國 立 中 央 大 學

電 機 工 程 研 究 所

碩 士 論 文

結合隱藏式馬可夫模型與類神經網路之國語語音辨識

指導教授 :莊 堯 棠 博士
研 究 生 :林 志 榮

中華民國 八十九 年 六 月

摘要

隱藏式馬可夫模型與類神經網路模型在語音辨識的研究領域上，各佔有一席之地，本篇論文提出一種混合兩種模型的新方法，以結合兩者的優點，提高辨識率及應用的範圍。實驗方面，以國語數字“0”到“9”當做訓練及測試的語音資料，並分成語者相關系統及語者無關系統兩部分進行。所提出的新方法以 HMM-NN-Net、HMM-HMM-Net 與 NN-NN-Net 三種狀態模型系統予以實現，並將實驗結果與隱藏式馬可夫模型相比較。

在語者相關系統方面，三種系統的辨識率皆在九成以上，其中 HMM-NN-Net、NN-NN-Net 狀態模型更可達百分之百，並且在經過對收斂條件適當的調整後，HMM-NN-Net 狀態模型的辨識率以微幅的差距超越隱藏式馬可夫模型。

在語者無關系統方面，HMM-NN-Net 狀態模型以 94.25％的辨識率領先其他模型，進一步證明了新方法的可行性。同時，利用 HMM-NN-Net 與 NN-NN-Net 兩種狀態模型的比較，對類神經網路收斂問題，做完整的分析。

Abstract

Hidden Markov model (HMM) was widely used for speech recognition and has been proved useful in dealing with the statistical and sequential aspects of the speech signal. However, their discriminative properties are weak if they are trained with the maximum likelihood. On the other hand, neural networks (NN) have powerful classification capability but are not well-suited for dealing with time-varying input patterns. In this study, a hybrid HMM-NN speech recognition system that combines the advantages of both models is presented. Three neural net state models, HMM-NN-Net, HMM-HMM-Net and NN-NN-Net, are developed for the proposed hybrid HMM-NN system. All the experimental results are compared with the one obtained from HMM.

In the speaker-dependent experiment, the recognition rates of all the three models are above the level of 90 percent. Furthermore, in spite of the results of HMM-HMM-Net models, all error rates approach to zero after adjusting the criterion.

In the speaker-independent case, HMM-NN-Net model achieves a recognition rate of 94.25 percent and has the best performance compared with other models. Besides, NN-NN-Net model requires less training time than HMM-NN-Net model although its recognition capability cannot compete with HMM-NN-Net model.

The experimental results indicate that the hybrid HMM-NN recognition system based on HMM-NN-Net model improves the performance of traditional HMM system. It is also found that the criterion of neural net state models was related to the recognition capability.

誌謝

首先感謝指導教授莊堯棠老師兩年來孜孜不倦的督促與鼓勵，其豐富的學識及嚴謹的治學態度，帶領我一步步踏實地完成碩士生涯的研究工作，寬大的胸襟及啟發式的教育理念，不但令人印象深刻，更是我努力學習的榜樣。

感謝口試委員余孝先先生、馮蟻剛教授、仝興倫學長及范國清教授在論文方面的建議與指導，特別是仝興倫學長，不論是在程式語言上的啟發、研究方向的確立或是觀念的釐清，皆不厭其煩地從旁予以協助，讓我在研究的過程中，減少許多錯誤的嘗試。

感謝國彰學長在生活上的照顧以及各項資源的提供，志鵬學長在課業上的協助與勉勵。此外，感謝同學克巽、泰宏、岱如及學弟志禮、佳偉、英智、鎮光，因為你們的存在，實驗室增添了許多溫馨與歡樂。

感謝我的父母、兄、姐以及最疼我的奶奶，因為有你們在背後的支持，才能讓我心無旁騖地完成學業。

感謝女友一直以來對我的包容與關懷，在我情緒低落時給我鼓勵，在我的研究遇到瓶頸時與我一起皺眉，謝謝妳陪伴我走過這段時間所有的汗水與歡笑，因為妳，我的生命更加絢爛。

最後，僅以此篇論文獻給我心中所有珍愛的人。

目錄

附圖目錄

表格目錄

第一章 導論

1.1 研究動機

在資訊科技蓬勃發展的時代裡，人類與電腦的關係已密不可分，綜觀目前人機溝通的方式，以按鈕、鍵盤、滑鼠等最為普遍，然而對多數非專業人員而言，這些輸入的工具雖可經由簡單的訓練而學會初步的操作，但顯然並不夠友善，即使是專業人員，也會因這些溝通工具在速度上的限制，降低了電腦的效能。倘若可以用語音輸入的方式，代替生硬的按鍵輸入，將可進一步拉近人與電腦之間的距離。目前語音的應用包括電腦文字輸入、電話查詢系統、汽車電腦系統、大樓門禁管制，以及近年來最熱門的無線通訊等。雖然語音的應用範圍如此廣泛，但語音辨識率的問題，卻影響了實際使用上的普及性，也就因此，目前多數語音辨識的研究仍著眼於正確率的提升。經過三十多年語音識別方法的發展改良，現今語音辨識系統的基本架構以隱藏式馬可夫模型（Hidden Markov Models，HMM）與類神經網路模型（neural networks，NN）最具代表性，兩種模型也各有其本身的優缺點，若能結合兩種模型的優點，可預期地，將能有效的提高系統的辨識能力。

1.2 文獻回顧

隱藏式馬可夫模型發展至今已在語音辨識上佔有舉足輕重的地位，基本理論是由 Baum 及其同僚在 1960 年至 1970 年間所共同發表〔1〕-〔5〕並由 Baker〔6〕及 Jelinek〔7〕-〔13〕實際應用在語音信號的處理上。隱藏式馬可夫模型會如此地受到歡迎，最大的原因在於模型中使用了不可觀測的推測程序，雖然狀態不可直接觀測，但是卻可透過其他可觀測的推測過程而獲得。對於語音自動辨識系

統而言，這項特性正足以克服語音信號中時變（time-varying）的問題〔14〕。

類神經網路模型應用在語音識別上，則是倚賴其強大的分類能力。Hinton 等人以時間延遲（time-delay）之類神經網路架構進行獨立字之辨識〔15〕，Bendiksen 與 Steiglitz〔16〕、Ghiselli-Crippa 與 El-Jaroudi〔17〕、Qi 與 Hunt〔18〕利用網路將語音信號以有聲、無聲、靜音（voiced-unvoiced-silenced）作分類，Kuhn 等〔19〕、Kim 等〔20〕、Hunt〔21〕、Ching 等〔22〕及 Chen 等〔23〕以循環類神經網路模型（recurrent neural networks）分別對數字及音節做辨識。

同時使用隱藏式馬可夫模型與類神經網路模型的語音辨識系統，為 Bourland 與 Wellekens 提出，以多層感知器（multi-layer perceptions，MLP）連結在隱藏式馬可夫模型之後，用來解決前後文脈相關問題之辭彙辨識〔24〕。

1.3 研究目標

目前以狀態（states）為基本單元的語音辨識系統，都是建構在隱藏式馬可夫模型理論的基礎上。由於採用了不可觀測的狀態序列，隱藏式馬可夫模型提供了一個非靜態的統計模型（non-stationary statistical model），相對於無法以時間長短作為聲音特徵的語音辨識系統，隱藏式馬可夫模型取代了動態時間校準（dynamic time warping），有效地解決了語音隨時間變化的問題。傳統隱藏式馬可夫模型是以兩組機率密度分佈建立對應的狀態模型，包括狀態轉移機率及狀態觀測機率，其中狀態觀測機率多使用混合高斯機率函數（a mixture of Gaussian distributions）來估計。雖然因此簡化了模型的訓練，卻也讓模型缺少了錯誤拒絕的能力。類神經網路模型中分

類能力的優勢，正好可以補強這項缺點，只可惜網路的複雜度及模型冗長的訓練時間常令人怯步。本篇論文提出三種方法，將隱藏式馬可夫模型與類神經網路模型合併，希望藉由兩種模型優點的結合，能達到以下兩點目標：

（1）相對於單純只使用混合高斯機率函數之隱藏式馬可夫模型，經由類神經網路模型的引入，提高辨識率。

（2）以不同的音框分配方式，對所有訓練及測試語句進行狀態歸類，以提高辨識率，並縮短類神經網路模型訓練時間。

1.4 方法簡介

完整的語音辨識系統必須包含兩個主要步驟：（一）以訓練樣本建立辨識模型，（二）輸入測試樣本進行辨識。以下依序就這兩項分別加以介紹：

1.4.1 以訓練樣本建立辨識模型

為了將隱藏式馬可夫模型中狀態序列的觀念引入類神經網路模型，設定每一個類神經網路狀態模型對應於一個獨立的隱藏式馬可夫狀態模型，換句話說，類神經網路狀態模型的個數相等於隱藏式馬可夫模型中的狀態數。網路的輸入層處理單元相等於一個音框的特徵值個數，輸出層處理單元個數則設為 1，用以代表每個狀態的觀測機率值。

有別於隱藏式馬可夫模型中以最大相似度（maximum likelihood）主動學習並估測模型參數，類神經網路辨識模型採用監督式（supervised）網路中之倒傳遞網路（back-propagation network）加以訓練，因此所有語音訓練資料在饋入網路前必須先進行音框分配的工作，以確定每一個音框所歸屬的狀態。分配音框的方式分為下

列兩種：

（1）隱藏式馬可夫模型：使用最大相似度觀念，以維特比演算法（Viterbi algorithm），找出最大觀測機率值之音框分配。

（2）自我監督類神經網路模型：重新設定一個新的類神經網路模型，利用網路中自我監督（self-supervised）的能力，加上維特比演算法，自行調整音框分配方式，直到分配方式不再改變。

待所有訓練語句以上述其中一種方法分配音框後，便可將已完成狀態歸屬記錄的所有訓練音框，分別饋入原先設定好的類神經網路狀態模型。若訓練音框屬於該狀態模型，則網路目標輸出值設為“1”，反之，網路目標輸出值設為“0”，以此方式重覆訓練所有狀態模型，直到所有訓練音框之網路實際輸出與目標輸出值間誤差皆小於要求，則辨識模型訓練完成。

1.4.2 輸入測試樣本進行辨識

透過對訓練樣本的充分訓練，可控制類神經網路狀態模型的輸出值在“0”與“1”之間，對應到傳統隱藏式馬可夫模型中，此網路輸出值相當於以混合高斯機率密度函數所計算出之相似度。將所有類神經網路狀態模型以符合邏輯方式排列成所有可能的狀態序列，如此便構成完整的辨識模組。測試語句音框依序饋入所指定的狀態模型並求取網路輸出值，若網路輸出值愈接近“0”，表示輸入樣本與該狀態相關性愈低，若網路輸出值愈接近“1”，表示輸入樣本與該狀態相關性愈高。至於測試語句音框選擇狀態模型的方式，可分為下列兩種：

（1）以辨識模組本身為依據，使用維特比演算法求得音框序列最佳狀態路徑。

（2）測試句語音框不先經過辨識模組本身，而是利用事先訓練完成之隱藏式馬可夫模型進行狀態模型的選擇。

最後，比較所有辨識模組的累積網路輸出值，具有最大值者便是所要求的辨識結果。

1.5 論文大綱

本篇論文的第一章為導論，說明研究動機、參考文獻、研究目標，並對論文中所提方法做簡單的介紹。第二章為理論基礎，以三個部份分別介紹語音資料的特徵值擷取、隱藏式馬可夫模型的理論以及類神經網路模型演算法。第三章針對本篇論文所提出結合隱藏式馬可夫模型與類神經網路模型之辨識系統做進一步詳細的說明，內容包含模型的訓練階段與辨識階段兩部份，分別介紹不同系統的建立方法與詳細流程。第四章為實驗結果與討論，分成語者相關與語者無關兩種系統對新模型進行測試。第五章為結論及未來展望。

第二章 理論基礎

2.1 特徵參數的求取

目前在語音辨識的研究上，語音特徵參數以梅爾倒頻譜係數（mel-cepstral coefficients）**與倒頻譜係數**（cepstral coefficients）**最為普遍，其辨識能力也勝過其他參數。在本篇論文中選用倒頻譜係數做為特徵參數，故以下僅對倒頻譜係數作介紹。**

首先將原始數位語音信號切割成多個獨立音框（frame）**，再將音框內的語音信號通過一階高通濾波器做預強調**（pre-emphasis）**處理，用以補償信號在接收時高頻能量的損失：**

$$y[0] = x[0], \qquad (2\text{-}1)$$

$$y[n] = x[n] - 0.95x[n-1] \qquad 1 \le n \le L \qquad (2\text{-}2)$$

其中的 L **為每一個音框的取樣個數。**

將每個音框乘上漢明視窗（hamming window）**，減少因能量的不連續性所造成的失真：**

$$\begin{aligned} s[n] &= y[n] \times h[n] \\ &= y[n] \times \left(0.54 - 0.46\cos\frac{2n\pi}{n-1} \right) \end{aligned} \qquad (2\text{-}3)$$

利用德賓遞迴程序（Durbin's recursive procedure）**求出線性預估係數**（linear prediction coefficients，LPC）a_j **：**

$$E^{(0)} = R(0) \qquad (2\text{-}4)$$

$$k_i = \left(R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j) \right) \Big/ E^{(i-1)} \qquad 1 \le i \le p \qquad (2\text{-}5)$$

$$a_i^{(i)} = k_i \qquad (2\text{-}6)$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \qquad (2\text{-}7)$$

$$E^{(i)} = (1 - k_i^2)E^{(i-1)} \qquad (2\text{-}8)$$

$$a_j = a_j^{(p)} \qquad 1 \le j \le p \qquad (2\text{-}9)$$

其中

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k) \qquad 0 \le k \le p \qquad (2\text{-}10)$$

p **為線性預估係數的階數。**

由線性預估係數求得倒頻譜係數 c_n **：**

$$c_1 = a_1 \qquad (2\text{-}11)$$

$$c_n = \begin{cases} a_n + \sum_{m=1}^{n-1}\left(1 - \frac{m}{n}\right)a_m c_{n-m} & 1 \le n \le p \\ \sum_{m=1}^{p}\left(1 - \frac{m}{n}\right)a_m c_{n-m} & n \ge p \end{cases} \qquad (2\text{-}12)$$

