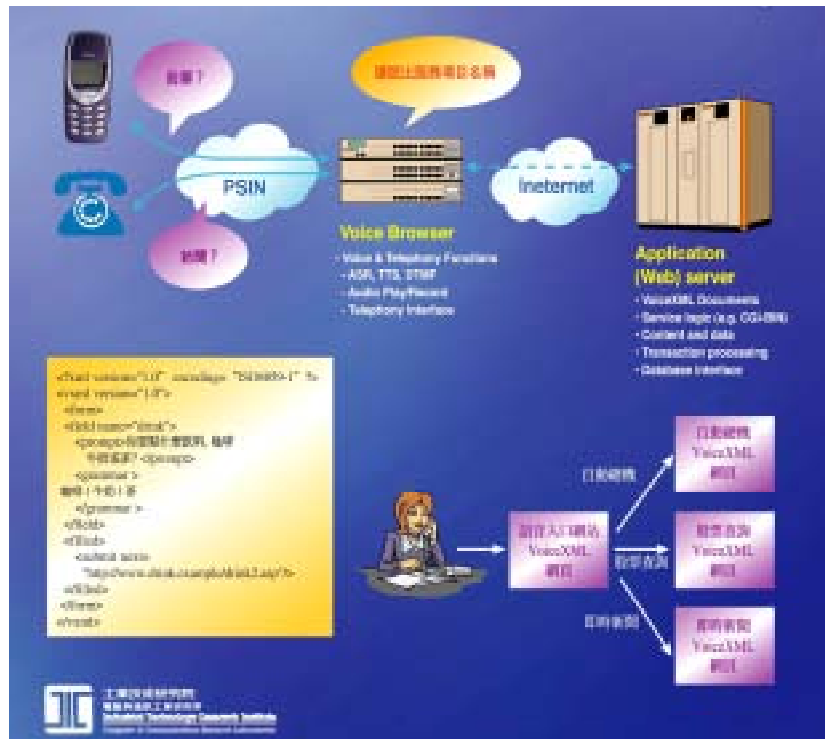
國內在口語對話系統的應用方面，也都是著重在資訊查詢系統，尤其是新聞檢索，但較注重在實用方面，包括電話語音自動總機系統，語音入口網站，e 化家庭，車用語音對話系統與嵌入式系統人機介面上，跨語言對話翻譯的活動較少。研究單位則包括台大、交大、成大、師大、北科大、中研院、工研院，資策會、中華電信、宏碁、台達電、賽微與聲碩等。

台大，師大與中研院的近年來致力於口語新聞多媒體檢索系統 [13,14]，如圖九，其可將每天播出的新聞多媒體自動作辨認，摘要與分類，並放置到資料庫中，讓使用者可以以口語做檢索，自動調出相關之新聞語音或影像。



圖九. 中文影音多媒體資訊檢索之雛形展示系統

工研院前瞻技術研究中心開發的語音入口網站（voice portal），如圖十，其為一中文電話語音對話系統，應用範圍廣泛，主要有旅遊資訊查詢、網際網路線上資訊查詢、電話語音查詢、個人語音秘書、金融資訊查詢、電子銀行、語音電子商務、醫院掛號等。



圖十. 電通所前瞻技術中心語音入口網站

工研院的 eAttendant [15]，北科大的語音電話總機系統[16]（見圖十一），與中華電信的自動 104 查號台等，則可透過電話線路查詢某個人名的分機號碼並自動轉接，皆已經實際上線供大眾使用多時。

## NTUT Automated Telephone Directory Attendant System

☎ +886-2-27717121 ext. 5104



- On-line service for the whole campus since May 2004
- ~400 calls per week so far
- ~2000 entries
- Research playground and real data collection

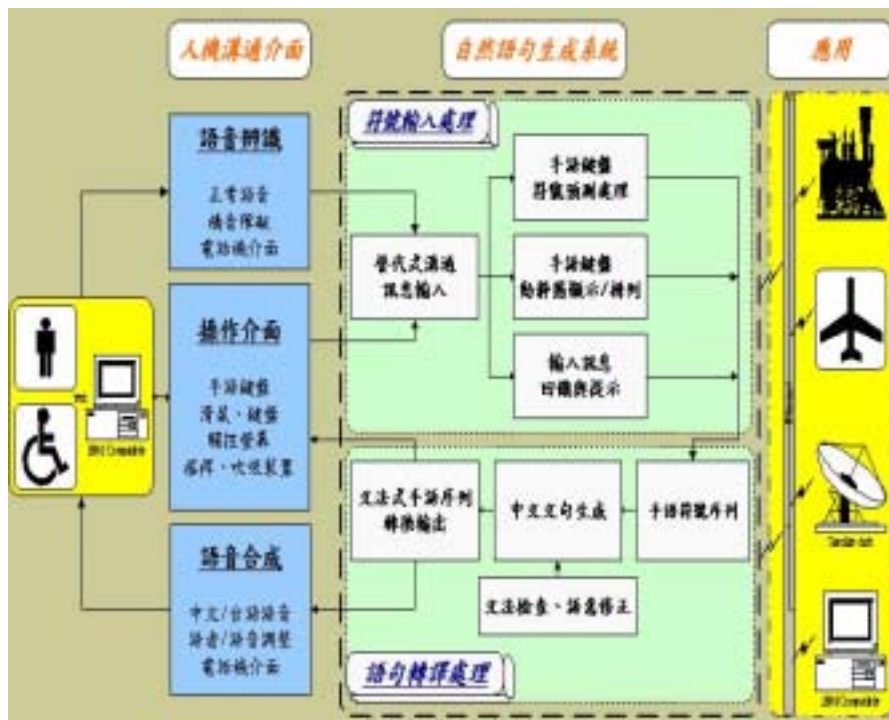
圖十一. 台北科技大學語音自動總機系統

交大與北科大的智慧型汽車導航系統 [16,17]，則是在車上環境，利用口語對話系統，控制全球衛星定位系統，電子地圖與行動電話通訊錄，以輔助駕駛人獲取行車時所需的導航資料，甚至是直接以語音撥打電話對外聯絡，如圖十二，可增加行車便利與安全。



