# 行政院國家科學委員會專題研究計畫　期中進度報告

## 子計畫三：語音人機介面於居家照護系統之應用(1/3)

計畫類別：整合型計畫
計畫編號：NSC93-2218-E-009-062-
執行期間：93 年 10 月 01 日至 94 年 09 月 30 日
執行單位：國立交通大學電信工程學系(所)

計畫主持人：王逸如
共同主持人：廖元甫
計畫參與人員：楊智和、許竣為、彭以榮

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 7 月 29 日

# 行政院國家科學委員會專題研究計畫成果報告

## 數位化居家照護系統研究
## 子計畫三:語音人機介面於居家照護系統之應用(1/3)

計畫編號：NSC93－2218－E－009－062
執行期限：93 年 10 月 1 日至 94 年 9 月 30 日
主持人：王逸如　國立交通大學電信工程系
共同主持人：廖元甫　國立台北科技大學電機系
計畫參與人員：楊智和、許竣為、彭以榮

## 一、中文摘要

老人居家生活及照護中的重要課題，是如何讓老年人能方便地使用各種居家照顧的設備與服務,因此在本計畫中將研究適合老年人使用之語音人機介面,其整合(1)語音辨認、(2)口語對話系統、(3)語者辨認系統與(4)文字轉語音系統,讓老年人可以直接以最自然的口語跟機器對話取得服務,此外以上四個系統亦可提供其他子計畫之人機介面使用。

此外，在本計畫中將分三年逐步收集之老年人的語音資料，與口語對話行為資料，以利相關之研究與系統發展。

本年度計畫中手現建立一套國語關鍵詞系統,並與子計畫三、之支麥克風陣列及居家看護機器人進行初步整合，並製作 Corpus-based 國語語音合成器之核心。

**關鍵詞**：語音辨認、關鍵詞辨認、語者調適、不匹配補償、口語對話系統、語者識別，語者確認、文字轉語音系統、老年人語音資料庫

## Abstract

One of the important issues for elderly homecare is to provide elderly people a universal and convenient human-machine interface to access various services and equipments. Among all possible interfaces, the friendliest way is through the speech technology.

This goal of the sub-project is to build a spoken dialogue interface. It will focus on integrating four core speech techniques, including (1) speech recognition, (2) spoken dialogue, (3) speaker identification/verification and (4) text-to-speech system. Beside, the core techniques can also support other sub-projects. Finally, an elderly speech database including speech signal and spoken dialogue behavior will be collected to assist the research and system development.

In the first year, a Mandarin keyword spotting system was first implemented and integrated with the microphone array and the homecare robot in subproject 4 and 5. Besides, the kernel of Mandarin corpus-based TTS system was also been developed.

Keywords: speech recognition, keyword spotting, speaker adaptation, mismatch compensation, spoken dialogue system, speaker identification, speaker verification, text-to-speech system, elderly people speech database.

## 二、 緣由與目的

老人居家照護已是各國政府相當重視的問題，不論在醫界或工程界，已有許多研究團隊投入 Homecare 方面的技術開發，期待能運用現代化科技的設備與服務，使 Homecare 的品質更加提升。

而老年人的視力與反應通常會隨年紀漸長而變弱,也就會越不容易使用目前資訊設備的標準人機介面（Human Computer Interface, HCI），例如螢幕、滑鼠與鍵盤等，而且多樣的人機介面，也徒增使用上的困難，因此要讓老年人可以更自然方便地使用各種居家照顧的設施,就必須具備自然語言／語音輸出入介面,讓老年人可以以口語直接使用今日科技世界『機器』所能提供的種種便利服務。