其中 n **為線性預估係數的階數，** p **為倒頻譜係數的維度。**

此外，還可由倒頻譜係數進一步求得轉移倒頻譜係數（delta-ceptral coefficients **）** Δc_n **：**

$$\Delta c_n(t) = \frac{\partial c_n(t)}{\partial t} = \frac{\sum_{k=-K}^{K} k \cdot c_n(t+k)}{\sum_{k=-K}^{K} k^2} \qquad (2\text{-}13)$$

2.2 隱藏式馬可夫模型

隱藏式馬可夫模型是一種以統計方法為理論基礎，針對語音信號特性所發展出來的辨識模型，對於不同語音特徵的識別取決於從模型中計算所產生的機率值。一個完整的隱藏式馬可夫模型可以（2-14 **式）表示：**

$$l = (p, A, B) \qquad (2\text{-}14)$$

模型中包含了 (p, A, B) **三個參數，各自代表的意義如下：**

（1）**初始狀態機率**（initial state probability）：

$$p = \{p_i = prob(q_1 = S_i)\} \qquad 1 \le i \le N \qquad \text{（2-15）}$$

（2）**狀態轉移機率**（state transition probability）：

$$A = \{a_{ij} = prob(q_{t+1} = S_j \mid q_t = S_i)\} \qquad 1 \le i, j \le N \qquad \text{（2-16）}$$

（3）**觀測符號機率**（observation symbol probability）：

$$B = \{b_j(O_t) = prob(O_t \mid q_t = S_j)\} \qquad 1 \le j \le N \qquad \text{（2-17）}$$

其中 $\mathbf{O} = \{O_1, O_2, \cdots\cdots, O_T\}$ **為觀測序列，** $\mathbf{S} = \{S_1, S_2, \cdots\cdots, S_N\}$ **為模組狀態序列，** $\mathbf{q} = \{q_1, q_2, \cdots\cdots, q_T\}$ **為觀測狀態序列，** T **為觀測序列長度，** N **為狀態總數。**

在語音信號中， $\mathbf{O} = \{O_1, O_2, \cdots\cdots, O_T\}$ **可視為各個音框的特徵參數，** T **為此一段語音的音框數，** N **為所對應的模組狀態數，同時，原始複雜的馬可夫鏈**（Markov chain）（**如圖** 2-1）**可簡化為符合時序的由左至右模型**（left-right model）（**如圖** 2-2）。**為減少計算量，令初始狀態機率** p **與狀態轉移機率** A **為** 1，**觀測狀態機率** B **則採用多變數高斯機率密度函數計算，其定義如下：**

$$D_i(\boldsymbol{t}_T) = (2\pi)^{-\frac{N}{2}} |R_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{t}_T - \boldsymbol{t}_{Ri})^T R_i^{-1}(\boldsymbol{t}_T - \boldsymbol{t}_{Ri})\right] \qquad \text{（2-18）}$$

其中 N **為特徵向量的維度，** t_T **為待測語音或訓練語音的特徵向量，** t_{Ri} **為模型中該狀態第** i **個混合數**（mixture）**的期望值，** R_i **為模型中該狀態第** i **個混合數的協方差矩陣**（covariance matrix）。

實際演算中，我們將觀測狀態機率取自然對數，則累積總觀測狀態機率可由相加個別觀測狀態機率而得：

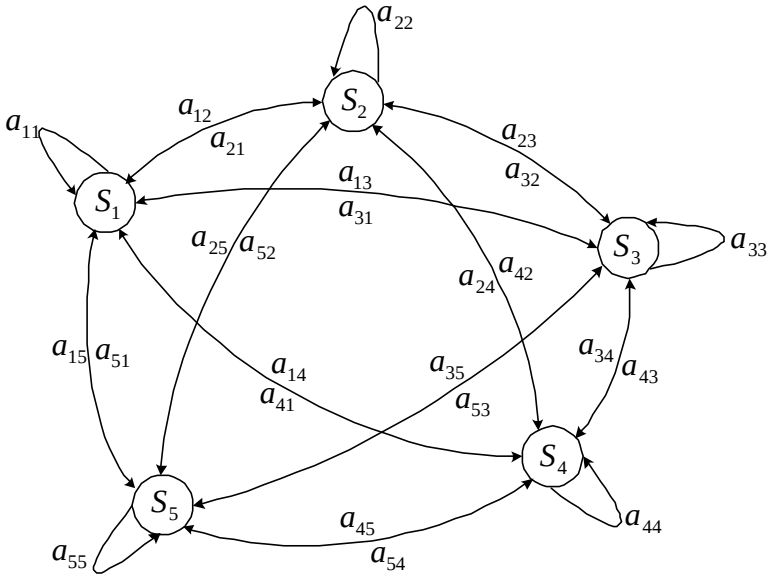


圖 2-1　狀態數為 5 的馬可夫鏈。狀態標示為 $S_1, S_2, \cdots, S_5$，所有狀態間的轉移皆不設任何限制。

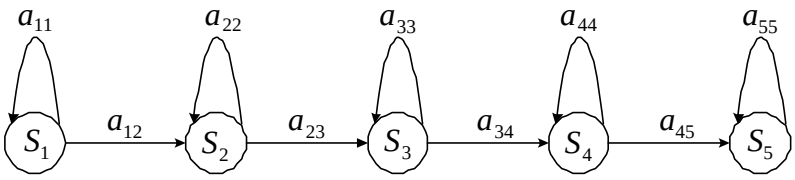


圖 2-2　狀態數為 5 的左至右馬可夫鏈。狀態標示為 $S_1, S_2, \cdots, S_5$，狀態只能停留在本身或向右傳遞。

$$\ln D_i(\boldsymbol{t}_T) = (-\frac{N}{2})\ln(2\boldsymbol{p}) + (-\frac{1}{2})\ln|R_i| + \left[-\frac{1}{2}(\boldsymbol{t}_T - \boldsymbol{t}_{Ri})^T R_i^{-1}(\boldsymbol{t}_T - \boldsymbol{t}_{Ri})\right] \quad (2\text{-}19)$$

假設 t_{Tj} **與** t_{Tk} **之間為獨立（** independent **），** R_i **可視為對角矩陣，再去除共同項** $(-\frac{N}{2})\ln(2\boldsymbol{p})$ **則上式可化簡為**

$$\ln D_i(\boldsymbol{t}_T) = (-\frac{1}{2})\sum_{k=0}^{N-1}(\ln r_{ikk}) - \frac{1}{2}\sum_{k=0}^{N-1}(\boldsymbol{t}_{Tk} - \boldsymbol{t}_{Rk})^2 r_{kk} \quad (2\text{-}20)$$

其中 r_{ikk} **為** R_i **的對角元素。**

將 $b_j(O_t)$ **以** $\ln D_i(\boldsymbol{t}_T)$ **取代，上式便是語音信號中個別音框相對該狀態之觀測狀態機率值，接下來再利用維特比演算法找出一條最佳的狀態序列，同時算出整段語音在該模組的最大總觀測狀態機率。**

維特比演算法分為四個步驟：

（1）初始（ initialization **）：**

$$\boldsymbol{d}_1(i) = b_i(O_1) \qquad 1 \le i \le N \quad (2\text{-}21)$$

$$\Psi_1(i) = 0 \qquad 1 \le i \le N \quad (2\text{-}22)$$

（2）遞迴（ recursion **）：**

$$\boldsymbol{d}_t(j) = b_j(O_t) \cdot \max_{1\le i\le N} \boldsymbol{d}_{t-1}(i) \qquad 2 \le t \le T, 1 \le j \le N \quad (2\text{-}23)$$

$$\Psi_t(j) = \arg\max_{1\le i\le N} \boldsymbol{d}_{t-1}(i) \qquad 2 \le t \le T, 1 \le j \le N \quad (2\text{-}24)$$

（3）結束（ termination **）：**

$$P^* = \max_{1\le i\le N} \boldsymbol{d}_T(i) \quad (2\text{-}25)$$

$$S^* = \arg\max_{1\le i\le N} \boldsymbol{d}_T(i) \quad (2\text{-}26)$$

（4）路徑回溯（ path backtracking **）：**

$$S^*_t = \Psi_{t+1}(S^*_{t+1}) \qquad t = T-1, T-2, \cdots\cdots, 1 \quad (2\text{-}27)$$

其中 T 為此一段語音的音框數，N 為所對應的模組狀態數，$d_t(i)$ 為時刻 t 到達狀態 i 的最大機率，$\Psi_t(i)$ 記錄時刻 t 到達狀態 i 時，時刻 $t-1$ 所到達的狀態，S^* 記錄最佳的狀態序列，而 P^* 為此整段語音最大總機率值。

語音識別的方法便是利用此隱藏馬可夫模型所求得的最大總機率值 P^* 作為辨識的依據。

2.3 類神經網路

類神經網路是一種模仿生物神經網路的資訊系統，具有適應性、容錯性以及由經驗中學習的能力。不同於隱藏式馬可夫模型是為解決語音信號時變特性所發展出來的理論，類神經網路廣泛地被使用在各種領域的樣本識別上。雖然類神經網路發展至今仍無法完全模擬人類複雜的思考模式，但卻可提供許多值得參考的決策依據。類神經網路模型依網路學習演算方式，可分為監督式學習網路（supervised learning network）和非監督式學習網路（unsupervised learning network）。

所謂監督式學習網路，必須將給定的一序列訓練向量或樣本，根據所選用的學習演算法則對網路加以訓練，調整網路中的權值，使所有與樣本相對應的網路輸出值符合預定的設定值。由於每一樣本已事先指定輸出值，監督式的名稱便由此而來。代表性的網路模型有感知機網路（perceptron）、倒傳遞網路、學習向量量化網路（learning vector quantization）等。相對於監督式學習網路，非監督式學習網路不須指明有哪些群集及群集內所包含的樣本，利用自我組織（self-organizing）方式，修改網路權值，將相似的輸入向量分配至相同的輸出單元，同時，每一個形成的群集將產生自己的代表

向量。代表的網路模型如自我組織映射圖網路（self-organizing maps），適應性共振理論網路（adaptive resonance theory）。

在語音辨識的應用上，大部分須在網路訓練時指定輸出值，並不適合於非監督式學習網路，因此只考慮監督式學習網路。以下就監督式學習網路中最常見的倒傳遞網路作介紹。

2.3.1 倒傳遞網路的定義與學習原理

倒傳遞網路使用梯度下降法（gradient descent method），以將輸出值與目標值的誤差予以最小化的過程完成網路的訓練，使用單一隱藏層之倒傳遞類神經網路，模型架構如圖 2-3。倒傳遞網路訓練的過程包含三階段：（1）將訓練樣本以前饋（feedforward）方式輸入網路，（2）以反饋（feedbackward）方式計算相對誤差及權值調整量，（3）調整權值。

定義：

$\mathbf{o_i} = (o_{i1}, o_{i2}, \cdots o_i \cdots, o_L)$	輸入層單元向量
$\mathbf{o_j} = (o_{j1}, o_{j2}, \cdots o_j \cdots, o_M)$	隱藏層單元向量
$\mathbf{o_k} = (o_{k1}, o_{k2}, \cdots o_k \cdots, o_N)$	輸出層單元向量
$\mathbf{t} = (t_1, t_2, \cdots t_k \cdots, t_N)$	目標輸出向量
$\mathbf{w_{ji}} = (w_{j1i1}, \cdots w_{ji} \cdots, w_{LM})$	隱藏層處理單元權值
$\mathbf{w_{kj}} = (w_{k1j1}, \cdots w_{kj} \cdots, w_{MN})$	輸出層處理單元權值
$\mathbf{è_j} = (q_{j1}, q_{j2}, \cdots q_j \cdots, q_M)$	隱藏層偏權值（bias）
$\mathbf{è_k} = (q_{k1}, q_{k2}, \cdots q_k \cdots, q_N)$	輸出層偏權值
$f(x)$	致動函數（activation function）
h	學習速率

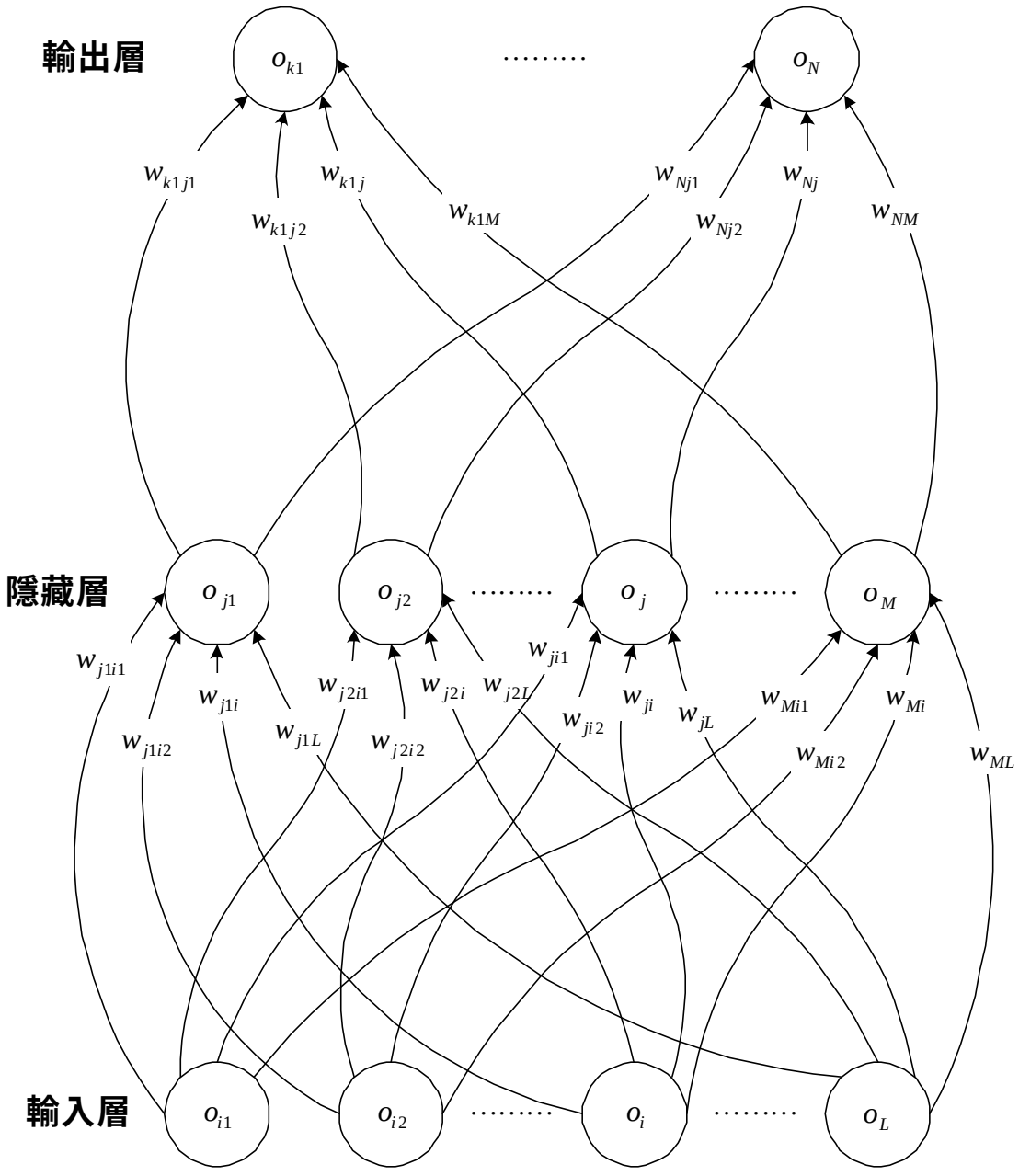


圖 2-3　單一隱藏層之類神經網路架構。輸入層單元數為 L，隱藏層單元數為 M，輸出層單元數為 N。

學習原理：

對隱藏層而言，輸入值為

$$net_j = \sum_{i=1}^{L} w_{ji} o_i + \theta_j \qquad 1 \le j \le M \qquad (2\text{-}28)$$