圖十二. 智慧型口語對話汽車導航系統，使用者導航中撥打電話，電話撥通時，鈴聲響起

成大則甚至結合手語辨認等，以多模式輸出輸入方式，製作聽語障輔具 [18,19]，如圖十三。其可同時接受口語與手語輸入，並給予相對的語音或手語輸出。



圖十三. 聽語障科技輔具

### 2.3. 大陸相關之研究

大陸語音相關之研究，因多家外商公司在大陸接連成立研究中心，包括微軟亞洲研究院[20]，IBM 中國研究院等等[21]，吸引大量優秀人才投入，正蓬勃發展中，不論研究人員數量與經費總數皆已遠超過台灣。主要的研究單位有中科院自動化所模式識別實驗室 [22]，清華大學

語音技術中心 [23,24]，安徽科大訊飛公司 [25] 等等。

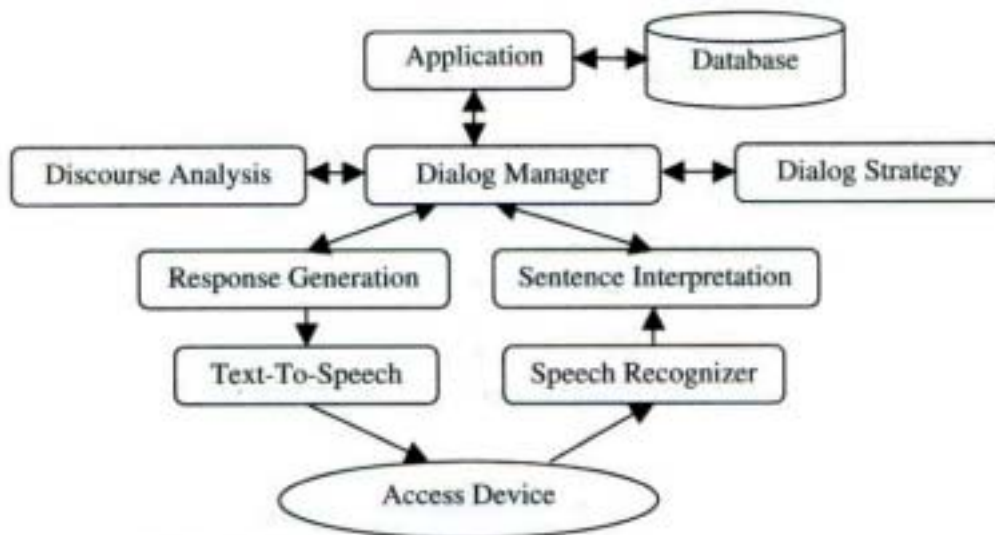
尤其是大陸為迎接 2008 年奧運的到來，正進行“面向奧運多語言資訊服務網路系統”的多國語言機器翻譯計畫，此計畫由大陸科技部 863 作為重點項目支持，涉及語音、翻譯等相關技術和產業化集成應用，能提供漢語、英語、法語、西班牙語等世界主要語言，還能提供多語言訊息同步發布、訊息查詢和語音交互式的電子商務服務，並重點提供基于位置的訊息服務，另外還提供人際交流的輔助工具，如口語翻譯機和自動翻譯電話。

### 3. 口語對話系統架構

以下進一步說明口語對話系統，其大抵需要下列如圖十四所列之元件，包括語音辨認（speech recognition），文句語意解析或稱為自然語言處理（natural language processing），對話管理（dialogue management），自然語言產生（natural language generation）與文字轉語音（text-to-speech）等等模組。

其運作如下，語音辨認模組接受使用者之語音輸入，將其轉換成文字，再由自然語言處理模組，依前後文解釋其語意，對話管理模組則依據使用者之語意，判別其行為意向，或是進一步與使用者互動，釐清使用者之計畫（plan），幫助其達成目的（goal）。互動過程中，則由自然語言產生模組，產生適合之回應文字，在交由文字轉語音模組，合成自然流利之語音。

因此口語對話系統架構複雜，且各模組實際上就幾乎是一個不小的研究領域，各有各的專門研究人員，因此要完成一良好的系統，通常需要各領域的專家一起合作來完成。以下將先進一步說明各模組的原理，再介紹如何連接各模組，尤其是語音辨認模組與自然語言模組的連接方式。



圖十四. 口語對話系統架構圖

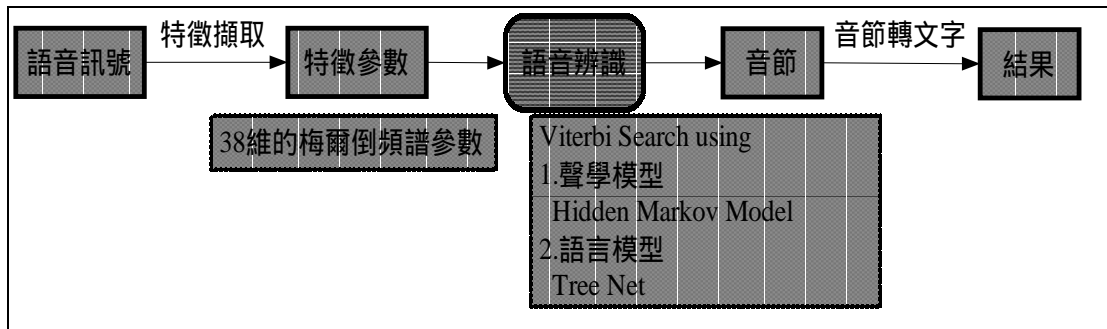


### 3.1.1. 語音辨認模組

語音辨識的功能在於將輸入的語音信號轉為文字輸出，其架構如圖十五所示。語音辨識首先藉由大量的語料，求取其特徵參數，訓練出語音聲學模型（Acoustic Model）及語言模型（Language Model），設計這些模型時依應用需求，可使用不同字彙量的連續語音辨識引擎。語音辨識模組就是利用這些引擎比對輸入的語音訊號，找出最可能的文句。

此外在辨認時通常會在聲學模型中加入加上一些無實際意義的音（particles）如嗯（hmm），ㄟ（eh）等填充模型（filler model），以容忍自發性語音的特性。或是在語言模型加上一些此領域的特殊文法限制，縮小比對範圍。或是加入填充詞（filler word），以對付不在辨認辭典中的詞彙（out-of-vocabulary, OOV），以加強系統的強健性。

此外在口語對話系統應用中，通常應用領域（domain）一但定義下來，則已經限制了所需辨識的詞彙數量，因此所需辨識詞彙通常不會太多。例如在車用語音對話系統開發中，對話情境通常設定在車用對話的領域，如車上語音命令控制，地圖、導航應用等而已。

