行政院國家科學委員會專題研究計畫　成果報告

子計畫三：行動語音人機介面的研究與開發(I)

計畫類別： 整合型計畫
計畫編號： NSC93-2218-E-009-041-
執行期間： 93 年 08 月 01 日至 94 年 07 月 31 日
執行單位： 國立交通大學電信工程學系(所)

計畫主持人： 張文輝

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： 本計畫可公開查詢

中　華　民　國 94 年 10 月 31 日

行政院國家科學委員會專題研究計畫報告

行動語音人機介面的研究與開發

ITS information access using voice over MANET

計畫編號：NSC 93-2218-E-009-041

執行期限：93 年 8 月 1 日至 94 年 7 月 31 日

主持人：張文輝　交通大學電信工程系 教授

一、中文摘要

(關鍵詞：語音對話系統，分散式語音辨認。)

人性化的隨身資訊服務是智慧型運輸系統必備的功能，網際網路的興起更成為資訊傳播的重要平台，使用語音作為人機介面則可以提升行車安全與便利。本子計劃在 MANET 無線網路架構下，建構一行動語音對話系統，讓駕駛員以聲控操作取得道路指引及購物消費的生活資訊。本年度研究規劃主要著重於聲控操作人機介面的製作。語音辨認系統採用分散式架構，車內終端機負責語音特徵參數的擷取與向量量化壓縮處理，遠端伺服器則利用隱藏式馬可夫模型的訓練與比對執行語音辨認處理。目前執行進度已完成分散式語音辨認系統，並於快速乙太網路和校園無線區域網路的環境下分別進行語音辨認及聲音回傳的操作。進一步模擬具有不同叢發特性的通道錯誤，用以測試環境對於分散式語音辨認的影響。

英文摘要

(Keywords: spoken dialogue system, distributed speech recognition.)

The purpose of this three-year research is to develop a spoken dialogue system that allows drivers to use voice-controlled commands to access the ITS information server through a mobile ad-hoc network (MANET). The first part of this project will focus on developing a distributed speech recognition system, in which speech features extracted from a local front-end are transmitted through a data channel to a remote back-end recognition server. In light of the low-bit-rate transmission, speech features are compressed using a split vector quantizer that produces the index of the nearest code-vector over digital wireless channel. Wireless chammels are characterized by error bursts due to the combined effects of inter-symbol interference and multipath fading. It is believed that further improvement of system performance can be realized through a precise characterization of the channel. Then at the next part of the project we wish to design a joint source -channel decoder that can work well in high bit error rate condition.

二、計劃緣由與目的

智慧型運輸系統的發展趨勢，將是結合無線通訊與網際網路，突破時空的限制，以提供車輛駕駛員更人性化的隨身資訊服務。為提供多樣化的應用服務，人機介面必須進行適度的互動溝通，讓電腦逐步瞭解、接受及回應使用者的查詢指示。問題是車輛駕駛員在行進間的讀寫能力受限，無法在電腦鍵盤輸入或讀取行車相關資訊，使用語音作為人機介面則可以大幅提升行車安全與便利[1]。因此，我們計劃在 MANET 架構下，製作一個行動語音對話系統，結合語音辨認技術使系統得以聲控操作，進而提供駕駛員查詢道路指引和消費購物的生活資訊。

考量無線通訊裝置能源與運算處理能力先天受限的問題，將利用分散式語音辨認架構實現人機對話應用的服務，其處理單元包含兩部分：前級採用歐洲電信標準局(European Telecommunication Standards Institute, ETSI)所制訂的分散式語音辨認架構[2]，針對每一音框抽取其特徵參數再執行向量量化的壓縮編碼處理，透過車內終端機負責發送到遠端的伺服器進行較複雜的後級辨認處理[3]。辨認的核心技術則以隱藏式馬可夫模型為架構，藉由大量的語音資料庫訓練辨認模型，伺服器透過模型比對產生辨認結果便回覆相關訊息給用戶端。

為銜接下一年度的研究，在已經完成

對話系統後，我們著手進行分散式語音辨識面臨不同的無線通道環境下其辨識能力的影響。利用 Gilbert 通道模型模擬具無線通訊位元錯誤特性的通訊環境，提供未來開發合併音源-通道解碼器設計所需的實驗環境[4,5,6]。