輸出值為

$$o_j = f(net_j) \qquad 1 \le j \le M \qquad (2\text{-}29)$$

對輸出層而言，輸入值為

$$net_k = \sum_{j=1}^{M} w_{kj} o_j + \theta_k \qquad 1 \le k \le N \qquad (2\text{-}30)$$

輸出值為

$$o_k = f(net_k) \qquad 1 \le k \le N \qquad (2\text{-}31)$$

實際輸出值與目標輸出值間誤差定義為

$$E = \frac{1}{2} \sum_{k=0}^{N} (t_k - o_k) \qquad 1 \le k \le N \qquad (2\text{-}32)$$

網路學習的目的即是使（2-32 式）達到最小化，因此定義輸出層權值調整量 Δw_{kj} 為

$$\Delta w_{kj} = -\eta \frac{\partial E}{\partial \Delta w_{kj}} \qquad 1 \le j \le M, 1 \le k \le N \qquad (2\text{-}33)$$

經由連鎖律（chain rule）可得輸出層權值調整量 Δw_{kj} 及偏權值調整量 $\Delta \theta_k$

$$\Delta w_{kj} = \eta \delta_k o_j \qquad 1 \le j \le M, 1 \le k \le N \qquad (2\text{-}34)$$

$$\Delta \theta_k = -\eta \delta_k \qquad 1 \le k \le N \qquad (2\text{-}35)$$

其中

$$\delta_k = (t_k - o_k) f'(net_k) \qquad (2\text{-}36)$$

同理，定義隱藏層權值調整量 Δw_{ji} 為

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial \Delta w_{ji}} \qquad 1 \le i \le L, 1 \le j \le M \qquad (2\text{-}37)$$

經由連鎖律可得隱藏層權值調整量 Δw_{ji} **及偏權值調整量** $\Delta \theta_j$

$$\Delta w_{ji} = \eta \delta_j o_i \qquad 1 \le i \le L, 1 \le j \le M \qquad (2\text{-}38)$$

$$\Delta \theta_j = -\eta \delta_j \qquad 1 \le j \le M \qquad (2\text{-}39)$$

其中

$$\delta_j = f'(net_j) \sum_{k=1}^{N} \delta_k w_{kj} \qquad 1 \le j \le M \qquad (2\text{-}40)$$

常用致動函數有二元 s **形函數**（binary sigmoid function），**函數圖形如圖** 2-4，**二極** s **形函數**（bipolar sigmoid function），**函數圖形如圖** 2-5。**若** $f(net_j)$ **使用斜率參數** s **為** 1 **之二元** s **形函數：**

$$f(net) = \frac{1}{1 + e^{-net}} \qquad (2\text{-}41)$$

$$f'(net) = f(net)[1 - f(net)] \qquad (2\text{-}42)$$

則 δ_j **和** δ_k **可寫成**

$$\delta_k = (t_k - 0_k) o_k (1 - o_k) \qquad 1 \le k \le N \qquad (2\text{-}43)$$

$$\delta_j = (\sum_{k=1}^{N} \delta_k w_{kj}) o_j (1 - o_j) \qquad 1 \le j \le M \qquad (2\text{-}44)$$

一般的通常網路學習過程採每饋入一個訓練樣本即更新網路一次，換句話說，每饋入一個訓練樣本便利用權值調整量及偏權值調整量對網路權值及偏權值進行修正。在輸出層部份：

$$w_{kj(new)} = w_{kj(old)} + \Delta w_{kj} \qquad 1 \le j \le M, 1 \le k \le N \qquad (2\text{-}45)$$

$$\theta_{k(new)} = \theta_{k(old)} + \Delta \theta_k \qquad 1 \le k \le N \qquad (2\text{-}46)$$

在隱藏層部份：

$$w_{ji(new)} = w_{ji(old)} + \Delta w_{ji} \qquad 1 \le i \le L, 1 \le j \le M \qquad (2\text{-}47)$$

$$\theta_{j(new)} = \theta_{j(old)} + \Delta \theta_j \qquad 1 \le j \le M \qquad (2\text{-}48)$$

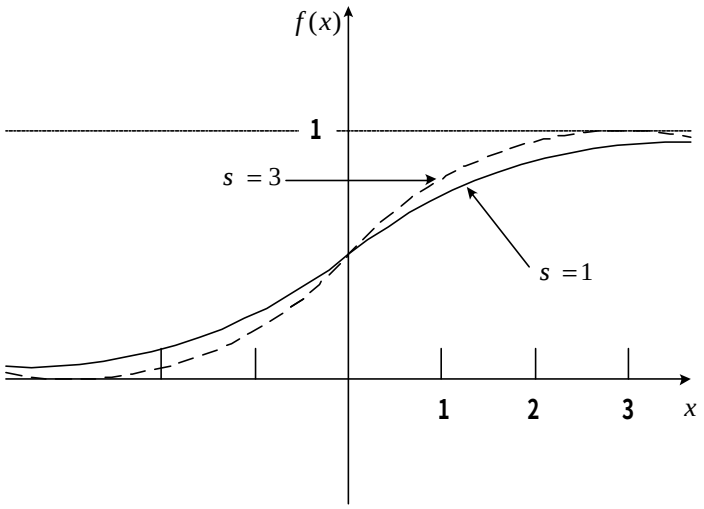


圖 2-4　**二元** s **形函數圖形。虛線及實線分別代表** $s=3$ **與** $s=1$ **之函數圖形。**

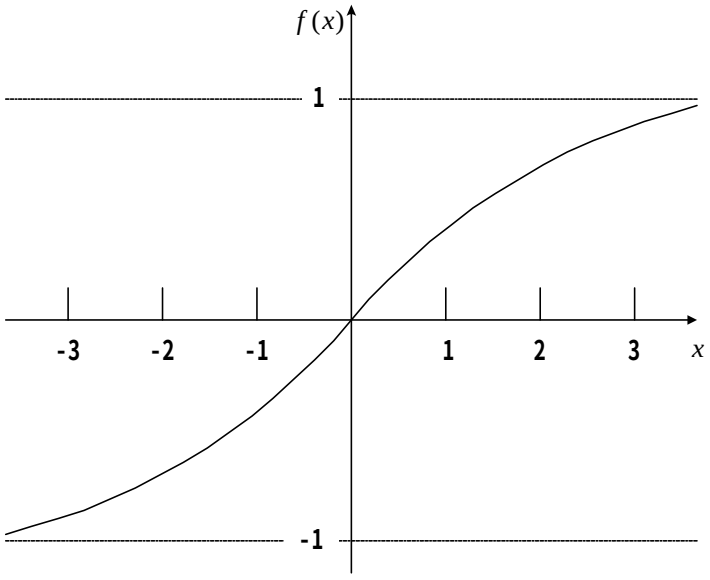


圖 2-5　**二極** s **形函數圖形。**

2.3.1 倒傳遞網路的訓練方法

圖 2-6 為倒傳遞網路訓練流程圖，詳細運算流程可分為 9 個步驟，如下所述：

步驟 0. 設定網路架構與收斂條件，並給定初始權值及偏權值。

步驟 1. 若不滿足終止條件，做步驟 2-9。

步驟 2. 對每一組訓練樣本，做步驟 3-8。

前饋傳遞：

步驟 3. 每一個輸入層單元（ o_i ）接收輸入信號，並將信號廣播到上一層的所有單元（隱藏層單元）。

步驟 4. 隱藏層單元的輸入值等於其輸入信號之加權總和（2-28 式），再以致動函數計算其輸出信號（2-29 式），將輸出信號傳送到上一層（輸出層單元）。

步驟 5. 輸出層單元的輸入值等於其輸入信號之加權總和（2-30 式），再以致動函數計算其輸出信號（2-31 式）。

反饋傳遞：

步驟 6. 每一個輸出層單元（ o_k ）接收此訓練樣本所相對應之目標輸出（ t ），計算輸出層誤差訊息（2-36 式），計算輸出層權值調整量（2-34 式）與偏權值調整量（2-35 式），並將誤差訊息傳送至下一層（隱藏層）。

步驟 7. 計算隱藏層誤差訊息（2-40 式），計算隱藏層權值調整量（2-38 式）與偏權值調整量（2-39 式）。

更新權值與偏權值：

步驟 8. 更新輸出層權值與偏權值（2-45 式、2-46 式），更新隱藏層權值與偏權值（2-47 式、2-48 式）。

步驟 9. 檢驗收斂條件。

一般的情形下，網路權值與偏權值的初始值可以亂數設定，通常介於 0.5 至-0.5 或 1 至-1 之間。針對基本類神經網路架構，已有許多修正方式被提出，以改善網路的效能。這些修正方式包括改變更新權值的程序、使用不同的致動函數等，至於使用時機，則與網路的功能及訓練樣本的特性有密切的關係。因為這些修正方式並非本文重點，故不再詳述。

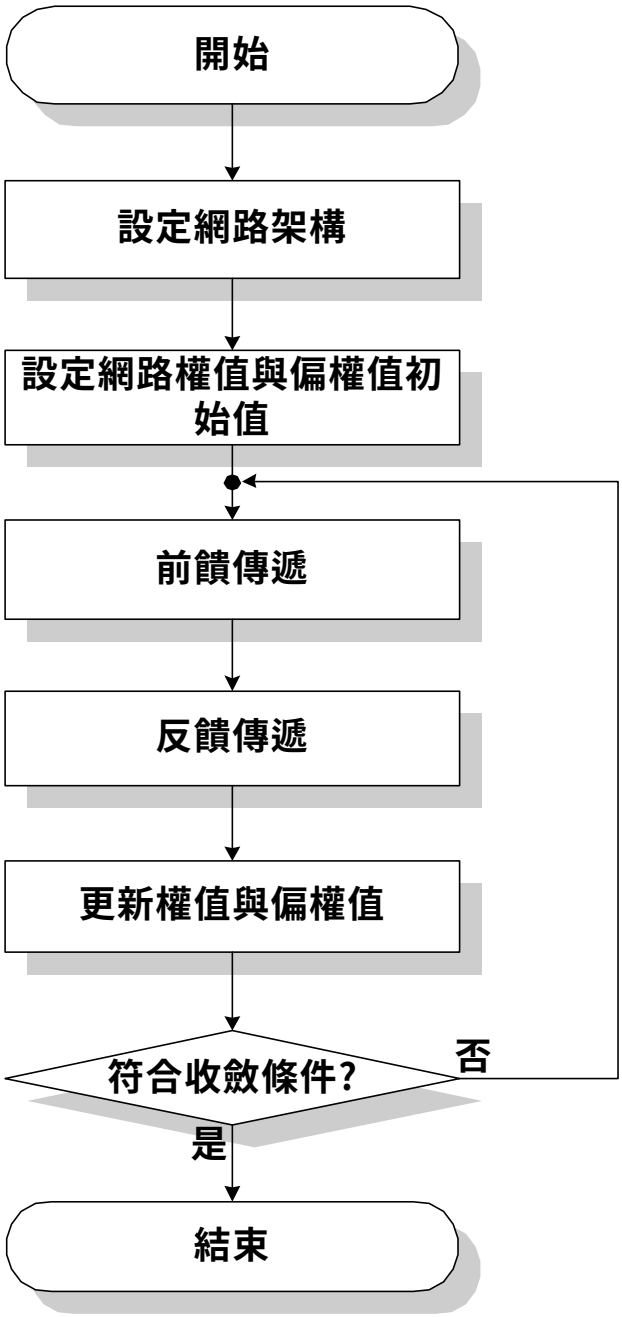


圖 2-6　倒傳遞網路訓練流程圖

第三章 結合隱藏式馬可夫模型與類神經網路模型之語音辨識系統

本章分為模型訓練階段及模型辨識階段兩部分，詳細介紹藉由結合隱藏式馬可夫模型與類神經網路模型兩者優點所提出的類神經網路狀態模型辨識系統。在隨後的章節將利用本章所介紹的方法，組合成三種不同的狀態模型辨識系統，做為實際驗証的依據。

3.1 模型訓練階段

由於本篇論文所使用的倒傳遞網路是監督式的類神經網路，因此在訓練類神經網路狀態模型時，所有訓練語句音框必須先被決定在固定的狀態中，才足以指定網路的目標輸出值，換句話說，在訓練狀態模型之前，必須額外建立一個音框分配系統。本篇論文以兩種不同的方式實現此音框分配系統：（1）以訓練語句先行建立隱藏式馬可夫模型，並在模型訓練完成後，記錄所有語句音框歸屬。（2）設定另一組類神經網路模型，利用其自我監督能力分配音框。兩種方式詳細的進行程序分述如後。