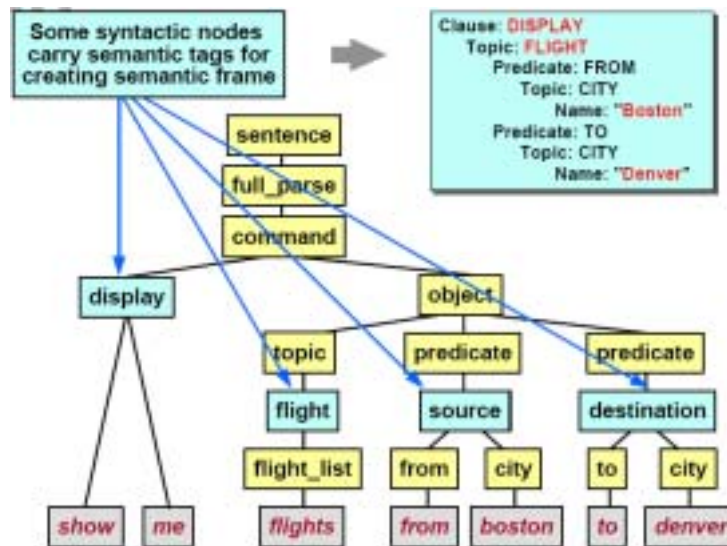


圖十五. 語音辨認模組架構

### 3.1.2. 自然語言理解模組

自然語言理解在對話系統所扮演的角色，在於將經由語音辨識引擎辨識出的文字轉為有意義的結構化資訊，填入對話控制模組的語意欄位（slots）中，如圖十六所示之文句解析過程。找出包括人、事、時、地、物等，將其轉換成使用者的意圖，如此才能做後續的對話處理。

此外，除自發性語言常會有不合乎文法規則的句子出現外，語音辨識錯誤的產生通常無可避免。因此自然語言理解必須能做強健性文句剖析（robust parsing），例如加入填充詞，或是只擷取有意義的部分，避開無意義的部分，以對付不合乎文法規則的句子，來加強系統的強健性。



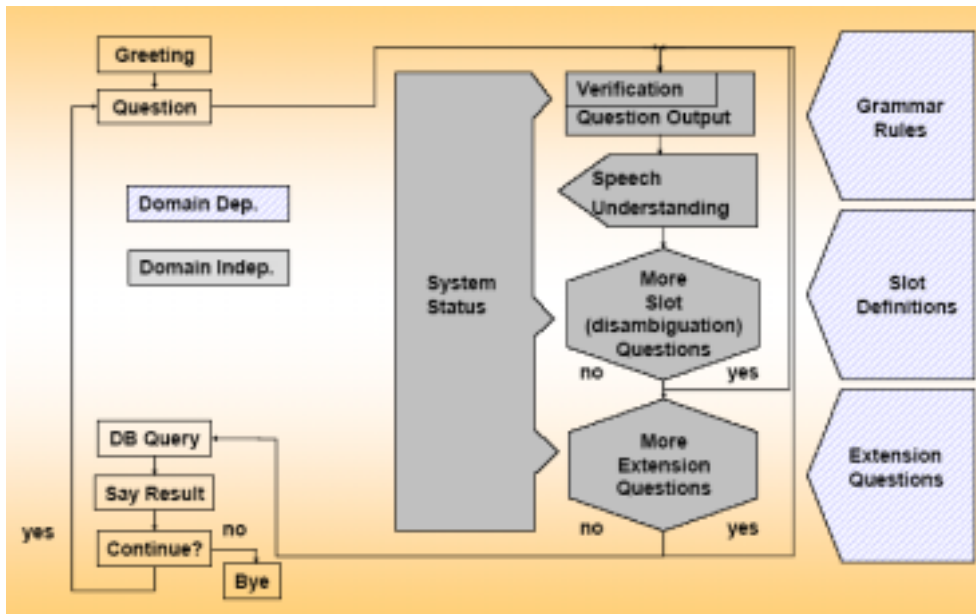
圖十六. 自然語言理解模組之文句剖析與語意資訊擷取

### 3.1.3. 對話管理模組

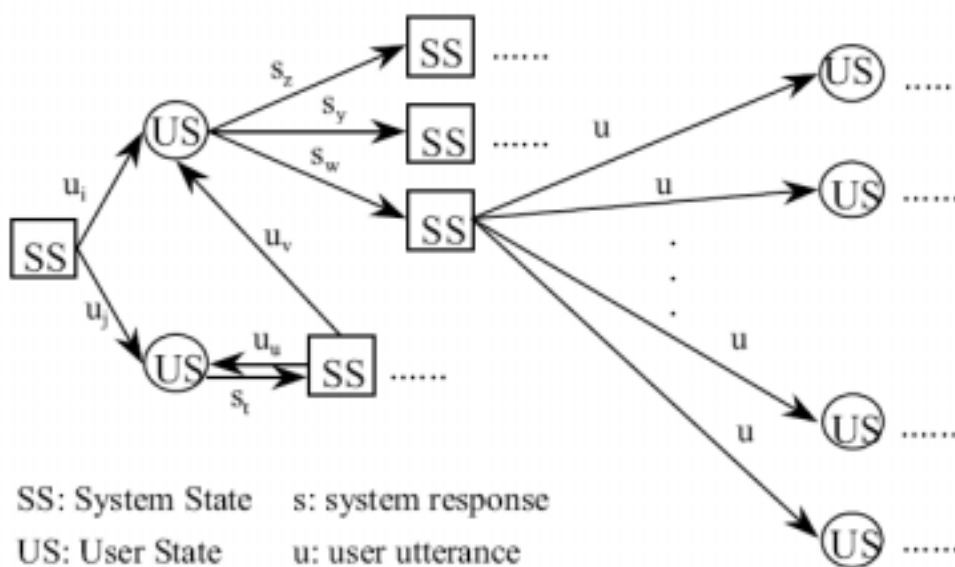
對話管理模組主要是管理整個對話流程，其架構如圖所示十七，通常先依所需理解的資訊，定義一具有數個欄位 (slots) 的表單，當作系統的狀態 (status)，再依據使用者可能產生的事件 (event)，定義可能發生的狀態 (state)。並如圖十八所示，定義每個狀態的進入條件 (condition)，相對應產生那些對話來回應 (prompt)，以及需從自然語言語言理解擷取更新那些欄位，或採取相對應的行動 (action) 等等。最後靠更新後的表單欄位的內容，驅動如圖十八的有限狀態機 (finite state machine)，在不同狀態中運行。

