行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　※
※　子計畫三：導盲機器人之聽覺系統與人機溝通介面研製（III）　※
※　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　　■整合型計畫
計畫編號：NSC89－2218－E－009－040－
執行期間：89 **年** 08 **月** 01 **日至** 90 **年** 07 **月** 31 **日**

計畫主持人：林進燈　　教授
共同主持人：

本成果報告包括以下應繳交之附件：

- □赴國外出差或研習心得報告一份
- □赴大陸地區出差或研習心得報告一份
- □出席國際學術會議心得報告及發表之論文各一份
- □國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學電機與控制工程研究所

中　華　民　國 90 **年** 7 **月** 31 **日**

子計畫三：導盲機器人之聽覺系統與人機溝通介面研製（III）

Hearing and Human-machine Communication System of a Guidance Robot for Blind Pedestrians

計畫編號：NSC89-2218-E-009-040
執行期限：89.7.31~90.7.31

主持人：林進燈　國立交通大學　教授
執行機構：國立交通大學電機與控制工程研究所

一、摘要

本研究計畫主要是建立導盲機器人的聽覺系統與口語人機溝通介面，以提供盲胞更方便的操作。在第一年我們研發出抗雜訊的口語辨識器，卻發現語音訊息偵測錯誤會嚴重降低語音辨識器辨識率的問題，因此在第二年我們也成功的研發出在噪音環境下能正確地偵測語音訊號的方法，這兩年我們針對的皆為辨識視障者對導盲機器人所下的口語命令，但導盲機器人也必須回饋視障者一些週遭資訊，在第三年我們便是針對這個部分，成功地建立中文語音合成器，讓視障者與導盲機器人之間能有一道溝通的橋樑。與一般方法所不同的，我們的研究方向著重於韻律訊息產生器的探討，所提出的遞迴式模糊類神經網路是一個結合自我建構模糊類神經推論網路(SONFIN)與多層遞迴式類神經網路的組織架構，經由一些試聽測試後，其結果顯示合成出的語音較以往更為自然。

二、目前研究進度

中文語音合成器的基本架構，主要分為三大部份：語音資料庫、語音合成器及韻律訊息產生器。接下來我們便深入各部份一一地來做介紹。

首先第一個部分為語音資料庫方面。在中文處理上一般有兩種方式：第一種是利用語言學上的知識，制定一套完善的文句剖析法則，或是利用一些經驗法則來分析文句，然而這種方法無法涵蓋所有的文句，因此仍有斷詞錯誤的情形發生。另一種則是蒐集大量的詞彙加以分析以建立一個詞庫，這種方法不太需要語言學上的知識，然而有限的詞庫亦無法包含所有的詞彙，因此有可能發生在詞庫中找不到匹配的詞而造成錯誤。本系統採用的是後者的方式。所使用的資料庫主要分為兩大部分：

（一）**詞典**：提供詞彙供電腦查詢之用，以中研院八萬詞目之詞庫為基礎，經語音處理實驗室整理後，所得到約十一萬詞之詞庫。

（二）**語料庫**：語料庫的內容主要是已經做正確斷詞並標上正確詞類的中文文句，可供建立語言模型的計算與測試之用，使用的是中研院二百三十萬詞平衡語料庫。

接下來的文句分析主要可分成文字前處理(Preprocessing)和自然語言剖析(Parsing)兩個步驟。

（一）**文字前處理**：原始的文字資料，內容可能包括各種數字、日期、時間等特殊的表示法，是無法直接按照其文字念出來的。因此通常需要先經過一到轉換的手續，稱為文字正規化(Text Normalization)，將這些特殊文字轉換成適當的朗讀文字。

（二）**自然語言剖析**：剖析的最終目的，是要瞭解整個文句的句法結構。其中包含了各種層次的語調單位、詞類以及各語詞單位的關係程度，產生抑揚頓挫的變化。因此各語詞的詞類以及彼此的關係程度，將影響其朗讀的韻律變化。

在語文分析過程中，常會有不易

理解的語句存在，歧義性(ambiguity)無疑是一大影響因素，語文的歧義性可以發生在許多層次上，諸如：語句結構、詞彙類別及詞彙意義等。在此，我們列舉一些常常發生的歧義情形：

（一）詞類歧義：有些詞彙具有數種不同的詞類性質，如下面的句子中「制服」可以當動詞或名詞。

- 我看到一件制服歹徒的案子。(動詞)
- 我看到一件制服有很漂亮的配飾。(名詞)

