行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　※
※　　　　可靠之基頻軌跡偵測及聲調辨認　　　　※
※　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：☑個別型計畫　　□整合型計畫

計畫編號：NSC90 － 2213 － E － 009 － 111

執行期間： 90 年 8 月 1 日至 91 年 7 月 31 日

計畫主持人：王逸如

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：國立交通大學電信工程系

中　華　民　國　91 年　10 月　30 日

行政院國家科學委員會專題研究計畫成果報告

可靠之基頻軌跡偵測及聲調辨認

計畫編號：NSC90－2213－E－009－111
執行期限：90 年 8 月 1 日至 91 年 7 月 30 日
主持人：王逸如　國立交通大學電信工程系
計畫參與人員：　呂儲仰、鐘祥睿、魯弘茂

一、中文摘要

本計畫中針對國語連續語音提出一個利用統計式之基週軌跡偵測方法，並製作一個國語聲調辨認器。在統計式之基週軌跡偵測方法中，在傳統利用語音信號之自相關係數來偵測基頻的方法中，對每一音框產生數個基頻候選值，適當的將語音信號中各音框為有聲/無聲音之判別由決定值轉(deterministic)換為機率量測值(probabilistic)，並訂定音框間之機頻轉移機率模型，則基週軌跡偵測將可視為一個尋找最大相似度(Maximum Likelihood)路徑問題,可利用 Viterbi 搜尋方法來尋找一條最佳基頻軌跡。上述方法經使用 MAT2500 語料庫實驗證實效能較 ESPS 中之基週軌跡偵測器為佳。其次，使用類神經網路來製作一個國語聲調辨認器，使用上述統計式之基週軌跡偵測方法所獲得之機頻軌跡來作連續國語語音之聲調辨認，可獲的 77%的辨認率。

關鍵詞：基週軌跡偵測、聲調辨認、軌跡追蹤、機率量測值

Abstract

In this project, reliable pitch detection and tone recognition was studied. First, a new statistical pitch contour tracking algorithm is employed to improve the conventional auto-correlation based pitch detection method. First, several pitch candidates were leaved in each frame. Then, the voiced/unvoiced classification of frames, the inter-frame pitch transition, and the inter-segment pitch jump were model by statistical models. The pitch contour tracking problem becomes to find a pitch contour with maximum likelihood value from all possible candidates. And, the optimal pitch contour can be performed by the Viterbi search with properly defined the probabilistic measure function. By experiments using MAT2500 speech database, the performance of proposed statistic pitch detector is proved better than the pitch detector in ESPS package. In the second part of this project, the neural network based tone recognition method proposed previously was established. The pitch contour found by the proposed pitch detector was used, in order to improve the tone recognition rate. A tone recognition rate of 77% was achieved.

Keywords: pitch detection, tone recognition, contour tracking, probabilistic measure

二、 緣由與目的

基週軌跡在國語是聲調辨認中所需使用的重要參數，雖然過去有許多基週偵測方法被提出，但其效能仍有待改善，尤其二倍頻及半倍頻的錯誤機率頗高，所以在此計畫中將提出更好的基週軌跡偵測方法，必進行聲調辨認器之製作。

三、研究方法

1. 統計法之基週期偵測器

本計畫中所提出的基週期軌跡求取演算法，主要將每個音框留下數個基頻候選值，再將：(1)各因框之 U/V 判別，(2)基頻或區段間基頻轉移建立機率模式，再以 maximum likelihood 的要求來尋找最佳基頻軌跡。

首先，使用改良式的自相關係數(autocorrelation coefficient)作為特徵參數之一，特徵參數還包括通過 BPF 的能量($E(n)$)以及自相關係數的平方和($\sum_{n=n_1}^{n_2} \rho^2(n)$)，及自相關係數對應的頻率值($f_i$)，每一個音框選擇數個

自相關係數之區域最大值作為基頻候選者，給定如下的特徵向量

$$\bar{x}_i(n)=\left[E(n),\ \sum_{i=n1}^{n2}{}^{\circ 2}{}_n(i),\ {}^{\circ}{}_{f_i}(n),\ f_i(n)\right], i=1\sim 6$$