3.1.1 隱藏式馬可夫模型音框分配系統

傳統隱藏式馬可夫模型乃是採用由多變數高斯機率密度函數（2-18 式）所求出之機率值做為辨識之度量，在實際演算中，採用了化簡過後的計算值（2-20 式）。假設 M 段具有相同狀態模型的訓練語音資料，在經過特徵參數（以本論文而言，為 14 維的倒頻譜係數加上 14 維的轉移倒頻譜係數）的求取後，各有 T_1，T_2，・・・，T_M 個音框，欲訓練模組狀態數為 N（相當於有 20 句相同或不相同的訓練者所念的國語數字“1”，模組狀態數為 6）。在第一次進入

維特比演算法之前必須先對各語句音框做狀態切割的動作（通常是做均勻分配），建立起始隱藏式馬可夫模型，再進行後面的訓練程序。假設在某次維特比演算法訓練完成後，第一句語音資料分配在狀態 S_1 之音框數為 3，第二、三句語音資料為 2，第 M 句語音資料為 3，則狀態 S_1 的隱藏式馬可夫模型可由 t_{R1}、R_1 二個參數來建立，其中 t_{R1} 為分配至狀態 S_1 中所有音框（總音框數為 P_1）的期望值，R_1 為協方差，模型參數如圖 3-1。利用連續疊代，調整模型參數，直到所有模型皆不再變動，則可記錄此時所有訓練音框的狀態歸屬，並結束模型的訓練。詳細流程如圖 3-2。

3.1.2 自我監督類神經網路模型音框分配系統

首先設定網路架構，網路輸入層每次饋入的是一個音框的特徵值，因此輸入層單元數為特徵值的維度（本篇論文為 28 維特徵參數），採用具有 50 個處理單元的單一隱藏層，輸出層單元數為該組訓練樣本欲訓練之狀態模組的狀態數（本篇論文中每個獨立數字的模組狀態數為 6），並以亂數設定網路權值與偏權值。以訓練樣本而言，此網路模型與連接在後之辨識模型不同之處在於此網路模型只用來分配具有相同狀態模組的訓練語句音框，不同的狀態模組須設定不同的網路，因此不具有相同狀態模組的其他訓練語句勿須包含在不相關的網路訓練樣本中。各語句在第一次饋入網路前，先做狀態切割的動作（本篇論文是做均勻分配），饋入網路後，該音框所屬的狀態單元目標輸出值設為“1”，其他的狀態單元目標輸出值設為“0”。一個完整的網路訓練流程，採一次完成一個語句的訓練，也就是說，必須不斷更新網路，直到同一語句中所有音框的實際輸出值與目標輸出值間誤差皆符合要求，始能訓練下一語句。待所有訓練語句皆完成一個網路訓練流程，以調整過權值與偏權值的網路重

新分配所有訓練語句音框。至於音框對於狀態的選擇，則以網路輸出值最接近“1”的狀態單元為優勝。為了符合狀態序列由左至右的條件，乃藉助維特比演算法尋找最佳狀態路徑。對於被重新分配狀態的音框，必須再饋入網路進行訓練，直到所屬狀態不再改變。網路每經過一次完整的更新動作後，須重新檢視所有訓練語句音框的狀態歸屬，若仍有變化則重覆網路訓練，若音框不再重新分配則終止網路訓練，並記錄音框分配情形。詳細流程如圖 3-3。

3.1.3 完整訓練流程

在完成訓練樣本音框分配之後，接著開始訓練類神經網路狀態模型。首先設定狀態模型輸入層單元數為特徵值的維度（本篇論文為 28），單一隱藏層的單元數為 50，輸出層單元數為 1。由於狀態模型的網路輸出是“正確”或“錯誤”的分類問題，因此，當訓練語句音框輸入正確的狀態模型時，網路目標輸出值設為“1”，當輸入錯誤的狀態模型時，網路目標輸出值設為“0”。相對於隱藏式馬可夫模型一個狀態建立一個模型，類神經網路狀態模型總數也必須相等於所有模組的總狀態數（本篇論文中總狀態數為 60）。假設有 M 段具有相同狀態模型的訓練語音資料，在完成音框分配的動作後，歸屬於狀態 S_1 之總音框數為 P_1，若欲訓練狀態 S_1 之網路模型，則將此 P_1 個音框（$\mathbf{x}_1$）分別饋入網路模型 S_1，同時網路目標輸出值設為“1”，再將其他不是屬於狀態 S_1 的音框饋入網路模型，網路目標輸出值設為“0”（如圖 3-4），對網路模型做疊代運算，調整網路權值及偏權值，直到所有網路實際輸出值與目標輸出值間誤差符合要求，則此狀態模型建立完成，訓練流程如圖 3-5。所有其他狀態模型也依此程序進行訓練，最後則完成一個，以狀態為單位的類神經網路辨識模型。