## 三、研究方法

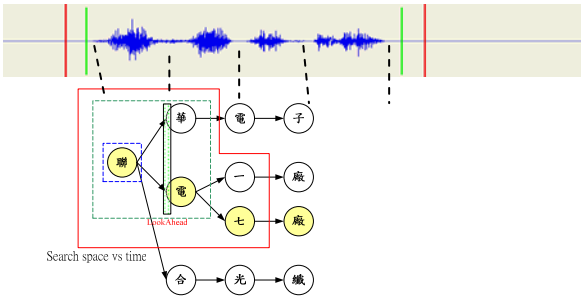
### 1. 國語關鍵詞辨認系統之製作

在計畫中，第一年為配合其他子計畫，所以在本計畫中先製作一個國語關鍵詞辨認

(keyword spotting)系統；在此將此系統建立的步驟簡述如下：

(1) 首先使用 TCC-300 國語語料庫[1]建立國語 100 final-dependent initial 及 40 final HMM 語音辨認模型。

(2) 製作一連續語音辨認器，為了簡化計算量以達及時辨認的目的，在系統中使用了光束搜尋法(beam search)[2]以降低搜尋空間增加速度，並加入 look-ahead 方法[3]以提升在刪除部分搜尋空間後之辨認效能，如圖一所示。




圖一、語音辨認系統中之光束搜尋(beam search)及 look-ahead 方法示意圖。

(3) 在關鍵詞辨認系統中，必須訂定待辨認之詞彙。在所製作之系統中使用一個簡單的文字檔來定義系統中所能辨認的關鍵詞，如下所示：

```
居家看護 4    /* task name, no. of keywords */
電腦  電腦    /* keyword, output answer */
開燈  開燈
關燈  關燈
打電話 打電話
```

在關鍵詞定義檔中所有關鍵詞均以 BIG-5 碼格式輸入，辨認器中會將其轉換成 pinyin 格式，用以查詢所需之 HMM 聲學模型。如此在其他子計畫中將可以輕易的依據應用建立其所需的待辨認關鍵詞詞典。

(4) 關鍵詞辨認系統中之填充模型之建立

關鍵詞辨認系統中之填充語音模型(Filler model)是用來描述使用者輸入語句中非關鍵詞部分的語音信號。若填充語音模型過於粗超會使得語句中關鍵詞部分被辨認成非關鍵詞，也就是所謂 miss detection；若填充語音模型過於精細會使得語句中非關鍵詞部分被辨認成關鍵詞，也就是所謂 false alarm。在本系統中所使用的是較為精細的國語 100 final-dependent initial 及 40 final HMM 次音節模型，再對辨認分數加上一個懲罰函式；在本系統中因為在一輸入語句中僅允許出現一個關鍵詞所以將不會造成 flase alarm。

(5) 所建立國語關鍵詞辨認程式，其形式為 Microsoft Windows 作業系統下的動態函式(DLL)，我們也撰寫了一個簡單的使用者介面(UI)來展示我們所發展的國語關鍵詞辨認系統；如圖二所示，此系統會輸出 Top-10 的辨認結果。

(6) 為了測試上述國語關鍵詞辨認系統之效能，我們利用現有的語料庫來測試，此語料庫為一個較為複雜的工作(task)－查詢新竹工業園區的廠商名稱(有 300 多個關鍵詞)，對 2000 句測試語料之辨認率約為 98%正確率，且系統可以輸出 Top-N 辨認結果。除語音端點偵測錯誤的情況外，在所有測試中正確答案均能進入前 5 名，這在日後加入對話系統時有很大的幫助。

(7) 由上述系統效能評估，對老人居家生活照護中之簡單人機介面之應用已可以上線，所以已開始與子計畫五之麥克風陣列與子計畫四之居家看護機器人進行整合中。經初步測試，經麥克風陣列輸入後對(5)中之測試語料之辨認率會下降至 86%。所以我們也開始為使用麥克風陣列輸入之語音辨認器做之通道效應補償的工作。

## 2. 麥克風陣列之通道效應補償

本子計畫中所製作之國語關鍵詞辨認系統在與子計畫五之麥克風陣列進行整合時，因製作語音辨認器時所使用之訓練語料是由麥克風輸入與使用麥克風陣列輸入時有不同的通道效應。所以在使用麥克風陣列輸入時需要做語音通道效應的補償。在計畫中所使用的是 MLLR(Maximum likelihood Linear Regression)方法[2]來調適語音 HMM 辨認模型來做通道之補償。