通常在對話流程中，如果使用者所提供的資訊仍然不足，對話管理模組會再依其所需要的資訊欄位詢問使用者，從而在有限狀態機中做轉移，與使用者應答做互動。等到系統在獲得使用者的意圖和完整的欄位資訊後，就可以採取相對應的行動，和各種不同類型的知識伺服器端連結，如應用資料庫、網路資源、知識處理、資訊檢索等，來達到使用者的目標。

實際上，有限狀態機的狀態定義最為複雜，也最令人頭痛，常有漏網之魚情形發生。因此不適合直接實作在程式碼中，目前多數的實作方式，皆採程式碼與對話狀態控制定義分離的方式，來保持良好的維護性。尤其是業界大多採用 VoiceXML 標準，以 VoiceXML 來定義欄位，狀態與對話流程，再以 VoiceXML interpreter 來編譯與建立有限狀態機 [26,27]。



圖十七. Block Diagram of DM architecture



圖十八. Finite State Conversation Network

### 3.1.4. 自然語言產生模組

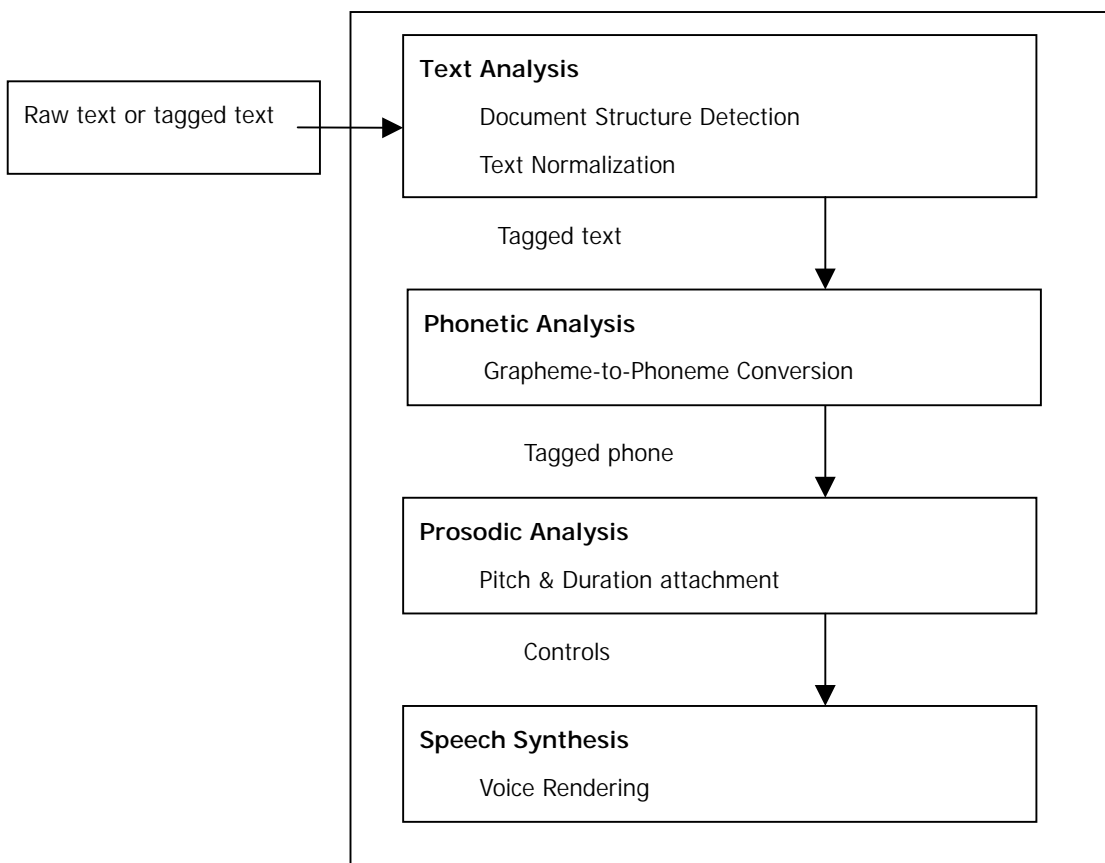
自然語言產生模組負責產生適當之回應文句，通常依對話管理模組的表單欄位資訊，與所處的狀態，使用樣本（template）方式產生文句，一般在每個對話狀態中定義各自的樣本。

### 3.1.5. 文字轉與語音模組

在對話系統中，通常使用文字轉語音技術將自然語言產生模組的文句轉換成語音，其優點在於可將任意輸入的文字轉為語音格式輸出，有較大的使用彈性。其架構圖通常如圖十九所示。其先對輸入文句做剖析，找出如斷詞位置，詞性，發音，音調等語言參數（linguistic features），

再依語言參數做韻律分析，產生適當之韻律訊息，最後合成自然流利之語音輸出。

目前語音合成多使用從大量語音資料庫擷取適當語音段落串接合成方式（**corpus-based TTS**）[28]，因為其是從真正語音資料選取適合音段，所以優點是聲音自然流利，缺點是所需資料量很大，至少需 500M bytes 的硬碟空間，有人甚至用到 20G bytes。