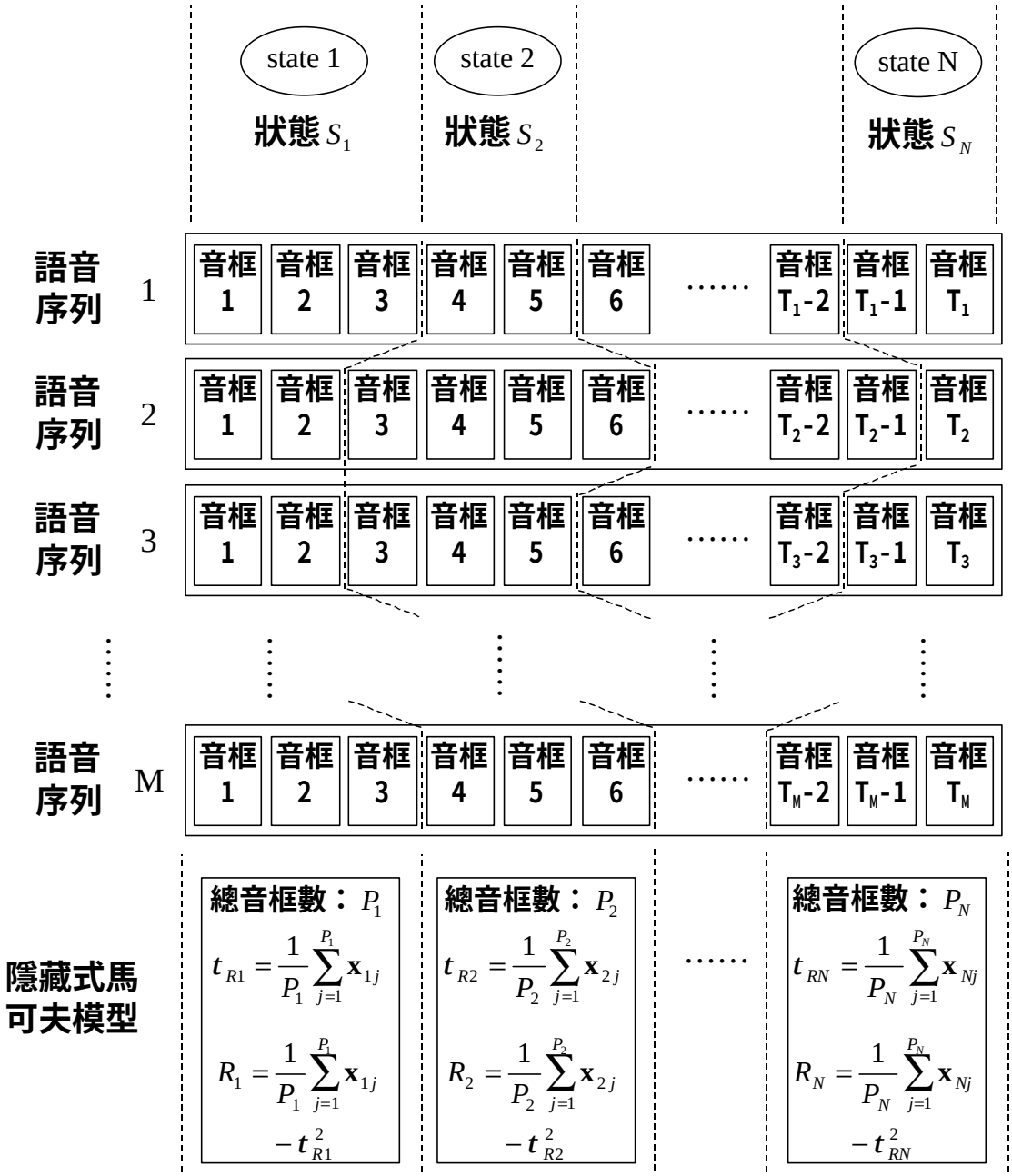


圖 3-1　隱藏式馬可夫模型音框分配系統模型參數圖

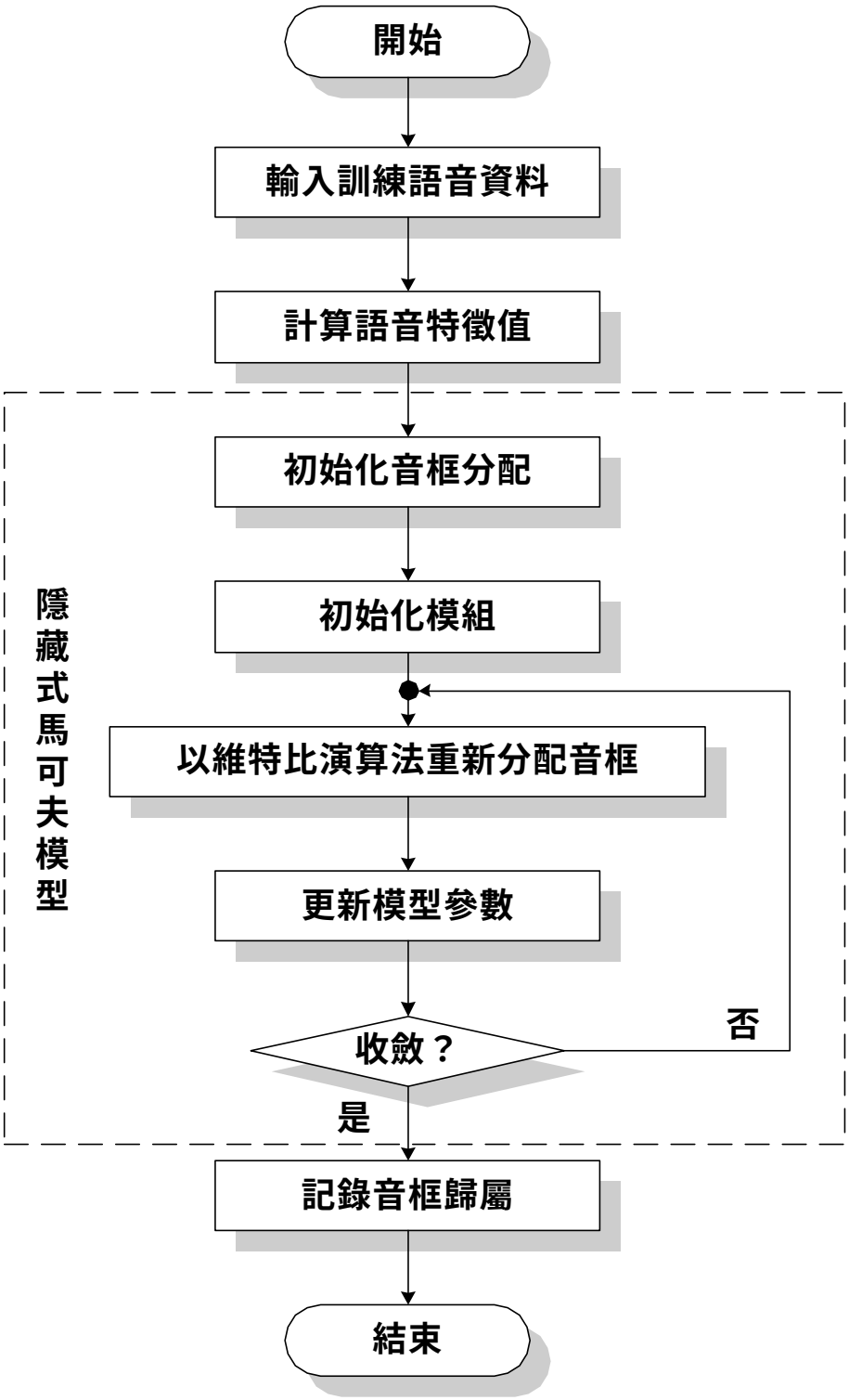


圖 3-2　隱藏式馬可夫模型音框分配流程圖

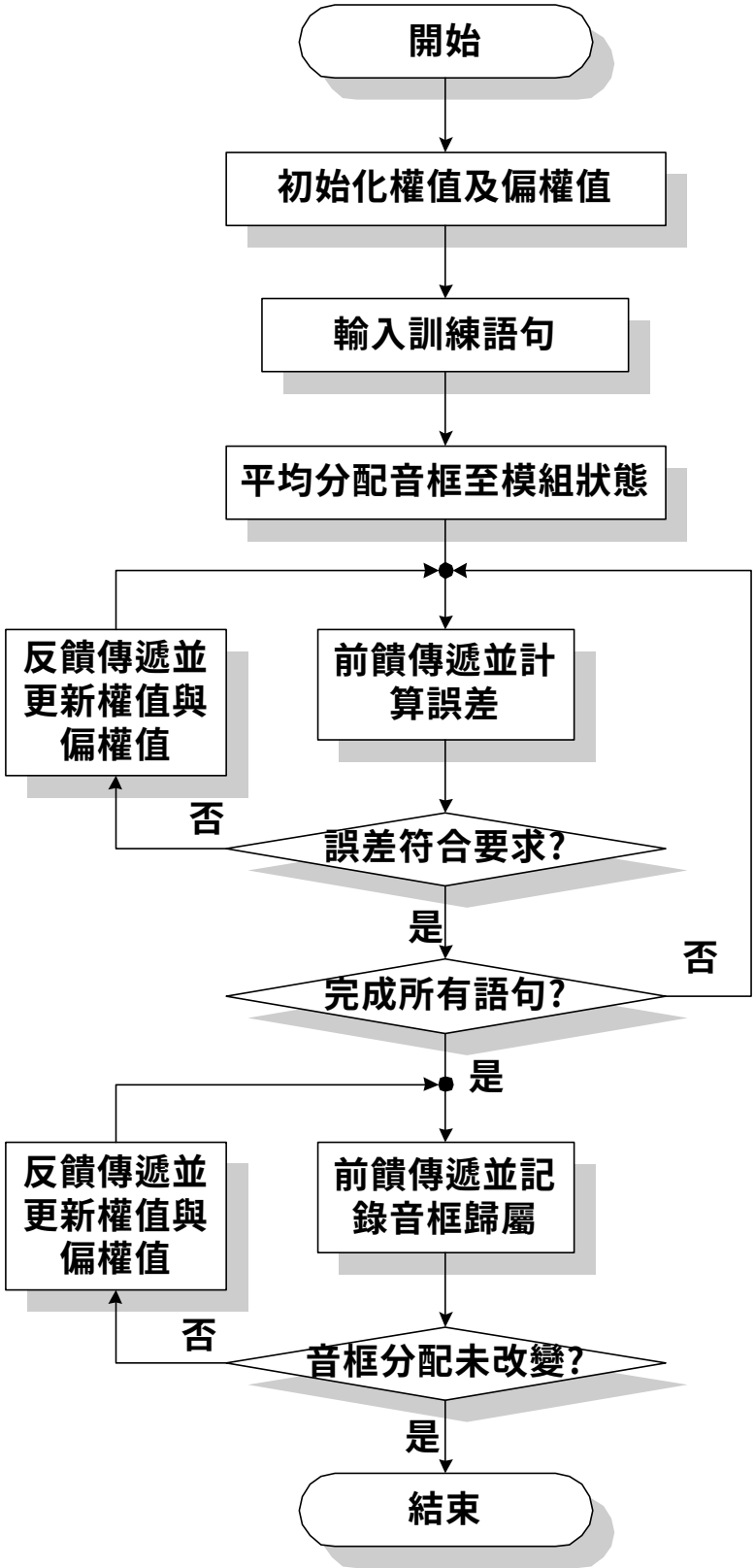


圖 3-3　自我監督類神經網路模型音框分配流程圖

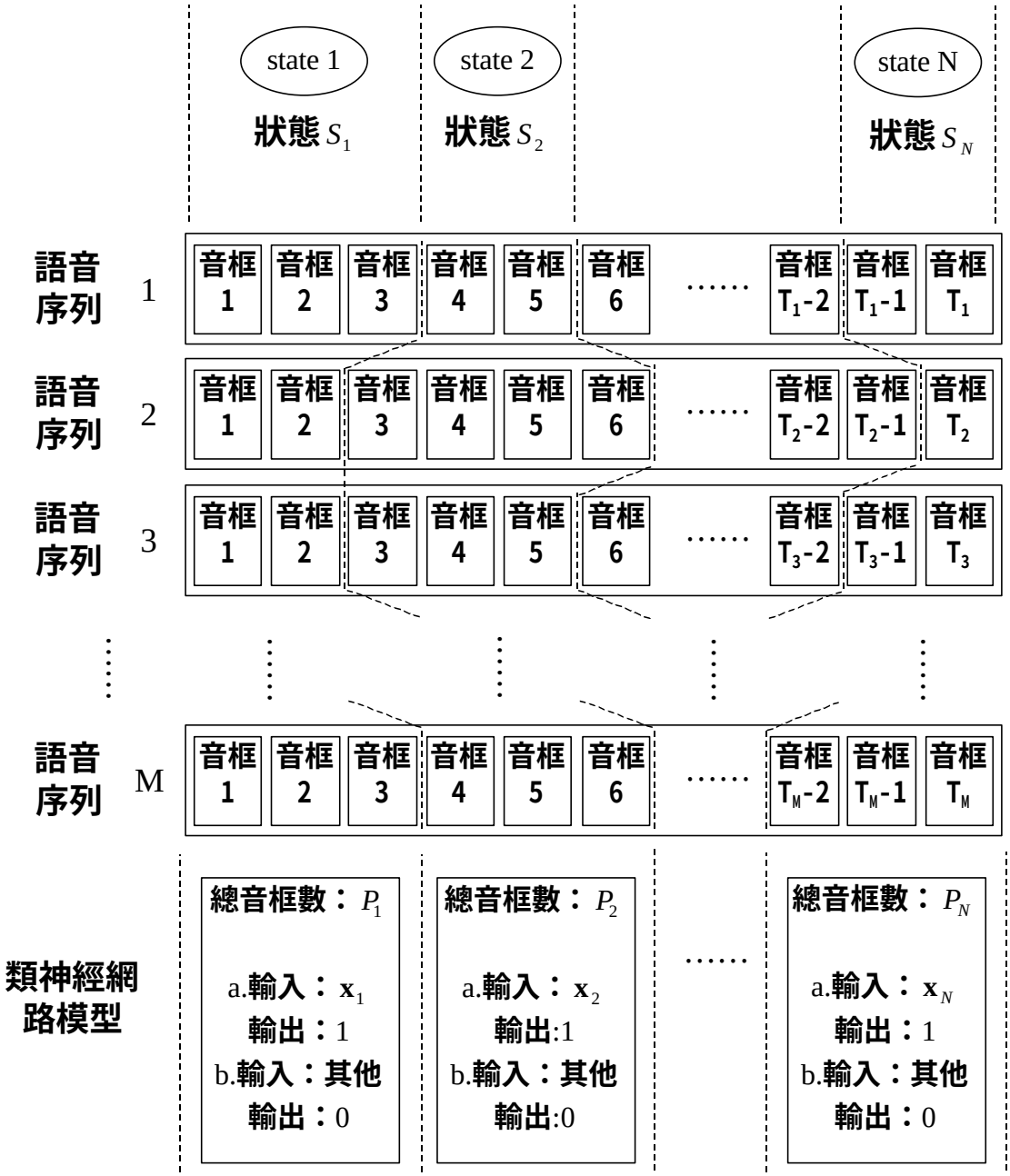


圖 3-4　類神經網路狀態模型網路目標輸出值設定圖

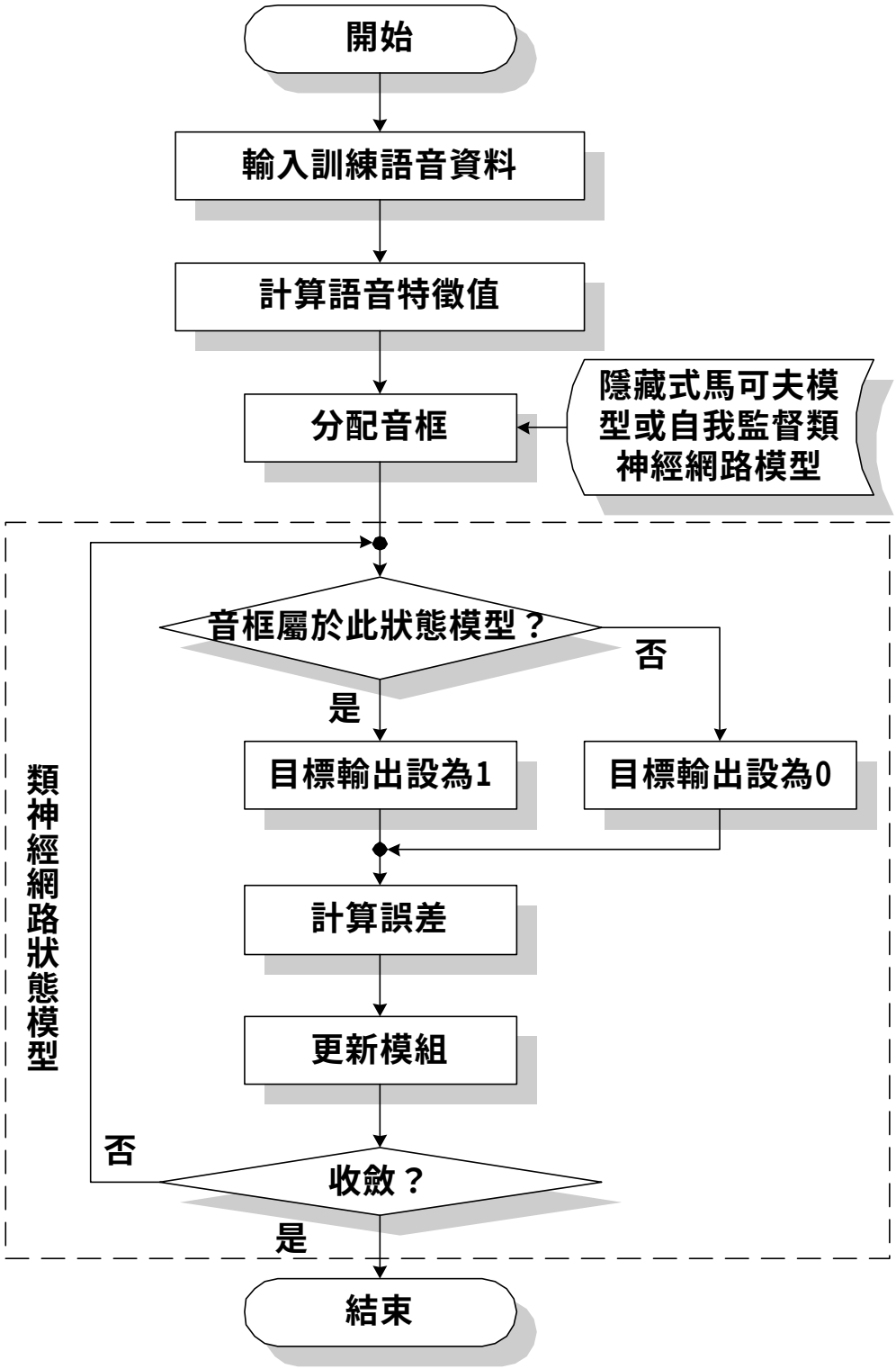


圖 3-5　類神經網路狀態模型訓練流程圖

3.2 模型辨識階段

在語音辨識階段不似模型訓練時複雜，不須對語句做初始化切割，也勿須對模型做疊代訓練。以下就隱藏式馬可夫模型辨識方法及類神經網路狀態模型辨識方法加以說明與比較。

3.2.1 隱藏式馬可夫模型辨識方法

對於測試語句，隱藏式馬可夫模型在挑選最匹配的模組時，乃藉助維特比演算法, 找出所有可能模組的最大可能狀態機率值路徑。假設目前的測試語句共有 T 個音框，首先，強制第一個音框 O_1 屬於第一個狀態 S_1，因此維特比演算法在此位置所記錄的個別機率值是第一個音框在第一個狀態的觀測機率值 $b_1(O_1)$，累積機率值 $d_1(1)$ 相等於 $b_1(O_1)$。倘若維特比最佳路徑中, 第二個音框 O_2 仍屬於第一個狀態，此位置所記錄的個別機率是此音框在第一個狀態的觀測機率值 $b_1(O_2)$，累積機率值 $d_2(1)$ 相等於前一位置的累積機率值 $d_1(1)$ 乘上此位置的個別機率值 $b_1(O_2)$。以此類推，若將最後一個音框 O_T 限制在最後一個狀態 S_N，則最後的累積機率值 $d_T(N)$ 相等於前一位置的累積機率值 $d_{T-1}(N)$ 乘上最後位置的個別機率值 $b_N(O_T)$，如圖 3-6。最後比較所有可能狀態模組累積機率值的大小，可得此測試語句的辨識結果，詳細的系統辨識流程如圖 3-7。

3.2.2 類神經網路狀態模型辨識方法

相較於隱藏式馬可夫模型，類神經網路狀態模型以網路的輸出值做為辨識的依據。假設目前的測試語句共有 T 個音框，第一個音框 O_1 屬於第一個狀態 S_1，維特比演算法在此位置所記錄的個別輸出值是第一個音框在第一個狀態模型的網路輸出值 $Y_1(O_1)$，累積輸出值 $d_1(1)$ 相等於 $Y_1(O_1)$。倘若維特比最佳路徑中，第二個音框 O_2 仍屬於

第一個狀態，此位置所記錄的個別輸出值是此音框在第一個狀態模型的網路輸出值 $Y_1(O_2)$，累積輸出值 $d_2(1)$ 相等於前一位置的累積輸出值 $d_1(1)$ 乘上此位置的個別輸出值 $Y_1(O_2)$。以此類推，則最後的累積輸出值 $d_T(N)$ 相等於前一位置的累積輸出值 $d_{T-1}(N)$ 乘上最後位置的個別輸出值 $b_N(O_T)$，如圖 3-8。最後比較所有可能狀態模組累積輸出值的大小，可得此測試語句的辨識結果。

至於測試語句最佳狀態路徑的求取上，採用兩種方式做比較，第一種方式，使用類神經網路辨識模型結合維特比演算法自行求出，詳細的系統辨識流程如圖 3-9。第二種方式，不採用類神經網路模型自己的最佳路徑，而是先將測試語句放入隱藏式馬可夫模型中，再以隱藏式馬可夫模型中求得之最佳路徑做為依據，詳細的系統辨識流程如圖 3-10。