（二）詞間歧義：由於中文句中，詞都沒有標記，因此連續的中文字有時會出現可以斷成好幾種詞的組合，也許只有一種組合才符合句意，但也許不同的組合可各自形成不同的句意。

- 他是一個 守本 分秒必爭 的好青年
- 他是一個 守本分 的好青年

此部份的技術乃利用馬可夫機率模型(Markov model)並結合中研院詞庫小組所發展之分詞標準，將中文斷詞與標詞類同時完成。

第二部分為基本的語音合成器。語音合成的過程，都是先產生一個單位音(Unit Voice)，再由合成器根據韻律資訊加以調整連接，而成為最後的連續語音波形。單位音可長可短，並不一定是一個單音，而是指語音合成的基本單元，與所採用的語音合成方法有密切的關係。這些合成單元包括音素(phoneme)、雙音素(diphoneme)、半音節、單音節(syllable)及長一點的單位音，例如詞(word)和片語(phrase)等。對中文語音合成來說，使用較大的合成單元可以產生比較好的語音合成品質，但是要大量的記憶空間。如果使用較小的合成單元，雖然比較節省記憶空間，但是需要考慮更多的相鄰單位之間連接的處理，而且所合成出來的信號品質通常不好。

此外，國語音系的特點是音節界線分明和音節帶有聲調音位。在國語中共有五種聲調，可分為「一聲」、「二聲」、「三聲」、「四聲」及「輕聲」，五種聲調的分別在於其基頻軌跡(pitch contour)均各自不同，一聲至四聲的基頻軌跡如圖二所示，而輕聲之基頻軌跡會隨前後音的不同而變化。在經過文句分析處理後，就會將各個音節標上發音的方式。因此，在我們的系統中，我們從語音資料庫中，選取 411 個適當的單音節作為合成單元。

語音合成的方法大致分為原音合成、語音模型兩類，我們採用前者來合成語音，所使用的方法為時域基頻同步疊加法(Time Domain Pitch Synchronous OverLap and Add, TD-PSOLA)。在時域上，先將原始語音訊號 $s(n)$，切割成基頻同步的短時信號 $s_m(n)$：

$$s_m(n) = s(n)h_m(t_m - n)$$

其中 m 表第 m 個短時封包，t_m 為原始訊號第 m 個基頻標記的位置，$h_m(t_m - n)$ 為 Hanning window，定義如下：

$$h_m(n) = 0.5 + 0.5\cos(\frac{2\pi n}{N_m}),\quad -\frac{N_m - 1}{2} \le n \le \frac{N_m - 1}{2}$$

N_m 為第 m 個封包的長度。當視窗長度大於一個基頻，這樣使得相鄰的短時信號總有一部份重疊，再按照音韻調整的需求，將前一步驟所得之短時信號轉換成與合成語音基頻標記位置 $\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_q$ 同步的合成短時信號序列 $\tilde{s}_q(n)$。

求得最後的合成語音信號：

$$\tilde{s}(n) = \sum_q s_q(t_q - n)$$

簡單重疊相加：

$$\tilde{s}(n) = \frac{\sum_q \sim_q \tilde{s}_q(n)}{\sum_q \tilde{h}_q\ (\tilde{t}_q - n)}$$

其中 $\sim_q$ 為正規化因子用來補償基頻變

換所造成的能量變化。$\sum_q \tilde{h}_q\ (\tilde{t}_q - n)$ 是用來補償相鄰視窗的重疊不相同所造成能量的變化。

而最小平均方重疊相加：

$$\tilde{s}(n) = \frac{\sum_q \sim_q \tilde{s}_q(n)\tilde{h}_q(\tilde{t}_q - n)}{\sum_q \tilde{h}_q{}^2(\tilde{t}_q - n)}$$

從頻譜上解釋這種合成方法是使合成短時信號 $\tilde{s}_q(n)$ 的頻譜與相對應合成信號的短時頻譜的平方誤差最小。

最後一個部分為韻律產生器。在發音的過程中，氣流留經震動中的聲帶然後進入咽腔與口腔或鼻腔，而造成聲音的強弱、高低和快慢等現象，表現在語音信號上的包括有基頻軌跡、音長、音量、停頓及句調等特徵。經由這樣的信號特徵，除了可以表現出說話者當時的情緒、生理狀況之外，還可以表現出一句話的抑揚頓挫，及和語意有關的訊息，如聲調分辨的訊息。