三、研究方法與結果

本研究主要是發展一適用於無線網路環境的語音對話系統。系統流程如圖 1 所示，包含分散是語音辨認與檢索語音回傳兩階段，即用戶端進行語音參數分析送出至遠端伺服器進行辨認與辨認結果檢索後回傳相對應語音資訊予用戶端播放。以下針對各個核心單元分別描述：

(1) 語音參數分析與資料壓縮：

依據 ETSI 在 2003 年所制訂的分散式語音辨認標準，系統的前級處理主要針對每一音框抽取其特徵參數再作壓縮處理，其設計之關鍵目的在於有效對抗行動通訊環境中面臨低傳輸位元率、高傳輸錯誤率以及背景雜訊干擾等相關影響，以期大幅提昇辨認的正確率。因此，核心技術包括雜訊抑制與壓縮編碼的設計[7]。訊號分析過程以音框長度 25ms 和平移 10ms 為單位依序處理，同時依據能量來判定每一音框的屬性是有語音或是僅有背景雜訊。雜訊抑制的工作原理，是藉由分析一段僅具單純背景雜訊的訊號，配合 Wiener 濾波器演算法設計一最佳化雜訊抑制濾波器，可有效減低背景雜訊對辨認結果的影響。因應無線網路低位元率傳輸的需求，編碼壓縮處理有其必要性。壓縮的處理以音框為單位，將 14 個參數依序兩兩一組個別進行向量量化，再送出最近似量化碼字的索引值。至於通道編碼的部分，則是依序將每兩個音框的資料視為一單位，產生其相對應的 CRC 錯誤偵測碼，再將 12 組音框對(frame pair)的相關資料依固定格式封裝，附上同步序號與檔頭訊息即完成傳輸所規定的封包資料。

(2) 語音辨認模型的訓練與測試：

當伺服器收到要辨認的用戶端封包資料後，會進行錯誤和緩(error mitigation)與特徵參數處理兩項工作。錯誤和緩的目的是要檢測及補償在網路傳輸過程中所引起的資料錯誤。檢測的機制有兩種，一是透過封包中所屬的 CRC 錯誤偵測碼逐一驗證每個音框對資料的正確性。另一個方法則利用相鄰音框間具有的相似性來進行檢驗，也就是事先訂下鄰近音框間個別特徵參數的相似範圍，當接收到的參數超出其門檻值時，則判定該組特徵參數已發生傳輸錯誤。補償機制的啟動是當檢測顯示有連續 2B 個音框發生錯誤，則前 B 個音框的參數用錯誤發生前一個正確音框的參數來取代，而後 B 個音框的參數則以錯誤發生後第一個正確音框的參數來取代，藉此可適度地緩和傳輸錯誤所引起的辨認失誤。至於伺服器端的特徵參數處理，旨在取得更多有助於語音辨識的特徵參數。在每一音框先線性整合能量與第零個梅爾倒頻譜係數，並與其他 12 個梅爾倒頻譜係數構成 13 個特徵參數，接著個別針對每個參數以差分方式求出其速度與加速度，整合而成為一組標準所需的 39 個辨認用語音參數。目前已完成系統的開發與測試，是利用 HTK 軟體製作中文單音節的辨認處理器，而辨認模型的參數訓練則藉由 Baum-Welch 疊代演算法求得。基於中文語言特性之考量，將使用一單純「左至右」型態的隱藏式馬可夫模型，聲母和韻母兩部分分別以 3 個和 40 個狀態模擬其統計模型，且每個狀態內均以 64 個高斯分佈混合模型來近似觀察值的機率分佈。進一步，更引入靜音之間歇與停頓的模型，其作用分別模擬發音過程前後端與過程中字和字之間過渡時期內訊號的統計模型。在中文單音節辨認模型訓練過程是利用 150 位男性和 150 位女性語者所錄製的 TCC-300 語料庫，以其中 270 位作為訓練語料進行 Baum-Welch 疊代演算法，額外 30 位的語料進行測試。

(3) 聲控操作遠距有聲網站的製作：

本系統是利用 Winsock API 作為開發平台，相關工作項目包括用戶端連線之初始過程、資料傳輸規格之定義、以及程式執行時程序控制和語音資料庫管理。使用者啟動用戶端程式後，鍵入伺服器端之網路位址以開啟雙方之間的連線與資料傳輸。為了提供系統運作之即時性訊號處