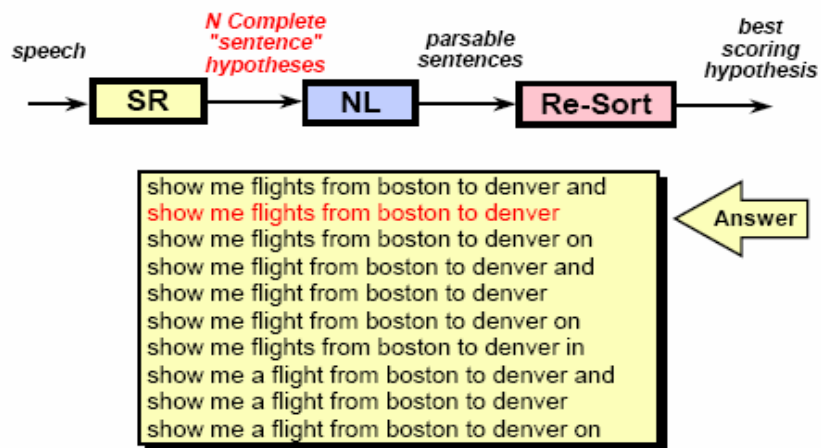


圖十九. 文字轉與語音方塊圖

### 3.2. 語音辨認模組與自然語言處理模組之連接方式

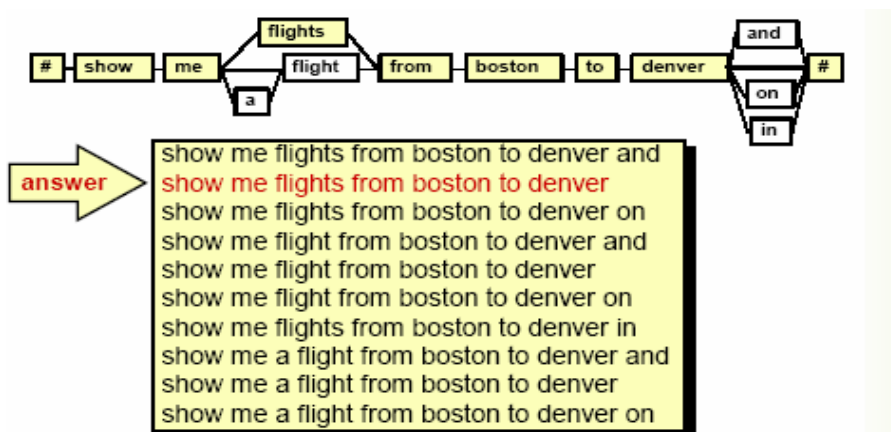
語音辨認模組與自然語言處理模組之連接，最簡單的是語音辨認模組只產生出一句文字（**sentence**），自然語言處理模組也只接受此文字句子。如此語音辨認模組與自然語言處理模組都可以保持原來個別發展的狀態，這是最簡單的連結方式。

但是考慮到口語對話系統的應用環境，可能是在車上，街上，或是在吵雜的環境上，除前面所提之自發性語言問題外，語音辨認模組還可能因雜訊而產生其他錯誤。因此通常口語對話系統會輸出最好的前幾名句子（**nbest**）與其分數大小，如圖二十，以求將正確答案涵蓋進來，然後再讓自然語言處理模組依自然語言知識計算出另一組分數。最後將兩分數加總後，重新排序（**rescoring**）所有句子的分數，找出真正的答案。



圖二十. 以 nbest 方式結合語音辨認模組與自然語言處理模組

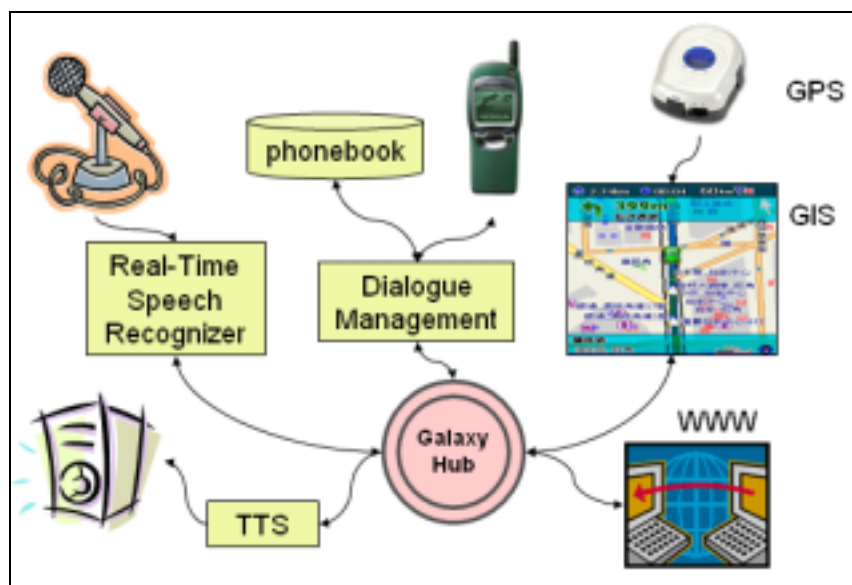
此 nbest 結合方式，只要稍稍更改語音辨認模組與自然語言處理模組，將其以迴圈方式重複執行即可，但會增加很多計算量。另一種方式是以圖二十一之 wordgraph 方式連接，先將 nbest 轉成 wordgraph，則可避免使用迴圈，大量減少重複的分數計算。



圖二十一. 以 wordgraph 方式結合語音辨認模組與自然語言處理模組

#### 4. 口語對話汽車導航系統建構實例

以下我們以一實際的例子，討論的系統建構過程，與介紹一些好用的公開領域工具程式庫。此處以交大與北科大的口語對話汽車導航系統為例，其應用情境如圖一所示，主要是在車上利用口語對話系統，控制全球衛星定位系統，電子地圖與行動電話通訊錄，以輔助駕駛人獲取行車時所需的導航資料，甚至是直接以語音撥打電話對外聯絡，增加行車便利與安全。其架構圖如圖二十二所示。包括一即時語音辨認器，以辨認出使用者的語音，並以文法樣版（grammar pattern）方式做簡單的語意理解；對話管理模組則做對話流程控制，並控制大哥大與汽車導航系統，或是透過網路擷取資訊；文字轉語音模組則將對話管理做出的回應轉換成語音輸出。以下先說明資料收集程序，再介紹在建構此系統時，所使用到的公開領域的工具程式庫。



圖二十二. Block diagram of the spoken car navigation system

#### 4.1. 口語資料收集與行為分析

在建構口語對話系統的過程中，最重要的是先定義要處理的應用領域，並在此領域中收集口語對話資料，分析所需要用到的詞彙與文法，以建立語音辨認模組，並決定如何解析每個句子的表層語意（*surface meaning*），與決定如何利用前後文找出每個句子的深層語意（*deep meaning*），以建立自然語言處理模組，分析所發生的口語行為。

當然要正確地收集在某一特定應用領域的口語行為，最好是有一實際系統，讓使用者使用，並將實際互動過程記錄下來。但在系統建構初期，不可能有任何系統可以讓使用者使用。基本上這是一個雞生蛋，蛋生雞的問題，因此必須改以 *Wizard of OZ* 方式收集[29]。作法是先找人做角色扮演，假裝是真實的系統，與使用者互動，並將互動過程完整記錄下來，以供日後分析。