MLLR 語音辨認模型調適基本上是將 HMM 語音辨認模型中各個 mixture 的 mean 經一線性轉換來做調適，也就是

$$\hat{\mu} = A\mu + B$$

，其中轉換矩陣 A, B 可由 EM(Expectation-Maximum)演算法求得。但是，我們需要大量的語料來求得較佳的 HMM 之調適辨認模型；語料量越大，我們可以使用較多組的線性轉換來調適。所以在計畫中，我們已將 TCC-300 中十分之一的語料(約 3 小時)，經子計畫五所製作之麥克風陣列再次錄製以作為作 HMM 辨認模型之調適所需之調適語料，並已開始進行訓練 MLLR 調適模型，以提升在麥克風陣列輸

入環境下的語音辨認系統效能。

## 3. Corpus-based 國語語音合成器之製作[4]

子計畫四之居家看護機器人除需要語音輸入外，還需要語音合成器能讓機器人『說話』。在現階段，將使用交大語音實驗室所發展的國語語音合成器;因為它是使用國語基本411音節來合成所有國語語音，所以所需之記憶體會較小，適合在機器人上。

現今的語音合成器多使用例如：PSOLA(Pitch Synchronous OverLap adding)等語音剪接方式來產生所需之合成語音,在語音合成器中任何對波型有做修飾都會造成合成音質下降的情況之下,使用大型語料庫所儲存的語音來做合成單元,如果可以找到較長之詞或詞組來當合成單元,當然是一個比較良好的選擇,因為在這樣的合成單元內,就已經包含本身的音韻,因此對訊號的修飾就可以盡量避免,在串接時,對於合成語音的自然度當然有一定的效果提升。但所需付出的代價是必須儲存數十 Gbyte 的語音波形檔。現今我們所錄製的大型語料庫其文字內容為中央研究院現代漢語語料庫(Sinica Corpus)[1]中的文章，至目前為止計有 8,017 個詞，語音波型檔僅數十 Mbyte。

(1) 合成器語音資料庫之文字及波形前處理

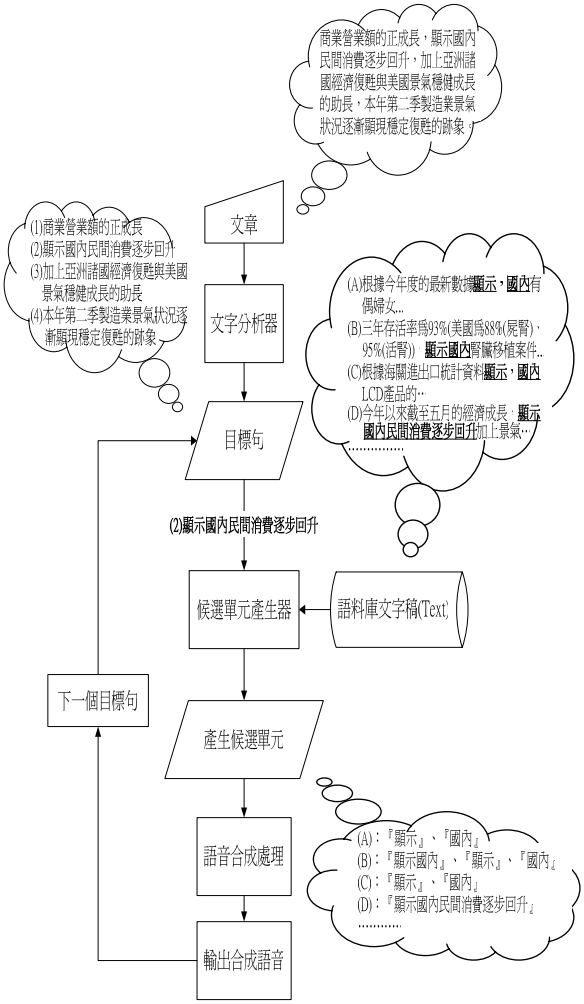
因為所使用之語料使用中央研究院現代漢語語料庫之文字資料所錄製,所以使具有斷詞結果乃至於語法結構之資料,所以不需對文字資料再做處理。而波形資料則是使用 HMM 辨認模型做 force-alignment 後再經微調獲得－例如：將因節切割位置移至能量較小處等。