理，分別在程式用戶端的參數抽取、量化編碼處理以及伺服器端辨認處理等相關位置，提供作為同步處理的排隊緩衝的暫存空間。而系統設計則利用圖形視窗介面控制不同的執行緒以完成語音輸入、參數抽取、進行辨認以及回傳結果等相關步驟，如圖 2 所示。行動終端機的執行緒分述如下：WaveIn 執行聲音載入時的取樣量化以及音框訊號的管理，AdvfrontEnd 執行雜訊抑制和參數抽取，Coder_VAD 執行參數量化編碼以及資料封裝的工作。遠端伺服器的執行緒則包括有：Socket 過濾並解析屬於本系統的特徵參數封包，Mainserver 則包括傳輸錯誤和緩、接收資料的解碼、辨認參數的後處理、辨認比對的運算以及辨認結果相關資料的回傳。最後，Client_recv 執行緒會接收回傳的語音資料並依序播放予使用者。

四、實驗結果與討論

本年度計畫主要目標是聲控系統的建置，首先在高速乙太網路的架構下，先完成系統的開發，其目的是考慮有線網路傳輸具有較高的傳輸品質，可提供幾乎沒有傳輸錯誤的平台，讓我們在開發過程中不受傳輸錯誤的干擾逐一完成每一項核心技術的測試與整合，測試結果顯示在通道錯誤幾乎為零的情況下連續數字音與中文 411 個單音節的辨認結果分別可以達到 97.2%與 61%之正確率。在確認過系統開發完整無誤之後，進一步將用戶端介面移植到無線網路設備上，經反覆測試也確認系統重建後的功能正常，但由於無線網路通訊品質不穩定，導致辨認結果會有程度不一的下降。

為瞭解傳輸錯誤對於系統辨認率的影響以及要銜接下一年度的研究，我們嘗試模擬分散式語音辨任在不同的無線通道環境下之運作狀況。利用 Gilbert 通道模型模擬具有無線通訊位元錯誤特性的通訊環境。模型架構是基於兩個狀態的馬可夫鏈，如圖 3 所示，其中一個狀態是不會發生錯誤的良好狀態，另一個則是具有錯誤機率為 h 的不良狀態，而每一次兩狀態相互轉變的機率為 g 和 b，利用參數{h,g,b}的調整可產生不同叢發性質的位元錯誤序列。藉此設計具有不同錯誤率的通道環境，再讓本系統遭遇這些環境，並觀察其在辨認結果上的影響。結果如表 1 顯示由於系統設計本身具有錯誤和緩的機制可以針對發生錯誤的音框資訊加以修正，所以位元錯誤率在 1%內的情況下，幾乎不影響語音辨認的效果，當錯誤率超過 5%後，辨認率便出現明顯的下降。透過通道模擬與辨認測試的實驗，初步提供我們對於無線傳輸通道錯誤在系統效能的影響程度。接著下一年度的計畫將著手於實際無線傳輸環境的量測，藉由量測的結果進行通道模型的估算，進一步開發合併音源-通道解碼器設計所需的實驗環境。

五、結論

本計畫第一年著重在語音對話介面系統的設計與開發，核心技術包括語音參數的分析、資料壓縮、通道錯誤和緩和語音辨認比對，前三者的設計完全符合 ETSI 標準所制訂的規格，得到傳輸量為 4800bps 的封包資訊，其內容包含連續 24 個音框所相對應的特徵參數以提供作為語音辨認之用。語音辨認的技術基於隱藏式馬可夫模型建構適合中文連續數字音與 411 個單字音的比對模型，透過 TCC 中文語料訓練辨認器的模型參數，經由測試數字串與 411 單字音的辨認率分別可達 97.2%與 61%。此外，我們利用通道模擬產生具有無線通訊特性的傳輸位元錯誤，並將其置入傳輸封包內用以測試不同的錯誤率對於語音辨認系統的影響，結果顯示在低錯誤率時，系統本身的錯誤和緩機制可以提供適當的抑制作用，而當錯誤率再增加時，便無法保證辨認的正確性。依據本年度研究之結果，利用分散式語音辨認平台將持續進行無線通道分析以及設計具有抵抗通道錯誤的和緩機制。

六、具體成果

本研究結合分散式語音辨認平台開發以及無線通訊傳輸通道模擬，進行辨認結果的測試以及錯誤和緩機制的設計。研究成果參與國際語音通訊相關研討會發表結果獲得肯定。

- C. L. Lee, and W. W. Chang, “Mem-

ory-enhanced MMSE-based channel error mitigation for distributed speech recognition," *Interspeech '2005-Eurospeech*, Lisbon, Portugal, Sep, 2005.