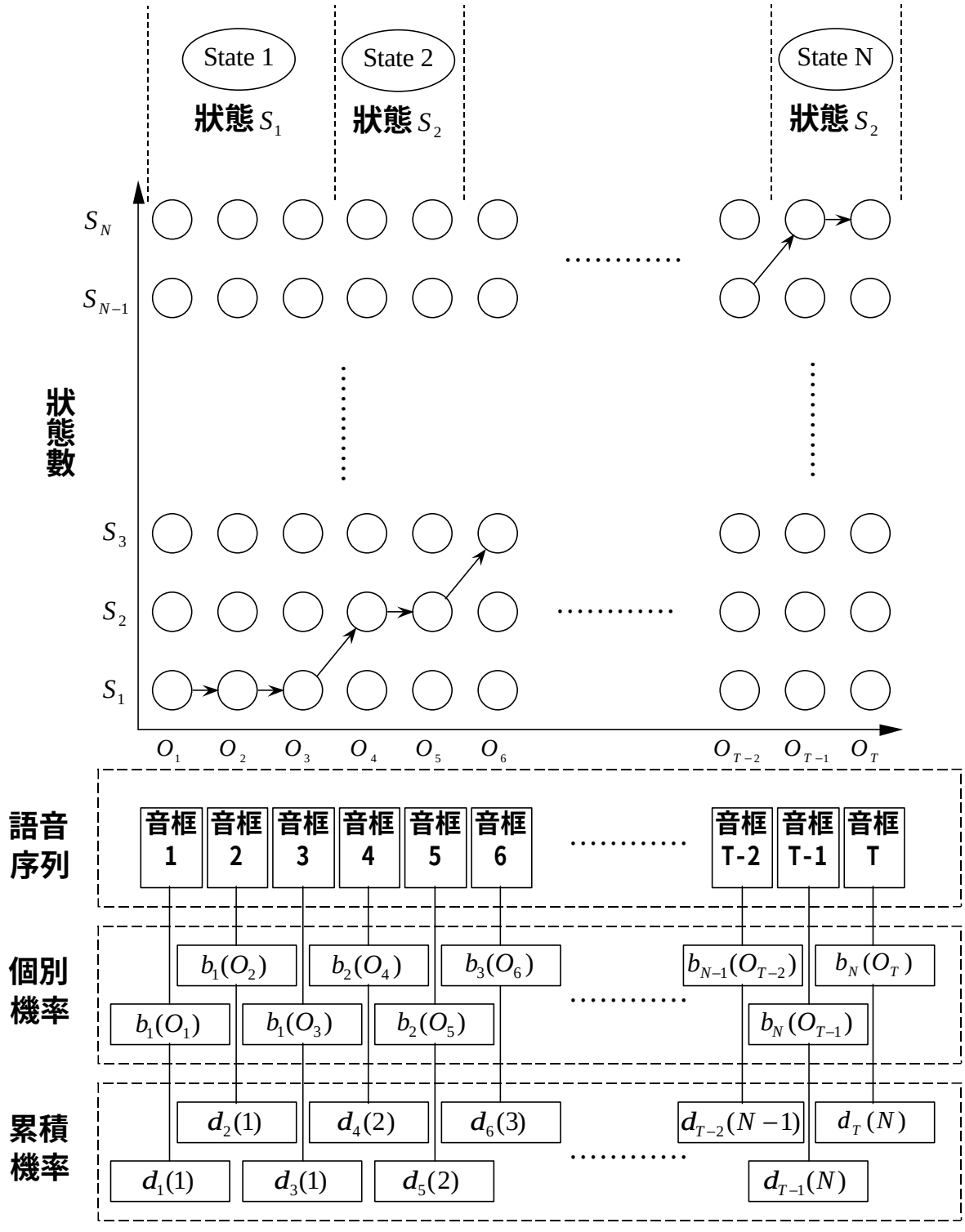


圖 3-6　隱藏式馬可夫模型辨識系統累積機率圖

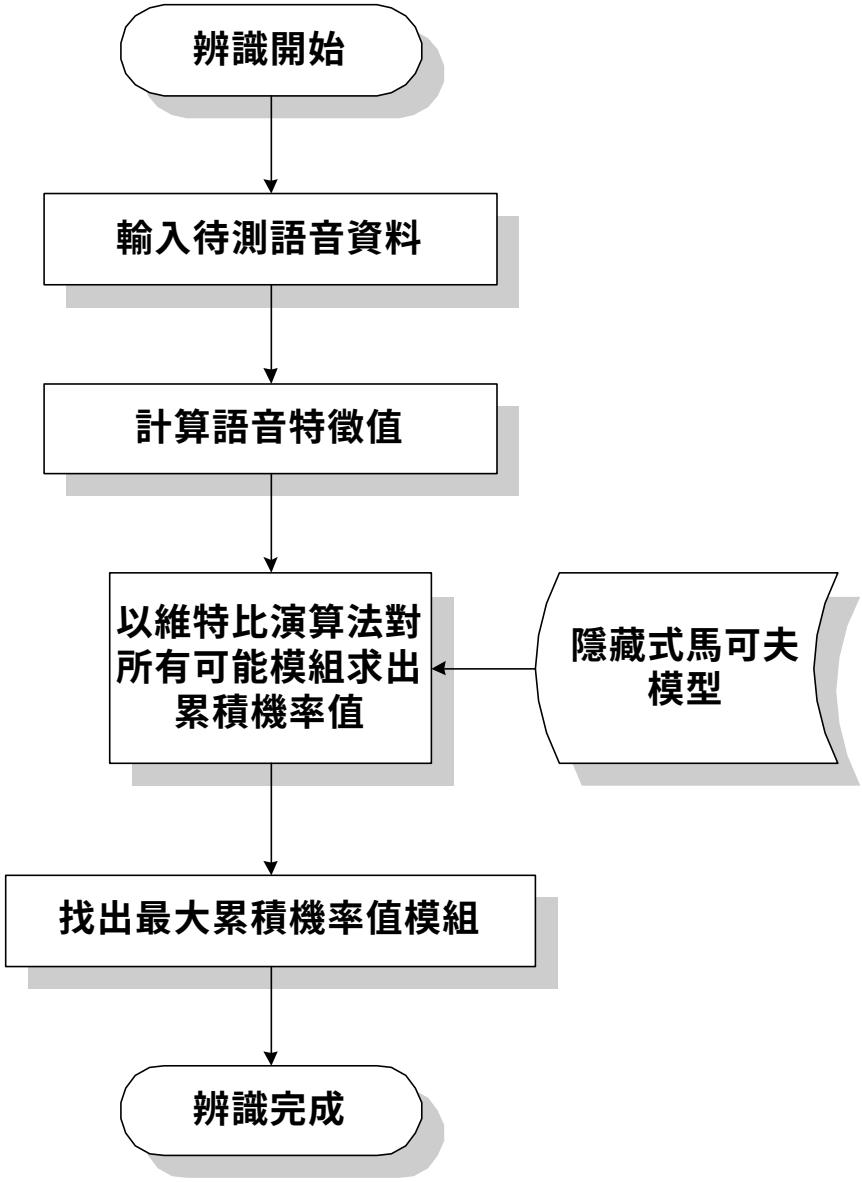


圖 3-7　隱藏式馬可夫模型系統辨識流程圖

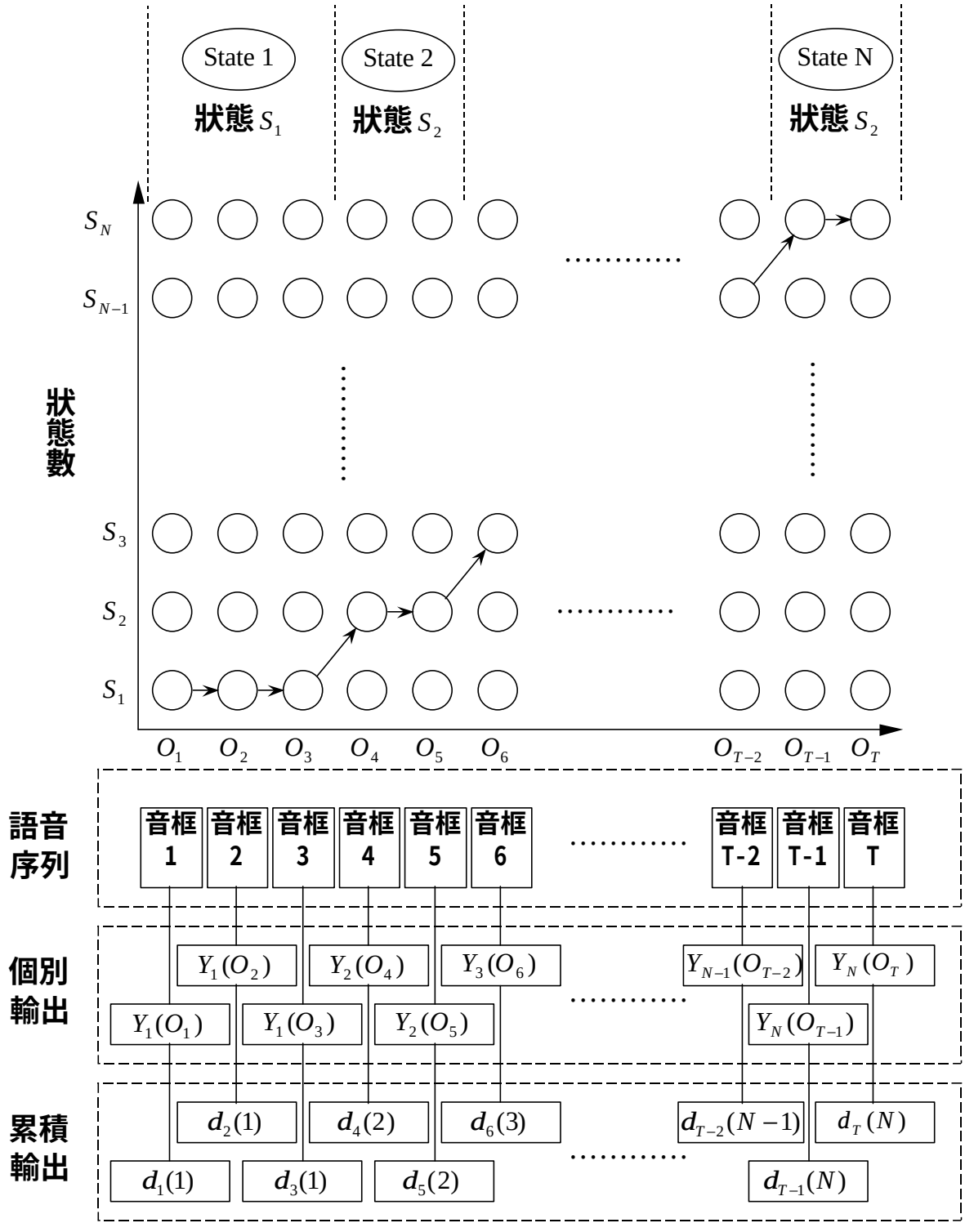


圖 3-8　類神經網路狀態模型辨識系統累積輸出圖

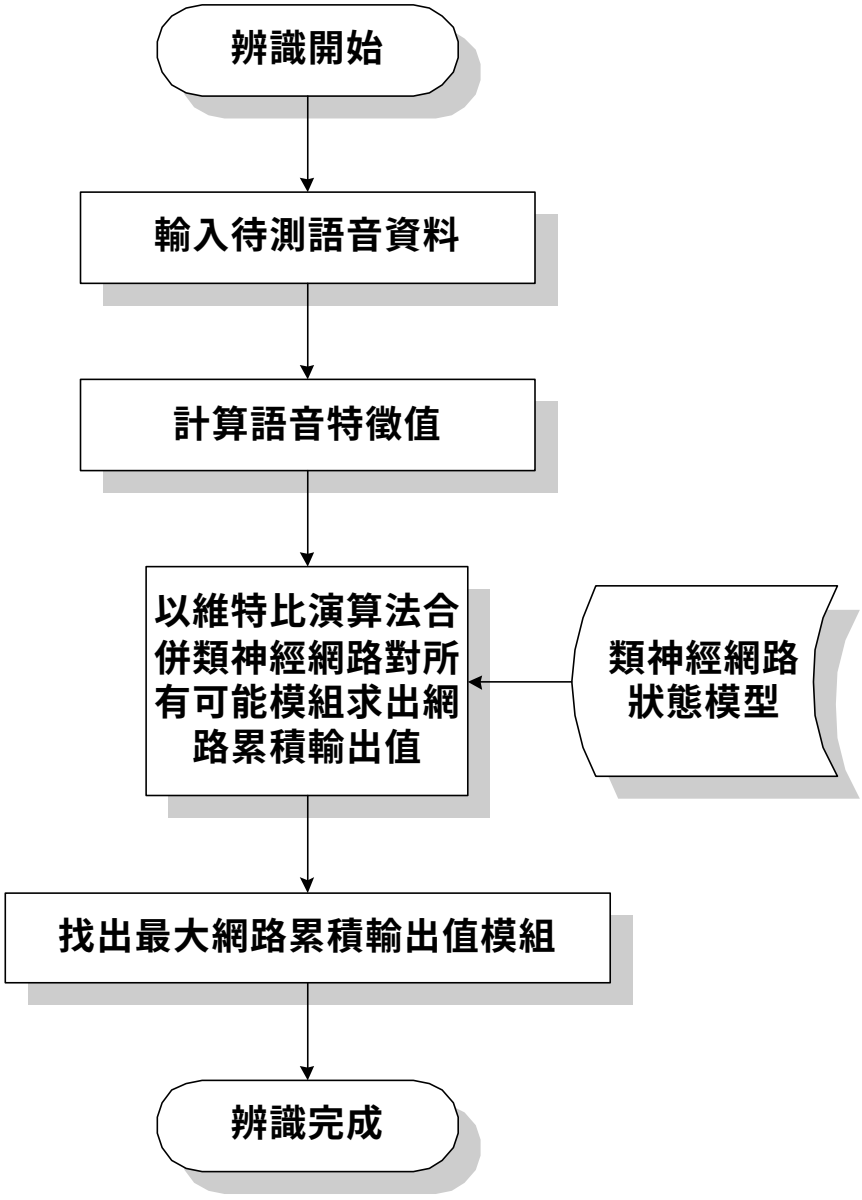


圖 3-9　類神經網路狀態模型系統辨識流程圖（1）

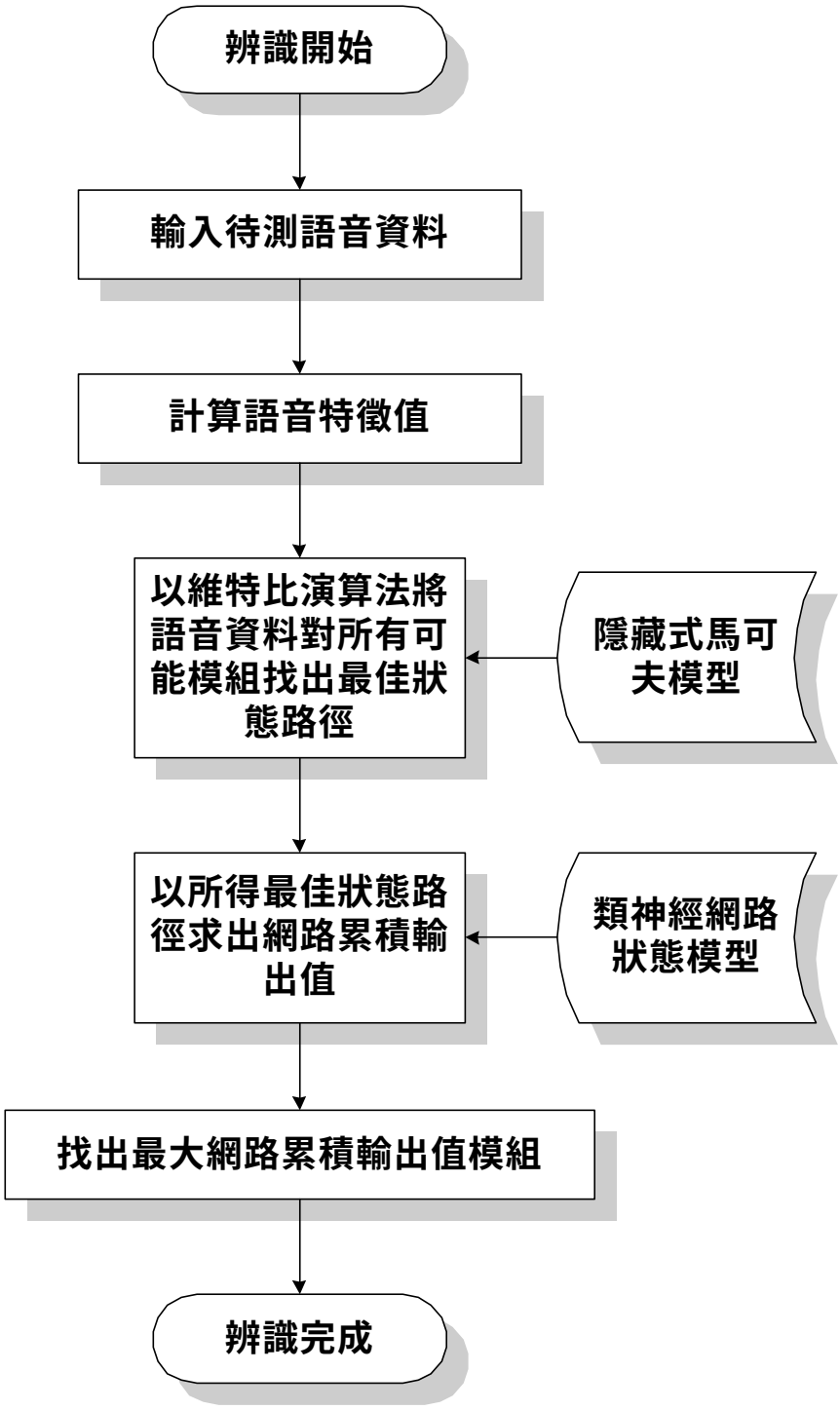


圖 3-10　類神經網路狀態模型系統辨識流程圖（2）

第四章 實驗結果與討論

4.1 系統設定

本篇論文以國語數字“0”到“9”做為辨識模組，由訓練及測試的語音資料來區分，辨識系統可分為語者相關（speaker-dependent）與語者無關（speaker-independent）兩組。語者相關的語音資料是透過麥克風經由聲霸卡直接存取於個人電腦上，語者無關的語音資料則取自 MAT-400 語音資料庫。取樣頻率為 8kHz，資料型態為 16-bit PCM 格式。語音資料以不定音框數切割，音框長度為 30ms（240 個取樣點），音框間重疊為 20ms（160 個取樣點）。語音特徵值為 14 維倒頻譜係數加上 14 維轉移倒頻譜係數。

每一個獨立國語數字之狀態數皆令為 6，因此，10 個國語數字的模型總狀態數為 60。類神經網路模型分為兩類，一類為負責分配音框之自我監督類神經網路模型，輸入層單元數為 28，隱藏層單元數為 50，輸出層單元數為 6，總模型數為 10，另一類則為用以辨識之類神經網路狀態模型，輸入層單元數為 28，隱藏層單元數為 50，輸出層單元數為 1，總模型數為 60。另外，做為分配音框及辨識率比較之隱藏式馬可夫模型，混合數設為 1。

用以測試的類神經網路狀態模型辨識系統如圖 4-1 所示，分為下列三種組合：

系統 1（HMM-NN-Net）：以隱藏式馬可夫模型分配訓練語句音框，建立類神經網路狀態模型，並以狀態模型本身為依據分配測試語句音框，完成辨識。

系統 2（HMM-HMM-Net）：以隱藏式馬可夫模型分配訓練語句音框，建立類神經網路狀態模型，並先行使用隱藏式馬可夫模型分

配測試語句音框，再以類神經網路狀態模型進行辨識。

系統 3（NN-NN-Net）：以自我監督類神經網路模型分配訓練語句音框，建立類神經網路狀態模型，並以狀態模型本身為依據分配測試語句音框，完成辨識。

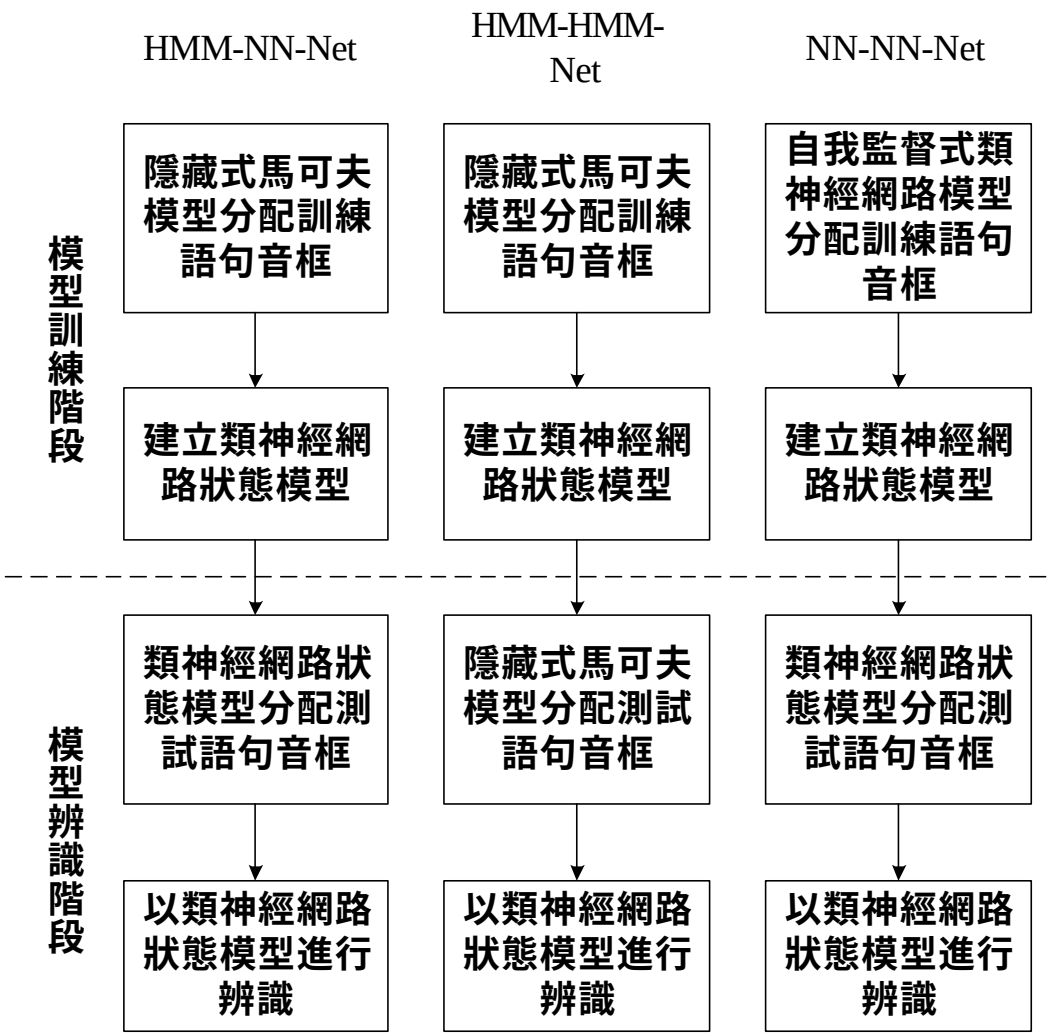


圖 4-1　三種類神經網路狀態模型系統

所有類神經網路狀態模型辨識結果，皆與隱藏式馬可夫模型辨識結果加以對照，比較系統辨識能力，以下分別就語者相關辨識系統與語者無關辨識系統做討論。

4.2 語者相關辨識系統