根據 U/V 特性的不同，如果音框為 Unvoiced 時，為了搜尋時的方便，使用以此音框中的自相關函數最大值所對應之頻率值（$f_i(n)=\arg\max_i {}^{\circ}{}_i(n)$），我們將其稱為虛擬頻率。特徵參數表示為

$$\bar{x}_{unvoiced}(n)=\left[E(n),\ \sum_{i=n1}^{n2}{}^{\circ 2}{}_n(i),\ {}^{\circ}{}_{f_v(n)}(n),\ f_v(n)\right]$$

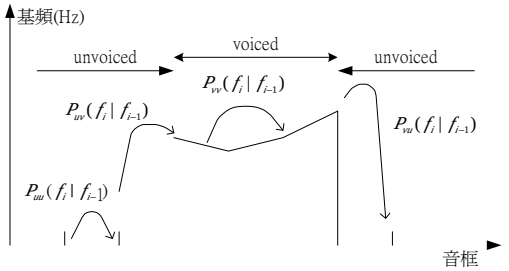

圖 1：基頻狀態移轉機率圖。

基於每個音框間的基頻移轉為一階 Markov Process 的假設，我們總共要建立 $P_{vv}(f_i(n)|f_j(n-1))$ 、 $P_{vu}(f_i(n)|f_j(n-1))$ 、 $P_{uv}(f_i(n)|f_j(n-1))$ 及 $P_{uu}(f_i(n)|f_j(n-1))$ 四種基頻轉移機率模型。我們首先建立 Voiced to Voiced 的基頻轉換機率，

$$P_{vv}(f_i(n)|f_j(n-1))=\frac{P(f_i(n),f_j(n-1))}{p(f_j(n-1))},\text{frame } n-1, n\in \textit{voiced}$$

我們將 $P_{vv}(f_i|f_{i-1})$ 假設為混和高斯機率分佈，則可以得到

$$P(f_i(n))=\sum_k \frac{w_k}{\sqrt{2\pi}\pm_{k,f_i(n)}}\exp\left(-\frac{1}{2}\frac{(f_i(n)-\mu_k)^2}{\pm_{k,f_i(n)}^2}\right)$$

$$P(f_i(n),f_j(n-1))$$

$$=\sum_k \frac{w_k}{(2\pi)^{L/2}|C_k|^{1/2}}\exp\left(-\frac{1}{2}(\bar{f}-\bar{\mu}_k)^T(C_k)^{-1}(\bar{f}-\bar{\mu}_k)\right)$$

其中 $\bar{f}=\begin{bmatrix} f_i(n) \\ f_j(n-1)\end{bmatrix}$ 。

當起始頻率 $f_i(n-1)$ 不同時，其轉移到下一個頻率 $f_i(n)$ 的範圍與機率皆不同，因此對 $f_i(n-1)$ 作變數變換。令 $f_i^{'}(n)=\frac{(f_i(n)-\mu_{f(n)|f(n-1)})}{\pm_{f(n)|f(n-1)}}$，原來的轉移機率就可以改寫成

$$P(f_i^{'}(n)|f_j(n-1))=\pm_{f(n)|f(n-1)}\cdot P(f_i(n)|f_j(n-1))$$

其中

$$\pm_{f(n)|f(n-1)}=\left[E\{f(n)^2|f(n-1)\}-E^2\{f(n)|f(n-1)\}\right]^{1/2}$$

原來因為起始條件頻率 $f_i(n)$ 不同所造成不平均的機率分佈，在經過變數變換後將可獲得改善。

若音框不屬於 Voiced 部分時，就沒有所謂的『基頻值』，我們便以『虛擬頻率』作為建立前後音框間基頻值移轉的機率模型。基頻值移轉機率可以表示成

$$P_{UU}(f_i(n)|f_j(n-1))=\frac{P_{UU}(f_i(n),f_j(n-1))}{P_U(f_i(n))},\text{frame } n-1, n\in \text{unvoiced}$$

我們也使用求取 Voiced to Voiced 的基頻轉換機率時所用的標準化因子。 $P_{vu}(f_i'(n)|f_j(n-1))$ 與 $P_{uv}(f_i'(n)|f_j(n-1))$ 的機率模型也以此法建立。另外，定義每個基頻區段間的基頻變化機率為

$$P_{seg\_jump}(F_{i+1,start}|F_{i,end},D_i)=\frac{P(F_{i+1,start},F_{i,end},D_i)}{P(F_{i,end},D_i)}$$ 。