(2) corpus-based 語音合成器核心之製作

在 Corpus-based 語音合成器中,合成單元的選取當然是越大越好,但主要是以詞組及詞為單位,所以必須以輸入文去語料庫中搜尋,對於每一個可能出現的詞或音節,去搜尋所有可能的組合方式,找出一組最佳的詞序列,如圖三所示。

但在一龐大的文字資料庫中搜尋能組成輸入文句所有可能的組合方式是一個極為複雜的運算。為了有效的搜尋，我們設計一套有效率的比對法,基本上是依據句子及前後中文字位置標記與詞段位置標記是否具備有連續性和相關性的特性，故稱之為連續相關比對法。基本上是使用 CLT (Character location table)來記錄語料庫中所有文字資料以便快速搜尋；接著，使用一個 working table (WT)去處理比對過程中所有可能是最長詞串的候選單元,如此就不需對不同長度的候選詞組重複做搜尋，如圖四所示。

對於在語料庫出現次數頻繁的中文字,例如：『的』，上述連續相關比對方法會導致比對時間的拉長,所以我們必須將這些字找出並做一些特殊的比對處理。我們使用了另一個表 FCLT (Frequent Character Location Table) 來紀錄這些常用字在句子中出現位置資訊的標記檔，用以代替要紀錄在 CLT 中的相同資訊；以便搜尋時可以跳過這些常見字，如：『的』，在比對時之處理流程則如圖五所示。如此將可減少搜尋所有可能的候選詞組所組成的合成單元 lattice 之流程所需之時間。