音韻處理在文句翻語音系統中是極為重要的一個部分。如果我們直接將文句相對應的語音組合起來，那麼所得到的合成語音自然流利度必定不佳。因此，為了使合成的語音更加自然流利，我們必須對輸入的文句結構加以分析，進而產生音韻變化的相關訊息。

韻律訊息產生器的功能將決定文句的朗讀是否具有自然流暢的抑揚頓挫變化。根據自然語言剖析之後所得到的句法結構資訊，韻律訊息產生器必須能產生(合成)相對應的韻律資訊。所謂的韻律，就是聲學上的音色、音高、音強、節奏等特徵的表現。前述特徵為主觀聲學特徵，其所對應的客觀聲學特徵一般是以頻譜封包(Spectral Envelope)、基頻(Fundamental Frequency)、能量、音長和停頓等來表示。

在過去有很多關於韻律訊息產生之方法被提出來，大致上可以區分成三類：規則法(rule-based)、統計法(statistical)、類神經網路法(neural network)。由於類神經網路法較規則法及統計法來得好，因此本系統採用改良式的類神經網路法。

我們所使用的為一個遞迴式模糊類神經韻律模型（圖一），它包含一具有學習能力的自我建構前向類神經模糊推理網路(SONFIN)及一四層遞迴式類神經網路，來模擬人類說話機制，以產生本語音合成系統所需要的韻律參數。因此我們可將此韻律訊息產生模組概分為兩部分：

（一）音韻發聲部分：根據文句分析可以得到音節層次的語言參數，如聲調、拼音等局部的發聲方式，由這些訊息，我們可以初步找出與韻律參數間的對應關係。

（三）文句規則分析部分：輸入的文句先經由文句分析抽取語言參數，再由此模組根據輸入的語言參數去學習人類說話時，整體文句部分的韻律規則。

我們先採用一具有學習能力的自我建構模糊類神經推理網路(SONFIN)，此 SONFIN 網路本身為一模糊系統。初始時，網路本身並無法則的存在，法則的產生與調整乃是由同時進行的結構與參數學習來完成。就結構學習而言，網路的前件部乃是根據對正型的分群法來作彈性分割。後件部的學習，起初是依據分群法來給定每條法則的單值。其後，在必要時，再依序加入較重要的元素(輸入變數)，這些元素並以線性組合的形式存在於後件部中。前件部與後件部的學習可產生一有效率、動態自我增長的網路。此為 SONFIN 網路的一主要特徵。至於參數調整，可由倒傳遞演算法導出。結構與參數學習同時進行的結果，使本網路具快速的學習能力。此外，為了加強 SONFIN 的知識表達

能力，可對輸入變數作線性轉換，如此可減少法則數的使用數目，或提高精確度。這些線性轉換參數也可以在參數學習過程中做動態調整。在輸入參數方面，由於考量人對於詞與句子的關係會依循某些模糊的規則，因此，我們使用整體性語言參數(如字在詞中位置、詞在句中位置、詞類及句長等)作為 SONFIN 的輸入參數。

再來是遞迴式類神經網路架構。最近幾年由於類神經網路的普遍應用，於是也有人把應用在語音合成上。由於類神經網路可採用錯誤回傳法則(Error back-propagaation)與逐步修正記憶之方式，因此我們利用類神經網路能夠自動耦合(Associating)與學習(Learning)兩組資料間的關係，並且將這些關係記憶在類神經網路中。此外，利用遞迴式類神經網路可以學習出產生資料串的方法。遞迴式類神經網路具有記憶以前輸出的效應，於是我們可以利用隱藏層遞迴式類神經網路，以語言參數當作輸入信號，來學習聲學參數的韻律模式，在此主要是掌握詞內韻律的狀態變化。在決定隱藏層神經元的個數方面，由於牽涉到區域韻律狀態變換的多少、整體的複雜度及訓練效率，我們很難決定神經元的個數，只能初步推判可能的韻律變化量，再根據實驗的結果來確定。