以兩個基頻區段為統計單位，第 i 個基頻區段的最後一個音框基頻值為 $F_{i,end}$，經過 D_i 個 Unvoiced 音框後，下一個基頻區段的第一個音框基頻值為 $F_{i+1,start}$。對一語句中基頻值的連接採取維特比搜尋(Viterbi Search)來尋找最佳路徑，其最佳路徑之累積機率值為：

$$P(f_{i^{(c)}(1)},f_{i^{(c)}(2)},\ldots,f_{i^{(c)}(m-1)},f_c)$$

$$=\max_{\forall i(n),\ n=1,\ldots,m} P(f_{i^{(K^*)}(1)},f_{i^{(K^*)}(2)},\ldots,f_{i^{(K^*)}(m-1)},f_c)$$

2. 中文連續語音聲調的辨認

使用前述統計方法求取一準確性較高的基頻軌跡，對國語連續語音做聲調辨認。方法上，採用『多層神經元』(Multi-Layer Perceptron, MLP)為主體的辨認器。 音節長度少於三個音框之音節視為刪除型錯誤，不予辨認。考慮音節間耦合的問題，使用前後文相關特徵參數。

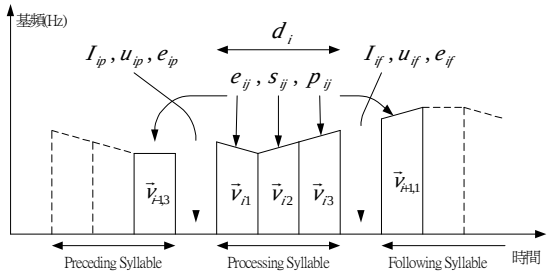

圖 4：特徵參數抽取示意圖。

其中的元素 e_{ij}, s_{ij}, p_{ij} 代表第 i 個音節中第 j 段的能量對數平均值、基頻軌跡的斜率(Slope)以及相對應的截距(Intercept)。d_i 代表第 i 個音節內基頻軌跡的長度。I_{ip} 用來表示前面是否有相鄰音節的二元指標，$\bar{v}_{i-1}$ 表示前一個音節的最後一段特徵向量，u_{ip} 是目前音節與前一個音節間隔的 unvoiced 區間長度，e_{ip} 為此 unvoiced 區間的能量對數平均值；I_{if} 用來表示後面是否有相鄰音節的二元指標，$\bar{v}_{i+1}$ 表示後一個音節的第一段特徵向量，u_{if} 是目前音節與下一個音

節間隔的 unvoiced 區間長度，e_{if} 為此 unvoiced 區間的能量對數平均值。

3. 統計法之基週期偵測器實驗結果與分析

使用「台灣之國語語音資料庫」(MAT)。透過電話網路，取樣頻率為 8kHz 取樣位元數為 16 位元。以 MAT2000 語料庫第五部份作為訓練語料；包括 1990 位語者，共 5,758,989 個音框。另外，從 MAT2500 語料庫中第九套裡挑選出 2893 句長句(包含 37,012 音節，其中 1512 位男生與 1381 位女生)，以人工逐句檢查的方式標示基頻值作為測試語料，同時也以 ESPS 套裝軟體求取基週期，比較彼此間的優劣。

U/V 判別正確性的比較

基頻軌跡的比較，通常初步比較的是 U/V 判別之優劣，因此先統計兩種方法所出現刪除型與插入型錯誤的機率，定義分別如下(1)刪除型錯誤：在音節範圍(使用 HMM 切音範圍)內完全沒有一個音框出現基頻的情況。(2)插入型錯誤：在 HMM 判定為靜音部分，分為兩類。類型 1：出現獨立的一段基頻軌跡。類型 2：基頻軌跡延伸到 silence 部分大於等於三個 frame。結果如表 1。

由統計結果可知，統計方法出現刪除型錯誤的機率較 ESPS 來得大，但是相對的在插入型錯誤上卻比 ESPS 小得多，代表由 ESPS 所求得的 Voiced 範圍較我們所提出的方法要大。

另以人工標示基頻值為準，比較統計法之基頻軌偵測器與 ESPS，希望了解兩種基頻求取方法在另一種 U/V 判別上的差異與正確性如何。測試音框數為 1253775 個音框。

[實驗一] 將求得的基頻值與人工標示的基頻值做比較，若出現的頻率值情況(Voiced or Unvoiced)與人工標示的頻率值狀況不相同，即判定錯誤。統計結果如表 2。