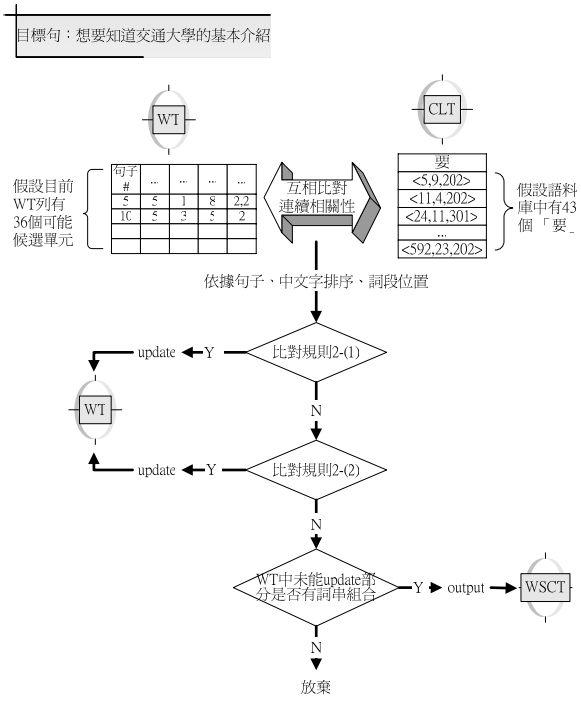



圖三、Corpus-based 語音合成器之示意圖。

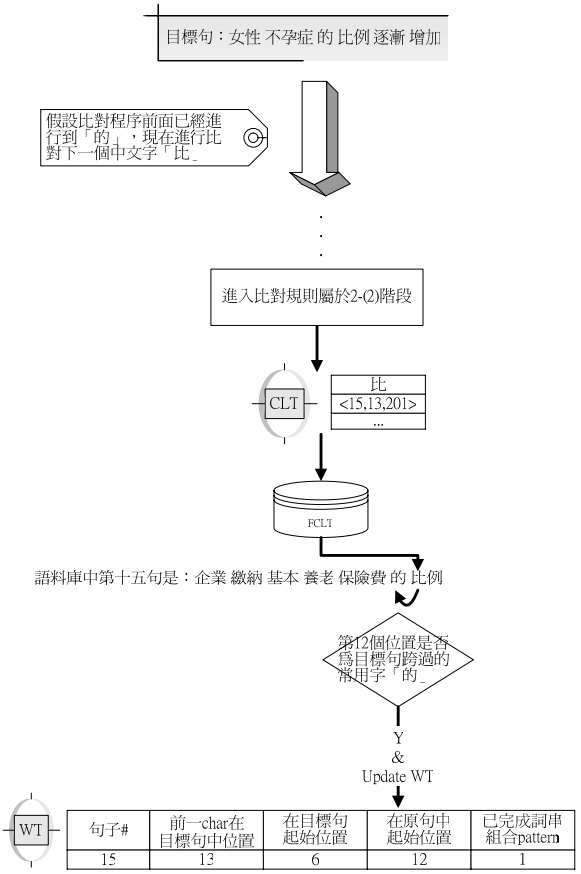
在獲得所有可能的候選詞組所組成的合成單元 lattice 後，我們對將各候選詞組串的聲學參數與由 prosodic information generator 產生之合成語句之聲學參數做比對，決定一組作佳合成詞組串，其聲學及語言參數與所需者合成語句經斷詞及國語語音合成器之 RNN 韻律產生器[5]所產生之聲學及語言參數最接近，所採用的比對條件是

$$C_{intra}(L_k^i) = \frac{w_l D_l(N_k^i)}{Q_l} + \frac{w_s D_s(S_k^i)}{Q_s} + D_p(P_k^i, P_t)$$

其中考慮了各詞串之字數 $N_k^i$，語意參數 $S_k^i$ 及韻律參數 $P_k^i$。




圖四、連續相關比對法之示意圖。




圖五、遇到國語常見字時之文字比對流程圖。

經實驗證實，所製作之 Corpus-based 語音合成器之語音較使用國語基本 411 音節製作的系統為佳。下年度中，我們將把此 Corpus-base 語音合成器與中文文字分析單元結合製作一個完整的 Corpus-based 國語語音合成系統，並繼續擴充所錄製之大型語料庫以提升合成語音之品質。

## 4. 老年人的語音資料庫之蒐集

到目前為止，我們的語音辨認系統還是使用年輕人的聲音做 HMM 升學模型的訓練及測試，對於老人的語音則需做語料之蒐集。對於一個語料庫蒐集的工作，一向是個繁瑣且耗費人力的工作，現今我們已做了所有準備工作：(1)錄音工具之選擇：我們改寫了工研院電通所尖端科技中心(ATC)為進行蒐集台灣之英文語料庫所製作之錄音介面；(2)錄音語料之選擇：本計畫已獲得清大王小川老師之授權，使用 MAT (Mandarin Across Taiwan)語料庫[1]所製作之語料文字，因為該語料內容為短詞、成語及短句，較適合老人錄製。該項老人語音資料庫錄音工作已經展開，但所獲的語音資料尚須經後處理人工更正發音內容與文字內容不符之錯誤後才可以使用。

## 5. 結論

本計畫中已製作了一套國語語音辨認器並已開始與子計畫五之麥克風陣列與子計畫四之居家看護機器人進行初步整合中；將由實際測試之結果再做調整。並且建立了 Corpus-based 國語語音合成系統之核心，下年度將與國語斷詞器結合已完成完整的 Corpus-based 國語語音合成系統。

# 四、計畫成果自評

在計畫書中所列舉之項目均已執行，並將結果開始與其他子計畫作初步之整合。

# 五、參考文獻


[1]、中華民國計算語言學學會，http://rocling.iis.sinica5 之語音資料庫 edu.tw/ROCLING/corpus98/index_cf.htm.