例如，在此口語對話汽車導航系統的例子中，我們以如圖二十三的系統設定，收集人機對話資料。其中駕駛人端只能見到目前位置附近的小範圍的導航地圖，其它導航資訊則完全由 *wizard* 端控制，以模擬一般全球衛星定位系統（*GPS*）導航系統的情境。*Wizard* 基本上模擬導航電腦，會適時給與駕駛人適當的資訊。但駕駛人也可以隨時與 *wizard* 對話來取得進一步導航資訊。我們在收集資料初期先邀集了約十數個使用者來試用，每人試用數次，並將過程全部錄音下來，再作標記（*tagging*）。



圖二十三. Data collection using Wizard of OZ setting

圖二十四列出以此方法收集到的一段典型對話資料，圖二十五則列出在此領域中，分析出來的可能口語行爲。由圖二十四中可看出口語行爲的多樣化與複雜性。理論上資料是越多越好，才可能涵蓋所有可能情形。但是以 Wizard of OZ 方式畢竟不太自然，所以當收集到一定量資料，可以建構出第一版的系統後，便應改以 Wizardless 方式，用第一版的真實系統，再次找人試用收集資料，以調整系統。整個流程如圖二十六所示。實際上這整個流程必須重複多次，直到達到收斂爲止。

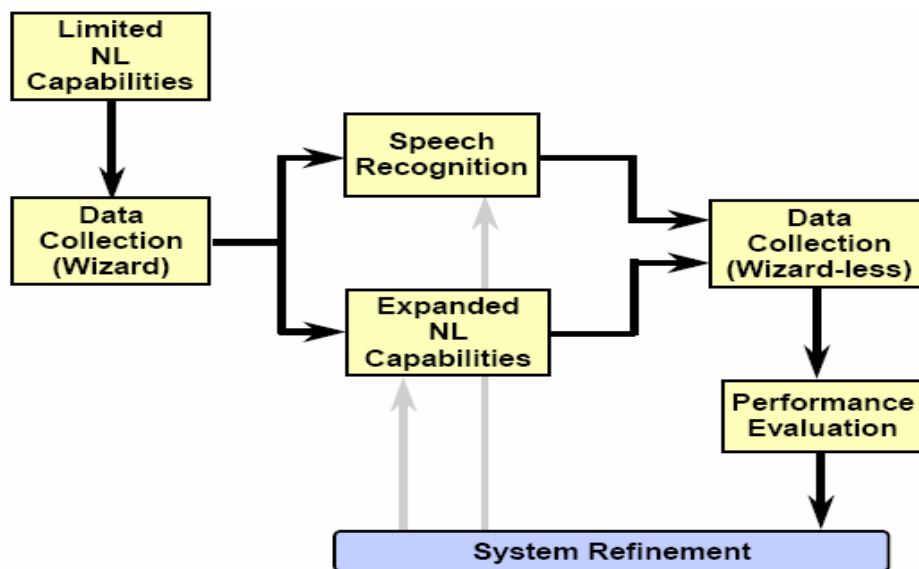
system_opening	s:	去哪裡
user_inquiry	u:	現在要去高雄從高雄火車站出發到高雄師大
system_hold	s:	MHME 請 E 系統規劃中請稍候
system_hold	s:	E 導航準備完成
system_navigate	s:	請沿 E 請沿中博地下道直走
user_info_request	u:	中國地下道在哪裡請問
system_answer	s:	就是你現在的位置
user_confirm_request	u:	O 我現在在中國地下道直走
system_navigate	s:	對往 E 到八德一路左轉
user_inquiry	u:	E 到什麼路左轉
system_answer	s:	八德一路
system_route_correct	s:	E 妳妳的方向錯誤是往妳的後後方走
user_repeat	u:	八 MHMHM
system_navigate	s:	請回轉

圖二十四. 典型的車用導航系統情境對話資料

system_opening (系統開場)	user_opening (用戶開場)
system_hold (系統查詢中)	user_inquiry (用戶查詢路線)
system_navigate (系統導航)	user_repeat (用戶重複)
system_answer (系統回答)	user_info_request (用戶要求資訊)
system_route_correct (系統更正路線)	user_confirm_request (用戶要求確認)
system_route_remind (系統提醒路線)	user_route_confirm (用戶確認路線)
system_route_confirm (系統確認路線)	user_answer (用戶回答)
system_destination_arrive (系統宣告到達)	user_destination_arrive (用戶宣告到達)
	user_self-talk (用戶自言自語)
system_closing (系統結束)	user_closing (用戶結束查詢)