語者相關辨識系統的語音資料由四位男性所錄製而成，每一位語者針對獨立數字“0”到“9”各錄製 24 遍，其中 4 遍做為訓練語句，另外 20 遍做為測試語句。因此，每個語者相關辨識系統各有 40 句訓練語句以及 200 句測試語句。

首先，類神經網路狀態模型收斂條件定為“實際網路輸出值與目標輸出值間平方誤差小於 0.16”，辨識結果如表 4-1。

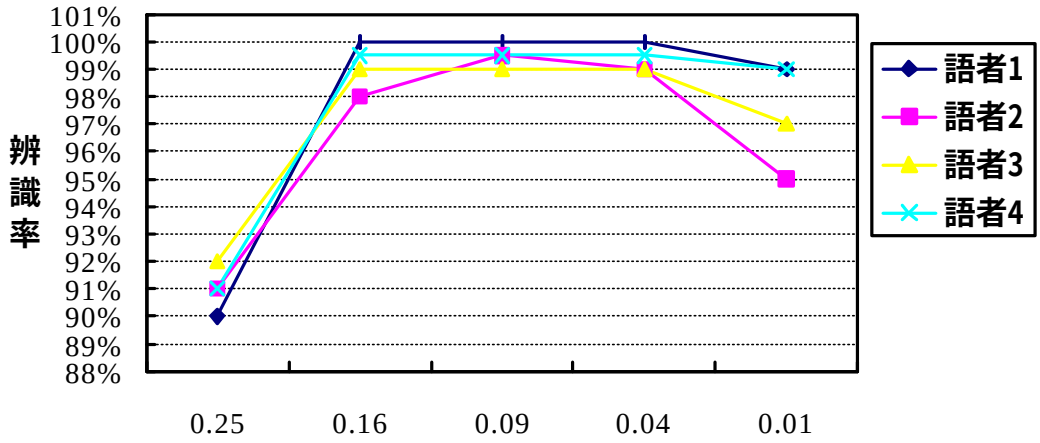
表 4-1　語者相關系統辨識結果

使用模型 \ 語者		語者 1	語者 2	語者 3	語者 4
隱藏式馬可夫模型	正確語句/測試語句	200/200	197/200	198/200	199/200
	正確率	100%	98.5%	99%	99.5%
HMM-NN-Net 狀態模型	正確語句/測試語句	200/200	196/200	198/200	199/200
	正確率	100%	98%	99%	99.5%
HMM-HMM-Net 狀態模型	正確語句/測試語句	190/200	185/200	181/200	192/200
	正確率	95%	92.5%	90.5%	96%
NN-NN-Net 狀態模型	正確語句/測試語句	198/200	194/200	197/200	199/200
	正確率	99%	97%	98.5%	99.5%

由實驗結果比較三種提出的類神經網路狀態模型與傳統隱藏式馬可夫模型，大致上以隱藏式馬可夫模型的辨識率較高，其次依序為 HMM-NN-Net 狀態模型、NN-NN-Net 狀態模型及 HMM-HMM-Net 狀態模型，其中 HMM-NN-Net 狀態模型除了在語者 2 的辨識率上比隱藏式馬可夫模型略低之外，對於其他語者的辨識率皆相同。此外，除了 HMM-HMM-Net 狀態模型對語者 3 的辨識率僅達到 90.5%，其他模型的辨識結果皆相去不遠且接近 100%。顯然地，所提出的三種類神經網路狀態模型，在語者相關辨識系統的應用上，皆可與傳統隱藏式馬可夫模型相匹敵。

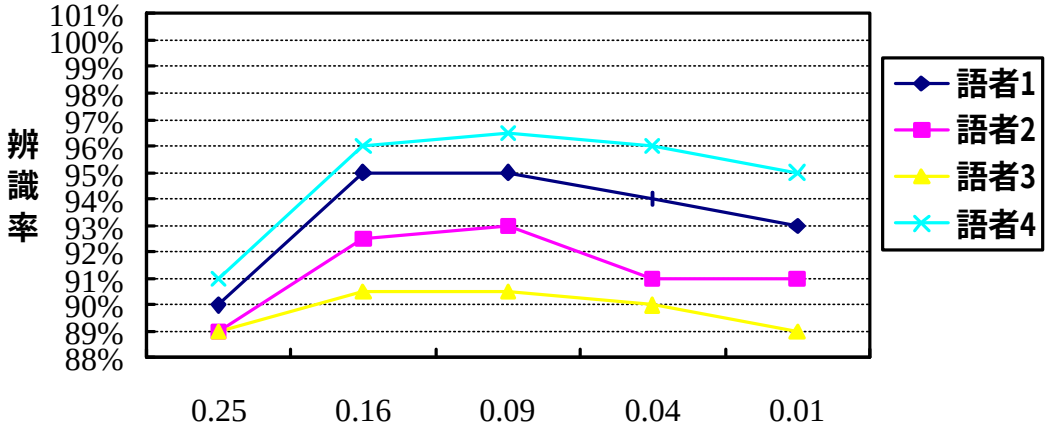
另一方面，不同的類神經網路收斂條件，對網路權值的調整常常有絕對性的影響，因此，適當地選擇收斂條件，將可獲得最佳的辨識模型。圖 4-2 至圖 4-4 列出三種狀態模型在不同收斂條件訓練下的辨識結果，其中實際網路輸出值與目標輸出值間最大平方誤差調整在 0.25 與 0.01 之間。

由實驗結果可以看出，誤差值控制在 0.16 至 0.09 之間有較佳之辨識率，收斂條件過於寬鬆或嚴苛皆不利於辨識系統，主要原因可能在於太寬鬆的收斂條件使得網路訓練被終止在尚未成熟的階段，類神經網路模型強大的分類能力無法完全發揮，相反地，當收斂條件太過嚴格，會造成網路的過度訓練，使得部分較為模糊的特徵，被網路強制區隔開來。此外，若將誤差值定為 0.09，HMM-NN-Net 狀態模型的辨識率可些微地超越隱藏式馬可夫模型。