若完全以參考語句基頻值來判定 U/V，兩著的判別正確率差不多。我們所使用的統計方法在 unvoiced 的條件下 U/V 判斷正確率較在 voiced 的條件下來的高，而 ESPS 剛好相反，因此 ESPS 所決定的邊界也較寬。

因為在人工檢查基頻軌跡值時，對 U/V 邊界的一兩個音框也常會有不容易判斷的疑慮，所以分別以不統計基頻軌跡邊界音框(Edge Frame)前後的 1 至 2 個音框來統計 U/V 判別的錯誤比較。

[實驗二] 在第 I 個音框出現邊界音框時，第 I-1 個 frame 到第 I+1 個 frame 間，不計算 U/V 判別辨認率。

[實驗三]：第 I 個音框出現邊界音框時，第 I-2 個 frame 到第 I+2 個 frame 間，不計算 U/V 判別辨認率

由表 2 可以看到，當我們把基頻軌跡邊界放寬 1～2 個音框後，統計方法的錯誤率會逐漸的下降，但 ESPS 所得到的結果並沒有改善的趨勢。因此統計方法的錯誤大多都是在基頻軌跡段的邊界，對於一般基頻穩定區段的影響並不大。

音節間基頻平均值的比較

每一音節的基頻值求其平均值，與人工標音的部分求其比值，以瞭解基頻值的準確性為何，可以避免因為少數音框的錯誤而造成整體的錯誤。比值從 0.5~1 之間取其倒數作為相同的範圍作統計，結果如表 3。

在表 3 中，統計法之測試比較音節數較少是因為它的 Deletion 較多之故。上表 3 的結果，我們所使用的方法相對於 ESPS 而言，有較高的準確值。

統計式基週偵測器之重估計

以基頻軌跡偵測器重新回去求取當初建立模型之語句的基頻軌跡，將此新得到的基頻軌跡重新估計一新的統計模型。利用此新的基頻軌跡偵測器再重新求取新的基頻值，將新的基頻值如上述之各項比較方式重新統計一次，結果如表 2、3。

由結果之比較，可以觀察到以下兩點：(1)由表 3 可知，經過重新估計後之基頻軌跡偵測器所求出的基頻值更精確。(2)但由表 1 與 2 發現，以此種方法會將基頻值較不可靠的音框判斷成 Unvoiced，因此造成刪除型錯誤的上升。

整合以上結果，可以對於以統計模型之基頻偵測器作一結論；由已知的基頻軌跡資訊建立一初步的機率模型，並以此機率模型完成新的高準確性基頻軌跡偵測器，而且此偵測器可以藉由越來越好的基頻軌跡來重新估計建立更新的模型，以得到準確性越來越好的基頻軌跡偵測器。

語音信號語句中，Voiced 與 Unvoiced 出現應不相同，因此根據 U/V 出現之事前機率(a priori prob.)作為加權值的調整，重新求取新的基頻軌跡後，再與上一節重估機率模型所求得的基頻值來做比較，得到結果如下表：

經由求取 *P(V)*與 *P(U)*的同時加上一偏移量來調整機率模型，因為每個音框出現 Voiced 機率變大，加上先前在 U/V 邊界判斷的較嚴格，因此調整後的機率模型也改善了 U/V 邊界容易誤判的情況。

因為多出來的 Voiced 音框是較不可靠的，所以求得的基頻值在音節的平均值與音框的基頻值比較上，正確率有些許的下降。但整體基頻值的正確性與 U/V 判別上都較 ESPS 來

的佳。

表 1：音節之刪除型錯誤與插入型錯誤出現機率統計。

	ESPS	統計方法	重估模型	調整模型
Deletion	0.004	0.017	0.023	0.017
Insertion(1)	0.006	0.005	0.003	0.004
Insertion(2)	0.035	0.001	0.001	0.001
總錯誤率	0.045	0.024	0.027	0.023

表 2：調整估模型前後 U/V 判別比較。

錯誤率	ESPS	統計方法	重估模型	調整模型
實驗一	11.40%	12.60%	13.00%	12.50%
實驗二	11.40%	8.20%	8.50%	7.90%
實驗三	10.80%	6.30%	6.70%	6.10%

表 3:調整模型前後音節間平均基頻值的比較。

	ESPS	統計方法	重估模型	調整模型
總音節數	36,751	36,233	36,026	36,246
比值範圍	所佔比例	所佔比例	所佔比例	所佔比例
0.9~1.1	94.80%	96.20%	96.70%	96.20%
0.8~1.2	96.30%	96.90%	97.40%	97.00%
0.7~1.4	97.10%	97.40%	98.00%	97.70%
0.6~1.7	97.70%	97.80%	98.40%	98.20%
0.5~2.0	98.70%	98.90%	99.30%	99.20%