1. 需為系統所有動作寫下標準句子。
2. 系統與用戶因為彼此互動角色不同標記並不對稱。
3. 此標記系統必須視標記狀況修訂。

圖二十五. 車用導航系統情境口語對話行為分類標記



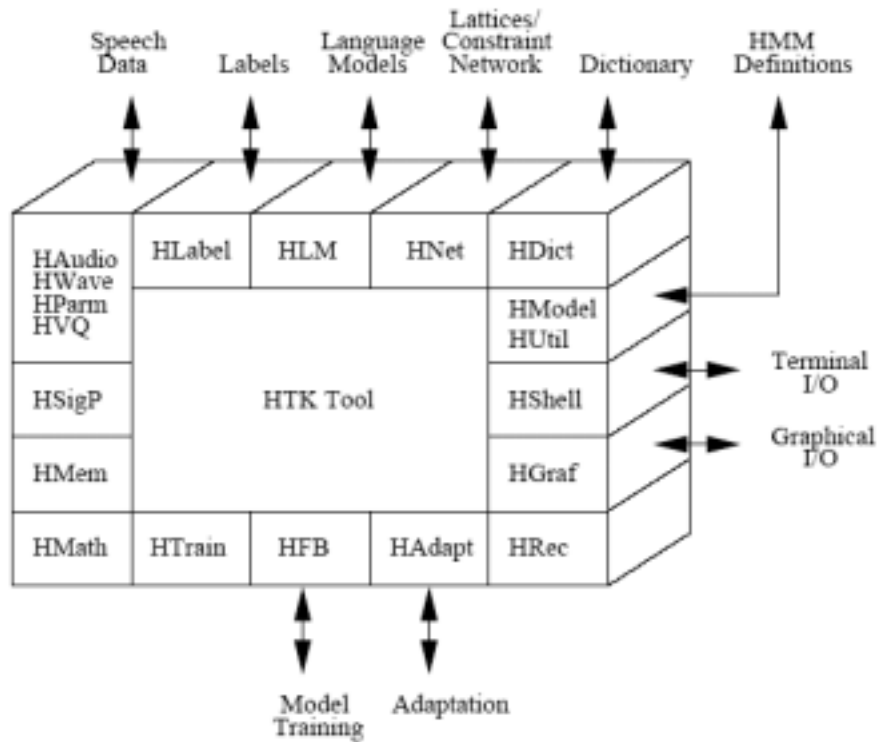
圖二十六. System Development Cycle

#### 4.2. 相關工具程式庫

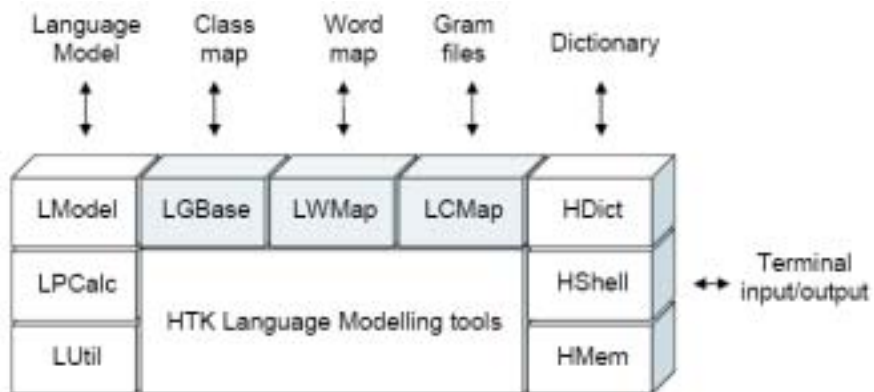
以下我們以交大與北科大的口語對話汽車導航系統建構過程為例，介紹一些好用的公開領域工具程式庫，尤其是這些工具程式庫，不但都有完整原始碼，充分的說明文件，更有教學文件（tutorial），按部就班教導使用者如何入門。

在此例中，口語對話汽車導航系統的架構如圖二十二所示，我們首先需建立語音聲學模型與語言模型，在這方面，我們使用英國劍橋大學釋放出來的 Hidden Markov Model Toolkit (HTK)，其包含最先進 (state-of-the-art) 的語音聲學模型工具庫，如圖二十七，與語言模型工具庫 (HTK Language Modeling Toolkit, HLM)，如圖二十八。其中 HTK 可以作大字彙語音辨認，包括使複雜的用 tri-phone 模型[30]，HLM 可以訓練 tri-gram 模型，也可以作 cluster-based 的語言模型[30]。



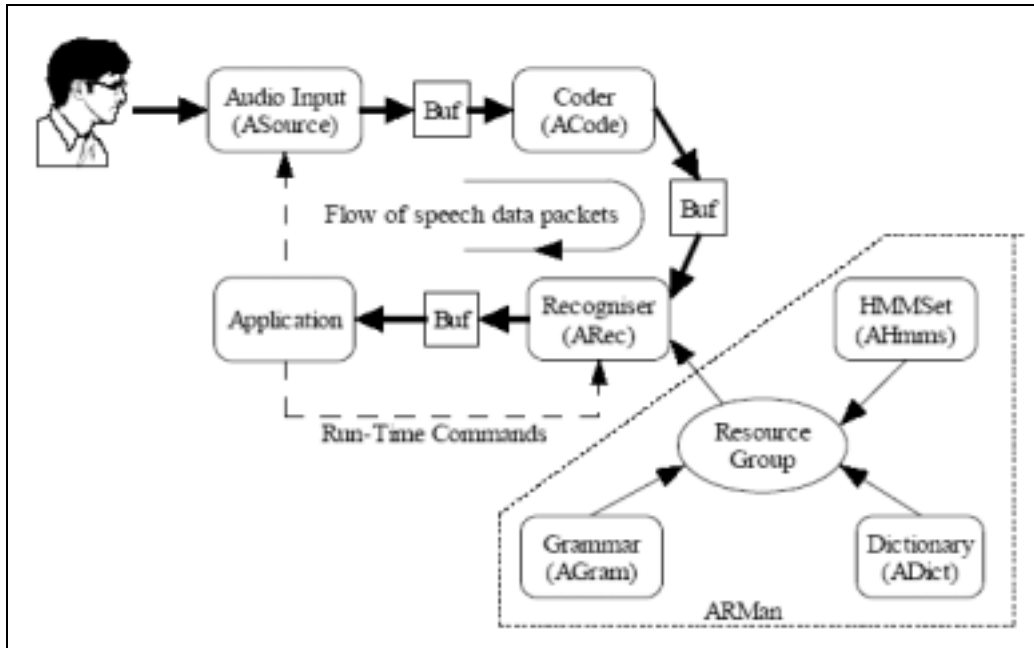


圖二十七. The HTK Operating Environment



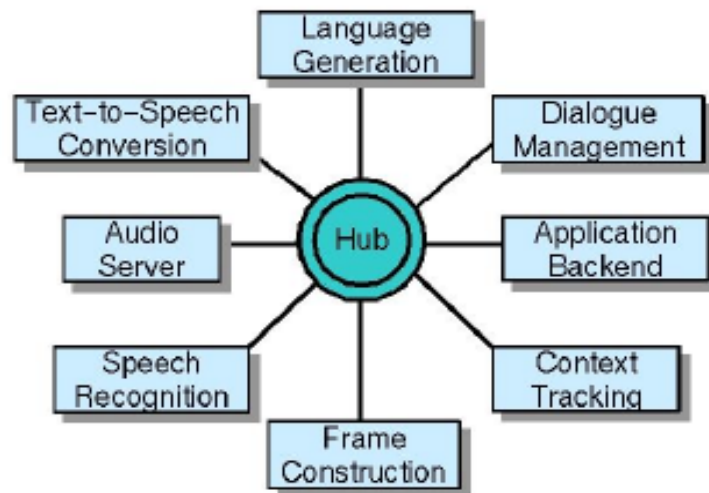
圖二十八. Language Modelling Reference

在作即時語音辨認模組時，我們也是採用同是英國劍橋大學釋放出來的 Application Toolkit for HTK (ATK) [31]，如圖二十九，以便與 HTK 和 HLM 相容。ATK 可以在 windows 作業環境下執行，其支援多工，可作大字彙辨認，也可以定義辨認用的文法網路 (grammar network)，限制辨認空間，提高辨認率。此外其也支援 confidence measure，可提供較佳之對話流程控制能力。