七、參考文獻

[1] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: a guide to theory, algorithm, and system development*, Prentice Hall, 2001.

[2] "ETSI ES 202 050 v1.1.3 Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech recognition; Advanced front-end feature extraction algorithm; Compression algorithm," ETSI Standard, 2003.

[3] C. Pelaez-Moreno, A. Gallardo-Antolin, and F. Diaz-de-Maria, "Recognizing voice over IP: a robust front-end for speech recognition on the world wide web," *IEEE Trans. on Multimedia*, vol. 3, pp. 209-218, June 2001.

[4] T. Fingscheidt and P. Vary, "Softbit Speech Decoding: A New Approach to Error Concealment, "*Speech and Audio Processing, IEEE Transactions* on , Volume: 9 , Issue: 3 , March 2001, pp. 240 – 251.

[5] W. Turin, *Digital Transmission Systems: performance analysis and modeling*, McGraw-Hill, 1999.

[6] W. W. Chang, T. H. Tan, and D. Y. Wang **"**Robust vector quantization for wireless channels," *IEEE Journal on Selected Areas in Communications*, Vol. 19, pp. 1365-1373, July 2001.

[7] P. Hedelin, P. Knagenhjelm, and M. Skoglund, "Vector quantization for speech transmission," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995.

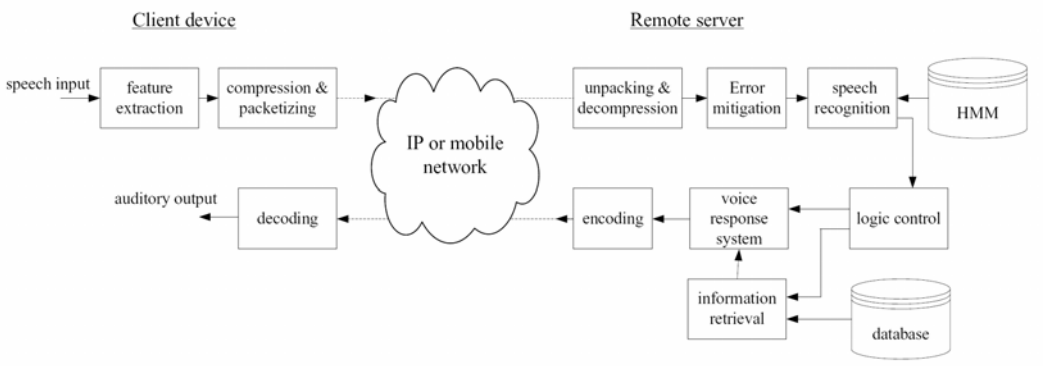

圖 1：對話系統方塊圖。

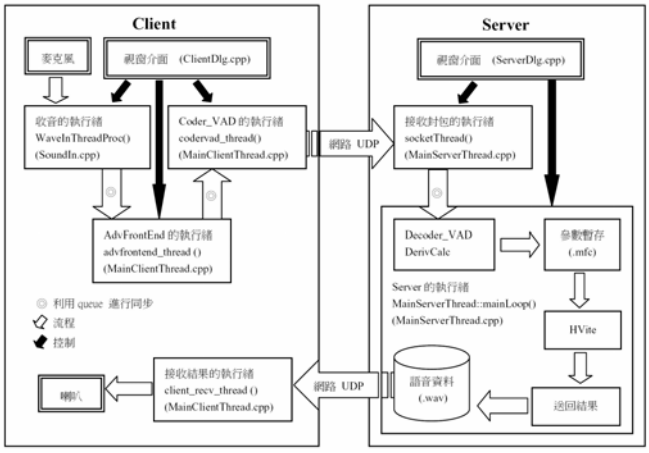

圖 2：介面開發執行緒關係圖。

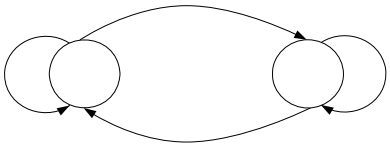

圖 3：Gilbert 通道模型

表 1：不同通道錯誤率對語音辨認結果的影響

位元錯誤率(%)	0.10	0.31	1.0
數字串辨認率(%)	97.0	97.0	96.9

3.16	10.0	17.78	31.62
94.9	92.8	82.4	48.3