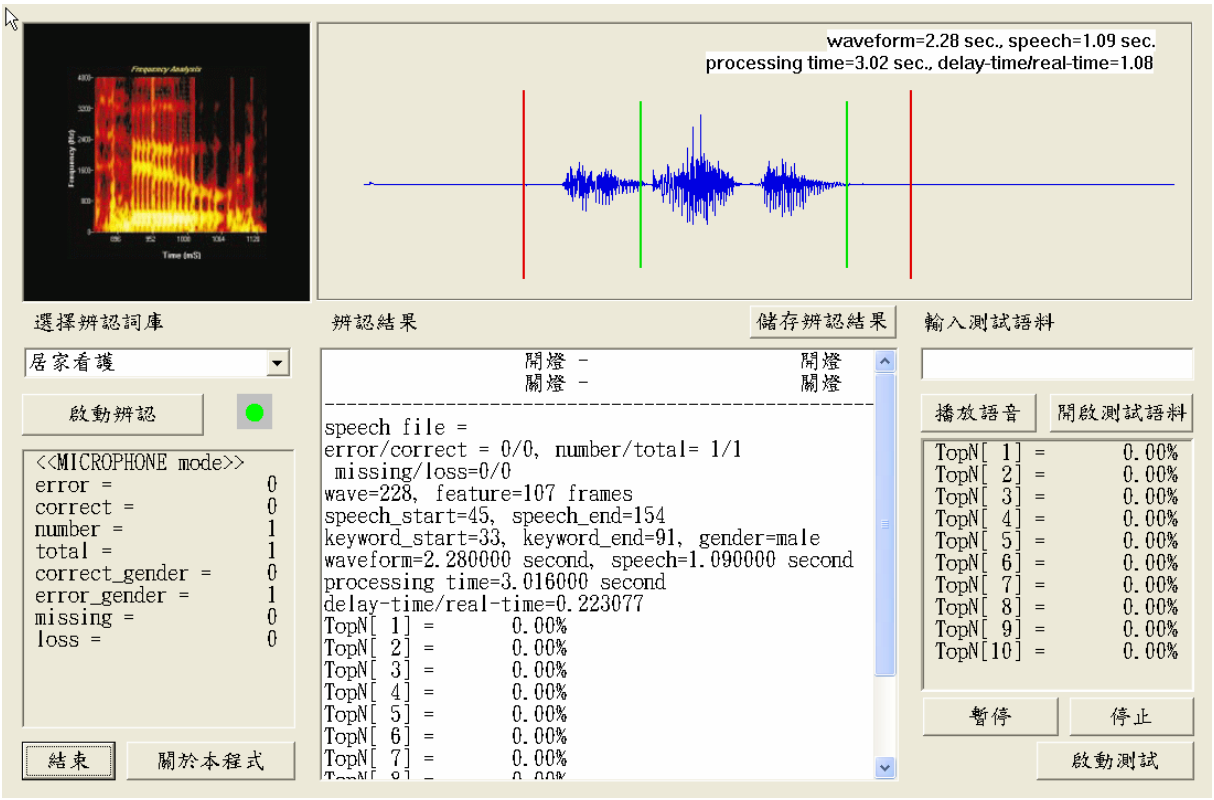
[2]、H. Ney, D. Mergel, A. Noll, and A. Paesler, “ Data driven search organization for continuous speech recognition,”, *IEEE Transactions on Signal Processing*, 40(2):272--281, February 1992.

[3]、Ortmanns, Stefan; Ney, Hermann 2000, “ Look-ahead techniques for fast beam search,” , *Computer Speech and Language* 14, 15.32, 2000.

[4]、C. J. Leggetter and P. C. Woodland, “ Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,”, *Computer Speech and Language*, pp. 171–185, 1995.

[5]、 吳佩穎，『以語料庫為基礎之中文文句翻語音系統中合成單元之選取』，交大電信所碩士論文，民國 94 年。

[6]、 Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, “ An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech, “, IEEE Trans. Speech and Audio Processing, Vol. 6, No. 3, pp. 226-239, May, 1998.




圖二、國語關鍵詞辨認系統。

（下方左、右子視窗為 batch 測試用，中間下方視窗之結果為線上辨認結果）。