圖二十九. Basic recognition system using ATK

最後，我們使用 MIT 釋放出來的 Galaxy communicator，將各模組的輸出輸入連接起來，Galaxy communicator，為一 Hub-Server 架構，如圖三十所示，基本上各 server 獨立執行，透過網路連接，靠 Hub 傳遞訊息。好處是其為一分散式系統，各 server 可在網路上的不同機器執行；一方面各 server 可獨立開發執行，免除開發時不同開發群間的溝通協調，與後續的程式碼維護負荷；一方面若一台機器計算能力不足，亦可以用多台機器來分散計算能力需求。



圖三十. Galaxy Communicator software infrastructure

最後完成的口語對話汽車導航系統如圖十二所示，我們將整個系統放到筆記型電腦上（Pentium4, 1.6GHz），以藍牙介面連接麥克風，GPS 與行動電話，拿到車上實際開車上路測試。除測試一般 GPS 汽車導航功能外，如“帶我走最佳路徑去台積電”，也測試較人性化的對話功能，見圖三十一，如詢問“再來怎麼走”；確認“下個路口左轉還是右轉”；撥打行動電話“我要打電話給王逸如”等等。



圖三十一. 口語對話汽車導航系統實機測試情境

## 5. 結論與展望

口語對話系統，是未來人機介面必備的技術，最終目的是希望可以直接以口語指揮電腦，幫我們完成一些複雜的工作。但因其技術複雜，人類的口語行為也非常活潑多變，因此需要各領域的研究人員，共同溝通協調。希望此文章可以激發出一些志同道合的人，一起跨越領域的藩籬，互相合作，完成最終目標。

## 6. 參考資料

- [1] [Harald Aust](#), [Martin Oerder](#), [Frank Seide](#) and Volker Steinbiss, "The Philips automatic train timetable information system", *Speech Communication*, Special issue on interactive voice technology for telecommunication applications, Volume 17, Issue 3-4 (November 1995), Pages: 249 - 262
- [2] Spoken Language Groups, MIT, <http://www.sls.csail.mit.edu/sls/sls-orange-nospec.html>
- [3] Hong-Kwang Jeff Kuo and Chin-Hui Lee, "Discriminative Training of Natural Language Call Routers", *IEEE Transactions On Speech And Audio Processing*, Vol. 11, No. 1, January 2003
- [4] Harald Aust, Olaf Schröer, "An Overview of The Philips Dialog System", *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne Conference Resort Lansdowne, Virginia, February 8-11, 1998  
<http://www.nist.gov/speech/publications/darpa98/pdf/ww20.pdf>
- [5] DARPA Communicator Evaluation Committee,  
<http://www.dcs.shef.ac.uk/~walker/eval/eval.html>
- [6] CMU Communicator, <http://www.speech.cs.cmu.edu/Communicator/>
- [7] CU Communicator, <http://communicator.colorado.edu/>
- [8] SRI Communicator System, <http://www.ai.sri.com/~communic/>
- [9] Galaxy Communicator, <http://communicator.sourceforge.net/>

- [10] Verbmobil, long-term project of the Federal Ministry of Education, Science, Research and Technology (BMBF, Projekträger DLR), Germany, <http://verbmobil.dfki.de/>
- [11] SmartKom, "Dialog-based Human-Technology Interaction by Coordinated Analysis and Generation of Multiple Modalities", <http://smartkom.dfki.de/>
- [12] Human-Robot Dialog System (ASKA), 音情報 理学講座, 鹿野研究室(Shijano Lab.), <http://isw3.aist-nara.ac.jp/IS/Shikano-lab/>
- [13] 台大語音實驗室, 台灣大學電機工程學系資訊工程學系, <http://speech.ee.ntu.edu.tw/>
- [14] 陳伯琳, 台灣師範大學資工系, <http://140.122.185.120/>
- [15] eWeather 中文電話語音對話系統, 工研院電通所前瞻技術中心, [http://www.ccl.itri.org.tw/products/transfer/multi/11t\\_050.htm](http://www.ccl.itri.org.tw/products/transfer/multi/11t_050.htm)
- [16] Speech and Communication Signal Processing Lab., National Taipei University of Technology, <http://www.cc.ntut.edu.tw/~enlab07/>
- [17] Speech Processing Lab., National Chiao Tung University, <http://speech.cm.nctu.edu.tw/>
- [18] 多媒體人機通訊實驗室, 成功大學資工系, <http://chinese.csie.ncku.edu.tw/chwu/sitemap.htm>
- [19] 汽車噪音環境內語音信號處理技術及多人對話系統之研究, 王駿發, <http://140.116.156.179/>
- [20] 微軟亞洲研究院, Microsoft Research Asia, <http://research.microsoft.com/asia/>
- [21] I B M 中国研究中心, IBM China Research Laboratory (CRL), <http://www.research.ibm.com/beijing/>
- [22] 中科院自動化所模式識別實驗室, <http://nlpr-web.ia.ac.cn/mainpage.html>
- [23] 清華大學語音技術中心, <http://cst.cs.tsinghua.edu.cn/>
- [24] "得意"口 对 系 包与"得意"智能互动信息服 , 得意音通公司, <http://www.d-ear.com/>
- [25] 安徽中科大訊飛資訊科技有限公司, <http://www.iflytek.com/>
- [26] VoiceXML Forum, <http://www.voicexml.org/>
- [27] Open VXI VoiceXML Interpreter , <http://sourceforge.net/projects/openvxi/>
- [28] Hideyuki Mizuno, Hisako Asano, Mitsuaki Isogai, Miki Hasebe and Masanobu Abe, "Text-to-Speech Synthesis Technology Using Corpus-Based Approach", *NTT Technical Review Online, Speech and Audio Signal Processing Technologies*, March 2004 Vol.2 No.3. <http://www.ntt.co.jp/tr/0403/files/ntr200403070.pdf>
- [29] J. Glass, J. Polifroni, S. Seneff and V. Zue, "Data Collection and Performance Evaluation of Spoken Dialogue Systems: The MIT Experience," *Proc. 6th International Conference on Spoken Language Processing*, Beijing, China October 2000. <http://www.sls.csail.mit.edu/sls/publications/2000/685.pdf>
- [30] Hidden Markov Model Toolkit (HTK), <http://htk.eng.cam.ac.uk/>
- [31] Application Toolkit for HTK ( ATK ), <http://mi.eng.cam.ac.uk/~sjy/software.htm>