4. 連續語音聲調辨認之實驗

使用 MAT2500 語料庫第四部份，訓練語料包含 13,219 句長句(2379 位語者)；測試語料部份，包含 2893 句長句(1473 位語者)，其中 1512 位男生與 1381 位女生。扣除基頻軌跡因為刪除型錯誤致無法使用外，總訓練音節數與第一聲至第五聲之聲調音節數分佈如表 4。我們也事先將語料庫裡的第三聲發成第二聲變調(tone sandhi)的狀況依變調規則做修正。

為了避免前後連續音節耦合的問題，以前後文相關之特徵參數做訓練與辨認，得到結果如表 5。

以前後文相關之訊息作為特徵參數，整體辨認率可達到 77%。除了原來一聲、二聲與四聲的辨認正確率仍然不錯外，主要就是三聲與五聲的明顯改進之故。在連續語音中，三聲與五聲最容易受前後音節耦合所影響，再加上前後文相關之特徵參數後，對辨認率有顯著的提昇。

表 4：第一聲至第五聲之音節分佈統計。

	Tone1	Tone2	Tone3	Tone4	Tone5	音節數
訓練語料	33848	43930	24368	59119	9940	171205
測試語料	7156	8904	5202	12570	2663	36495

表 5：前後文相關特徵參數之聲調辨認結果。

聲調	聲調辨認正確率(%)				
(Tone)	Tone1	Tone2	Tone3	Tone4	Tone5
Tone 1	77.08	10.91	1.33	10.35	0.32
Tone 2	6.58	77.74	9.64	4.26	1.79
Tone 3	1.46	18.13	65.74	11.17	3.5
Tone 4	5.22	2.21	2.9	88.84	0.83
Tone 5	5.89	17.04	19.82	15.84	41.4
正確率					77.07%

5. 結論

在計畫中，以統計方法架構之基頻軌跡偵測器，能夠有比傳統基頻求取法或者是 ESPS 所求之基頻值要準確。同時也可以由求得的基頻值，藉由重新估計建立更新的模型，配合對於 Voiced 出現的機率($P(V)$)加上一個偏移量的調整，在刪除型與插入型錯誤以及整體基頻值的正確率間做取捨，可以得到一良好之基頻軌跡偵測器。在此基頻軌跡下做聲調的辨認，正確率可以達到 77%左右。

四、計畫成果自評

在本計劃中(1)建立了一套統計式之基頻軌跡偵測器；同時也完成了統計式 U/V 偵測器，經比較發現其結果較 ESPS 中之基頻軌跡偵測器為佳。(2)使用經網路製作國語聲調辨認器，並獲得 77%之辨認率。其中統計式之基頻軌跡偵測器部分之研究成果已發表於 ICASSP-2002[7]。

五、參考文獻

[1] J.D. Markel. “The SIFT Algorithm for Fundamental Frequency Estimation,” IEEE Trans. On AE, Vol.20, pp.367-377, Dec. 1972.

[2] 翁以哲，”使用統計模式之基頻軌跡偵測器”，國立交通大學碩士論文，民國九十年六月。

[3] Yih-Ru Wang and I-Bin Liao, “An Overwiew of Mandarin-Speech Tone Recognition,” Journal of the Chinese Institute of Electrical Engineering, Vol.7, No.2, pp.145-155, 2000.

[4] Sin-Horng Chen, Yih-Ru Wang, “Tone Recognition of Continuous Mandarin Speech Based on Neural Networks,” IEEE Trans. on SA, Vol.3, No2, pp.146-150, March 1995.

[5] L.R. Rabiner, “On the use of Autocorrelation Analysis for Pitch Detection,” IEEE Trans. On ASSP, Vol. 25, pp.24-33, Feb. 1977.

[6] Hong Zhang, Taiyi Huang, Junshou Song, “A New Method of Fundamental Frequency Extraction in Frequency Domain,” ICSP ‘98, pp.690-693.

[7] Yih-Ru Wang, I-Je Wong, and Teng-Chun Tsao, ' A Statistical Pitch Detection Algorithm, ', Proc. of ICASSP-2002, Orlando, USA, Vol. 1, pp. 357-360,

May, 2002.