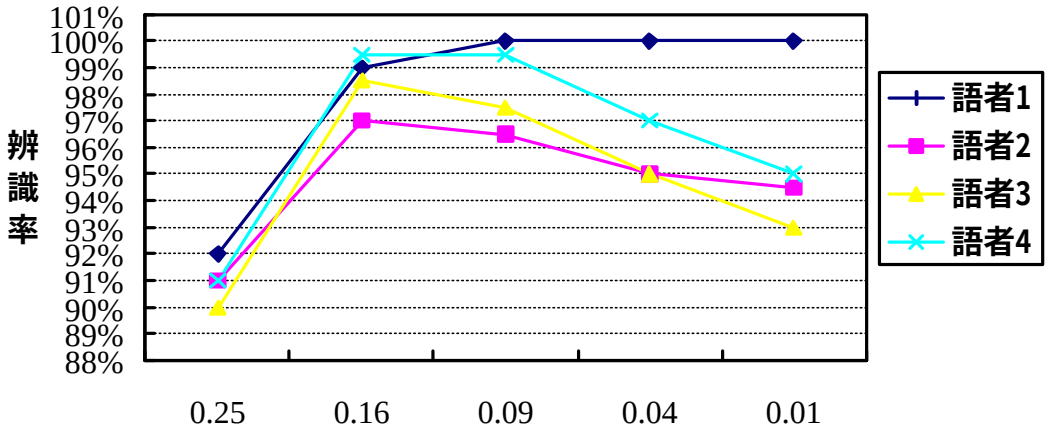
誤差值	0.25	0.16	0.09	0.04	0.01
語者1	90%	100%	100%	100%	99%
語者2	91%	98%	99.5%	99%	95%
語者3	92%	99%	99%	99%	97%
語者4	91%	99.5%	99.5%	99.5%	99%

圖 4-2　HMM-NN-Net 狀態模型在不同收斂條件下的辨識結果



誤差值	0.25	0.16	0.09	0.04	0.01
語者1	90%	95%	95%	94%	93%
語者2	89%	92.5%	93%	91%	91%
語者3	89%	90.5%	90.5%	90%	89%
語者4	91%	96%	96.5%	96%	95%

圖 4-3　HMM-HMM-Net 狀態模型在不同收斂條件下的辨識結果



誤差值	0.25	0.16	0.09	0.04	0.01
語者1	92%	99%	100%	100%	100%
語者2	91%	97%	96.5%	95%	94.5%
語者3	90%	98.5%	97.5%	95%	93%
語者4	91%	99.5%	99.5%	97%	95%

圖 4-4　NN-NN-Net 狀態模型在不同收斂條件下的辨識結果

4.3 語者無關辨識系統

語者無關辨識系統的語音資料包含 200 句訓練語句及 400 句測試語句，由 60 位男性所錄製而成，每一位語者針對獨立數字“0”到“9”各唸一遍，其中 20 位語者負責訓練語句，另外 40 位語者提供測試語句。

由語者相關辨識系統之實驗結果已得知，為了得到較好的辨識結果，類神經網路狀態模型收斂條誤差值須選擇在 0.16 至 0.09 之間，又考慮語者無關辨識系統使用較多且變異較大之訓練資料，誤差值定為 0.16，將可節省網路訓練所須時間。傳統隱藏式馬可夫模型及三種類神經網路狀態模型針對語者無關系統的辨識結果列於表 4-2。

表 4-2　語者無關系統辨識結果

辨識率 / 使用模型	正確語句/測試語句	正確率
隱藏式馬可夫模型	368/400	92%
HMM-NN-Net 狀態模型	377/400	94.25%
HMM-HMM-Net 狀態模型	347/400	86.75%
NN-NN-Net 狀態模型	350/400	87.5%

相較於語者相關辨識系統，不同模型在語者無關辨識系統中的所獲得的實驗結果，有較明顯的差異。HMM-NN-Net 狀態模型的辨識率為 94.25%，為四種模型中最高者，其次依序為隱藏式馬可夫模型、NN-NN-Net 狀態模型及 HMM-HMM-Net 狀態模型。綜合語者相關與語者無關辨識系統兩者的實驗結果可進一步得知，若使用正確的網路收斂條件，HMM-NN-Net 狀態模型在所有實驗的模型中有最佳的辨識率。此外，比較三種類神經網路狀態模型的辨識結果，可得兩項推論：

（1）在網路訓練階段，若藉由隱藏式馬可夫模型分配語句音框，可獲得較佳之狀態模型。

（2）在網路辨識階段，使用狀態模型自身的音框分配，有較好的區分能力。

類神經網路雖然有眾多優點，但是網路複雜的收斂問題，卻往往為使用者帶來很大的困擾，表 4-3、表 4-4 列出 HMM-NN-Net 狀態模型與 NN-NN-Net 狀態模型在網路訓練階段，各模型於相同的收

表 4-3　HMM-NN-Net **狀態模型疊代次數**

	狀態 1	**狀態** 2	**狀態** 3	**狀態** 4	**狀態** 5	**狀態** 6
數字“0”	751	1214	10129	3697	818	469
數字“1”	275	1506	2424	4036	1320	725
數字“2”	371	1829	3966	2124	1569	607
數字“3”	733	1340	929	1559	958	730
數字“4”	558	1550	1781	1710	1179	619
數字“5”	206	12238	2776	1359	1551	587
數字“6”	711	727	556	1069	1945	853
數字“7”	543	1612	677	1826	1638	1107
數字“8”	450	1051	3115	5224	11065	576
數字“9”	365	487	958	1688	1509	699
總合	110644					

表 4-4　NN-NN-Net **狀態模型疊代次數**

	狀態 1	**狀態** 2	**狀態** 3	**狀態** 4	**狀態** 5	**狀態** 6
數字“0”	746	880	3490	1724	911	849
數字“1”	306	732	5316	1088	563	337
數字“2”	358	1439	2158	1193	777	259
數字“3”	1730	2510	1725	847	1008	674
數字“4”	535	1659	2059	2056	580	422
數字“5”	68	647	834	1223	960	540
數字“6”	428	859	828	530	2572	828
數字“7”	777	1404	463	2102	760	577
數字“8”	171	278	1001	1200	1423	346
數字“9”	1280	703	706	1098	1601	1077
總合	66215					

斂條件及學習速率（ h ）下所須的疊代次數。

由表 4-3 及表 4-4 中可看出，使用自我監督類神經網路模型為訓練語句音框分框依據的 NN-NN-Net 狀態模型比使用隱藏式馬可夫模型為依據的 HMM-NN-Net 狀態模型擁有更快的收斂速率，合計所有模型的疊代次數，NN-NN-Net 狀態模型約比 HMM-NN-Net 狀態模型減少了 40%的訓練時間。兩者間差異造成的原因主要在於 NN-NN-Net 狀態模型本身的網路架構與所使用的自我監督類神經網路模型是相同的，因此，經由自我監督類神經網路模型歸類在同一狀態內的音框，對自己所屬的 NN-NN-Net 狀態模型也有較高的依存度，NN-NN-Net 狀態模型在網路訓練時，可輕易地區隔屬於不同狀態的音框。相對於 NN-NN-Net 狀態模型，HMM-NN-Net 狀態模型雖然使用了高斯機率分佈函數加以有效的分配訓練語句音框，但屬於不同狀態音框間的特性卻無法被網路模型快速地分辨出來，因此必須花費較長的時間訓練辨識模型。

第五章 結論與未來展望

5.1 結論

如何提升系統的識別能力，仍是目前語音辨識研究領域的主流，本篇論文的主要內容在於將隱藏式馬可夫模型中不可觀測的狀態觀念引入類神經網路模型，建立類神經網路狀態模型辨識系統，企圖藉由隱藏式馬可夫模型與類神經網路模型的結合，發揮兩者的優點，降低辨識系統的錯誤率。經由前一章的實驗結果，證明了所提方法在實際應用上的可行性，並且能確實達到預期的要求。

首先，在語者相關辨識系統方面，由於傳統隱藏式馬可夫模型的辨識率已趨近於百分之百，無法再期待所提的三種類神經網路狀態模型在辨識率上有明顯的改善，但將三種網路模型與隱藏式馬可夫模型相互比較，已約略可以看出，使用隱藏式馬可夫模型分配訓練語句音框並以狀態模型本身為依據分配測試語句音框的 HMM-NN-Net 狀態模型有較佳的辨識能力。此外，經由實驗也證實了網路收斂條件與辨識率間的相關性，太過於寬鬆或嚴謹的條件，皆會因不同的原因造成辨識率的下降。

在語者無關辨識系統方面，所有模型的辨識率皆不及語者相關辨識系統，但另一方面也突顯了 HMM-NN-Net 狀態模型在識別能力上的優勢。隱藏式馬可夫模型雖然在語者相關辨識系統上保有不錯的辨識率，但是在語音特徵較模糊的語者無關辨識系統上，便無法再與類神經網路模型相匹敵。針對狀態模型辨識率的提升，將所提出的三種類神經網路狀態模型辨識系統相互比較，可進一步得知，藉由隱藏式馬可夫模型於網路訓練階段分配語句音框，可獲得較佳之狀態模型。同時，使用狀態模型自身的音框分配做為網路辨識階

段的依據，有較好的區分能力。

此外，完全使用類神經網路架構進行訓練與辨識工作的 NN-NN-Net 狀態模型，雖然在辨識能力上不及 HMM-NN-Net 狀態模型，但可減少許多網路訓練所須時間。

5.2 未來展望

本篇論文以國語數字“0”到“9”的測試，證實了所提類神經網路狀態模型是一種可有效提升辨識率的新方法。由於類神經網路狀態模型改善了傳統隱藏式馬可夫模型與類神經網路模型在語音辨識上的缺點，在實際的應用上將更富多樣性。未來的研究方向，將延伸辨識模型至國語單音（isolated word）、音節（syllable）及次音節（sub-syllable）的建立，使模型更趨完備，進一步再完成連續語音及線上辨識系統。此外，在辨識率與網路收斂時間的兩重考量下，如何求得最佳化的網路收斂條件，也是值得重視的研究目標。

參考文獻

〔1〕 L. E. Baum and T. Tetrie, “Statistical Inference for Probabilistic Functions of Finite State Markov Chains,” Ann. Math. Stat., Vol. 37, pp. 1554-1563, 1966.

〔2〕 L. E. Baum and J. A. Egon, “An Inequality with Applications to Statistical Estimation for Probabilistic Functions of A Markov Process and to A Model for Ecology,” Bull. Amer. Meteorol. Soc., Vol. 73, pp. 360-363, 1967.

〔3〕 L. E. Baum and G. R. Sell, “Growth Functions for Transformations on Manifolds,” Pac. J. Math., Vol. 27, No.2, pp. 211-227, 1968.

〔4〕 L. E. Baum, T. Petrie, G. Soules, and N Weiss, “A Maximization Technique Occurring in The Statistical Analysis of Probabilistic Functions of Markov Chains,” Ann. Math. Stat., Vol. 41, No. 1, pp. 164-171, 1970.

〔5〕 L. E. Baum, “An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes,” Inequalities, Vol. 3, pp. 1-8, 1972.

〔6〕 J. K. Baker, “The Dragon System-An Overview,” IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 23, No. 1, pp. 24-29, Feb. 1975.

〔7〕 F. Jelinek, “A Fast Sequential Decoding Algorithm Using A Stack,” IBM J. Res. Develop., Vol. 13, pp. 675-685,1969.

〔8〕 L. R. Bahl and F. Jelinek, “Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition,” IEEE Trans. on Information Theory, Vol. 21, pp.

404-411, 1975.

〔9〕 F. Jelinek, L. R. Bahl, and R. L. Mercer, “Design of A Linguistic Statistical Decoder for The Recognition of Continuous Speech,” IEEE Trans. on Information Theory, Vol. 21, pp. 250-256, 1975.

〔10〕 F. Jelinek, “Continuous Speech Recognition by Statistical Methods,” Proc. IEEE, Vol. 64, pp. 532-536, Apr. 1976.

〔11〕 R. Bakis, “Continuous Speech Word Recognition via Centi-second Acoustic States,” in Proc. ASA Meeting (Washington DC), Apr. 1976.

〔12〕 F. Jelinek, L. R. Bahl, and R. L. Mercer, “Continuous Speech Recognition: Statistical Methods,” in Handbook of statistics, II, P. R. Krishnaiad, Ed. Amsterdam, The Netherlands: North-Holland, 1982.

〔13〕 L. R. Bahl, F. Jelinek, and R. L. Mercer, “A Maximum Likelihood Approach to Continuous Speech Recognition,” IEEE Trans. on Pattern Analysis and Machine Intelligence., Vol. 5, pp. 179-190, 1983.

〔14〕 L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proc. IEEE, Vol. 77, No.2, pp. 257-286, Feb. 1989.

〔15〕 K. J. Lang, Alex H. Waibel and G. E. Hinton, “A Time-Delay Neural Network Architecture for Isolated Word Recognition,” Neural Networks, Vol. 3, pp. 23-43, 1990.

〔16〕 A. Bendiksen and K. Steiglitz, “Neural Networks for Voiced/Unvoiced Speech Classification,” IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, No. 90, pp. 521-524, 1990.

〔17〕T. Ghiselli-Crippa, A. El-Jaroudi, “A Fast Neural Net Training Algorithm and Its Application to Voiced-Unvoiced-Silence Classification of Speech,” IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, No. 91, pp. 441-444, 1991.

〔18〕Y. Qi and B. R, Hunt, “Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and A Network Classifier,” IEEE Trans. on Speech and Audio Processing, Vol. 1, No. 2, pp. 250-255, Apr. 1993.

〔19〕G. Kuhn, R. L. Watrous and B. Ladendorf, “Connected Recognition with A Recurrent Network,” Speech Communication, Vol. 9, No. 1, pp. 41-48, Feb. 1990.

〔20〕S. J. Lee, K. C. Kim, H. Yoon and J. W. Cho, “Application of Fully Recurrent Neural Networks for Speech Recognition,” Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, pp. 77-80, 1991.

〔21〕A. Hunt, “Recurrent Neural Networks for Syllabification,” Speech Communication, Vol. 13, pp. 323-332, 1993.

〔22〕T. Lee, P. C. Ching and L. W. Chan, “Recurrent Neural Networks for Speech Modeling and Speech Recognition,” Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 5, pp. 3319-3322, 1995.

〔23〕W.-Y. Chen, Y.-F. Liao and S.-H. Chen, “Speech Recognition with Hierarchical Recurrent Neural Networks,” Pattern Recognition, Vol. 28, No. 6, pp. 795-805, 1995.

〔24〕H. Bourlard and C. j. Wellekens, “Links between Markov Models and Multilayer Perceptrons,” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 12, No. 12, pp. 1167-1178, Dec. 1990.