三、實驗測試

第一部分的語音資料庫，我們定義了以下幾種參數：

$$\text{斷詞率} = \frac{\text{正確斷出的詞數}}{\text{測試詞數}}$$

$$\text{詞類標示率} = \frac{\text{詞類標示正確的個數}}{\text{測試詞數}}$$

$$\text{句斷詞率} = \frac{\text{整句的斷詞結果均正確之句數}}{\text{測試句數}}$$

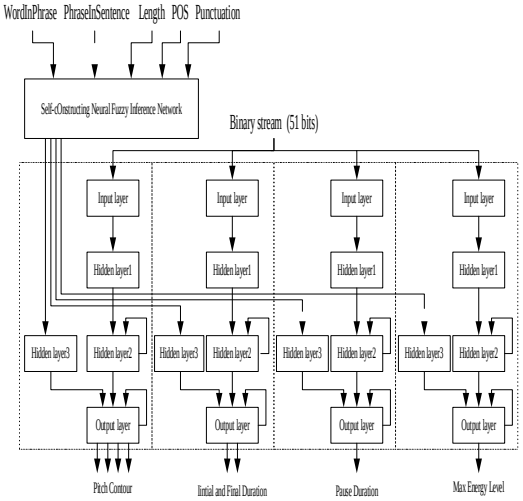
表一列出中文文句剖析器斷詞與標詞之結果。

第二部分為語音合成器的測試，可由以下三大方向來做測試：可辯度、理解力、自然度。由於以上的度量方式沒有一定的標準，所以我們採取主觀評量的方式，測試的成員是以實驗室成員為主並包含其他研究室的同學。測試的結果以理解力最好，大都能夠瞭解整段文句的大意；在可辯度方面，可大致鑑別詞句及字的意思。在自然度上，由於受語音資料庫及合成法的影響，在細微之處不算很好，但就一般而言，受測者多表示可以接受。我們在圖三比較原音與合成音的波形差異。

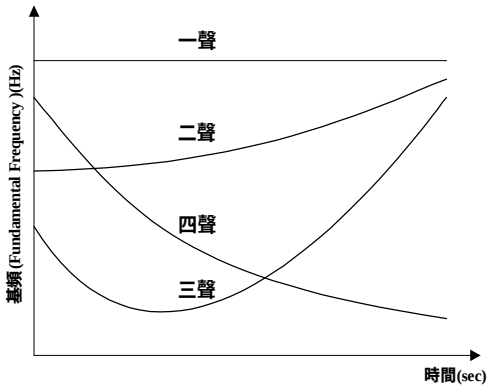
第三部分為韻律產生器，表二所列為本計畫發展之模糊推論韻律規則模型所訓練之韻律參數的均方根誤差，與之前研究的加強式類神經網路韻律模型的均方根誤差(表三)相比較，可以明顯觀察出此模型在韻律參數的學習方面，有很顯著的改善。

四、結論與討論

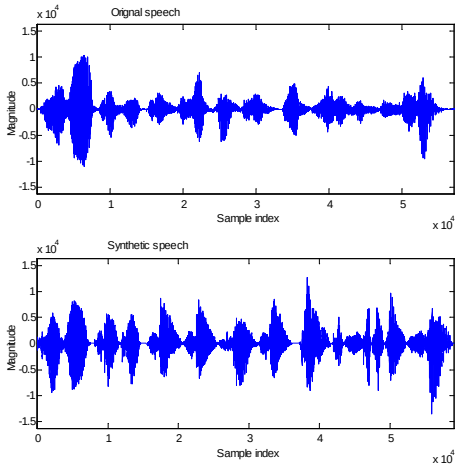
我們利用一具有學習能力的自我建構前向類神經模糊推理網路(SONFIN)及一四層遞迴式類神經網路，正確的提升韻律參數的學習。再配合語音資料庫及語音合成器，構成了一套語調「自然」的中文語音合成器。因此我們在第三年的計畫中，也成功的建立一套以中文語音合成技術為主的安全警告系統，使導盲機器人能利用語音將所偵測到的環境資訊或機器人本身的問題及運動狀況清楚地回報給視障者。



圖一　模糊推論韻律規則模型之架構



圖二　　多層遞迴類神經網路之架構



圖三 (a)原音(b)合成語音的差異

評估方式	正確率
斷詞率	90.24%
詞類標示率	78.67%
句斷詞率	42.87%

表一　　斷詞與標詞之正確率

均方根誤差值	訓練語料	測試語料
基頻	0.86ms/Frame	1.06ms/Frame
音量準位	3.96dB	4.09dB
聲母時長	19.81ms	20.26ms
韻母時長	34.38ms	36.30ms
停頓時長	42.22ms	44.79ms

表二　　模糊推論韻律規則模型之韻律參數均方根誤差

均方根誤差值	訓練語料	測試語料
基頻	1.0ms/Frame	1.5ms/Frame
音量準位	4.96dB	10.78dB
聲母時長	--------	--------
韻母時長	94.38ms	96.07ms
停頓時長	87.54ms	95.76ms

表三　加強式類神經網路韻律模型的均方根誤差