以CELP為基礎之文句翻語音中韻律訊息之產生與調整

吳宗憲，陳昭宏，莊欣中

國立成功大學 資訊工程研究所

chwu@server2.iie.ncku.edu.tw

摘要

在本論文中，我們對於所收集的語料庫中的各個單音，分析其基週變化特性，藉由向量量化之觀念，歸納出十二組基週軌跡參考樣本，以代表對應於四聲及輕聲的音高週期之變化，並藉由一拜氏網路機率統計模型，對連續語音資料加以分析，以描述文句與語音韻律變化之關係，並在語音合成過程中，決定一適當的基週參考樣本以供作韻律上的調整。另外，我們提出了一套韻律訊息產生及調整的方法，供韻律調整模組對CELP語音合成器中之激發源脈衝加以調整。經過 20 位測試者評估之後，在平均可辨度方面達到 96.6%，而在自然度方面，評分在等級「可」以上的佔了 84.4%。

一、緒言

一般說來，語音合成的方式主要可分為兩類：第一類的作法是將可能使用的語音信號事先錄製下來，當系統欲說出某一文句時，僅須找出相對應的語音信號，將其輸出即可。這一類語音合成方式的複雜性低、運算量較小，但合成之語音易於達到自然、流利、清晰的要求，適合應用在少量文句的語音合成系統之中。另一類語音合成方式，是先將基本語音合成單元及合成規則存放於記憶體之中，利用這些基本語音合成單

元組合成與輸入文句相對應的語音信號，並配合語音合成規則加以調整音高(pitch)、音長(duration)、音強(energy)、停頓(pause)及句調(intonation)等音韻特徵。這一類語音合成方式的複雜性較高，但所需之記憶體空間較小，可以合成任意文句，因此應用範圍極為廣泛，本論文中所提之文句翻語音系統便是屬於此類合成方式。

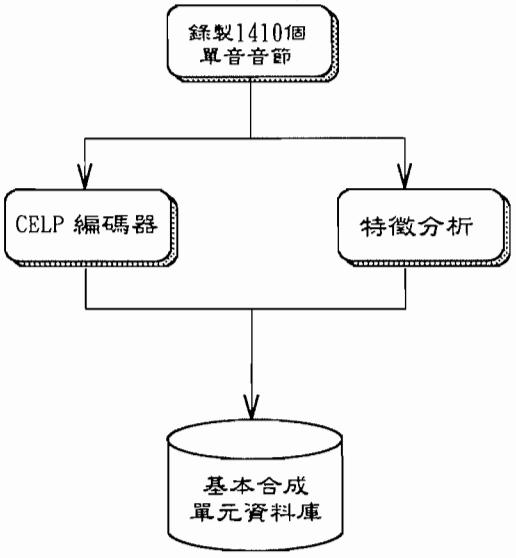
另外，語音編碼技術通常可以分為三種，即波形編碼（Waveform coding），參數編碼（Parameter coding）及混合式編碼（Hybrid coding）。波形編碼方式，一般應用在時域上（Time domain），如DM、DPCM、ADM 及 ADPCM 等，皆是於時域上去模仿語音波形，可以產生較高音質的語音，但壓縮率較低。參數編碼方式是應用在頻域上（Frequency domain），以模擬人類聲道特性為基礎，如 Linear Prediction Coding （LPC）等，雖然可以有較高的壓縮率，但所得到的音質是較不清晰的。而混合式編碼則是結合上面兩種方法的優點，如 Multi Pulse Excited Coding （MPE） 及 Code Excited Linear Prediction Coding (CELP) 等。

本論文所製作的文句翻語音系統，主要是以 408 個國語單音音節，配合聲調(tone)的變化，作為基本的語音合成單元，所以我們預先錄製了 1410 個由女性發聲的國語單音（含四聲及輕聲），利用碼本激發線性預測語音編碼技術 (CELP) 高壓縮率及其合成音質幾近原音之特性，將所有語音資料編碼壓縮後儲存，其壓縮率可達 13.3 倍。

過去的語音合成系統，通常是採用條列法則 (Rule-based) 來決定如何調整音韻的變化，但是必須藉由人工去分析大量的語音資料，這是一件非常費時且困難的工作。在本論文中，我們藉由一拜氏網路(Bayesian Network) 機率統計模型，對連續語音資料加以分析，以描述文句與語音韻律變化之關係。供韻律調整模組對語音合成器中所產生的激發源脈衝 (excitation pulse) 加以調整，以期使得輸出的合成語音更為自然、流利。

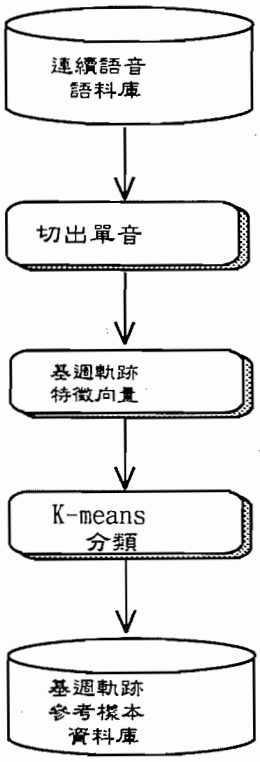
二、系統架構

基本合成單元資料庫的建立，如圖(一)所示，我們錄製了由女性發聲的 1410 個國語單音音節，包含了四聲及輕聲的變化，錄製時是以 11.025k 的取樣頻率，並控制各個單音音節的平均音量大小之誤差在 20% 以下，而且單音音節長度皆固定在 0.27 ms。接著將這 1410 個單音音節透過 CELP 語音編碼器加以處理，編碼後之參數儲存於基本合成單元資料庫中。另外，我們找出音節的母音段中（final），所有的基週中心點標記（pitch position），以及穩定區音框位置，並將這些資料亦存於資料庫中。



圖（一） 基本合成單元資料庫建立流程

在基週軌跡參考樣本方面，我們收集了約一百八十句的連續語音，然後由其中切分出各個單音音節，分析其基週變化之特性，透過 K-means 分類及向量量化之後，歸納出具代表性的十二組基週軌跡參考樣本，以對應四聲及輕聲的變化，之後，可供系統中韻律訊息產生及調整之用，見圖(二)。



圖(二) 基週軌跡參考樣本資料庫建立流程

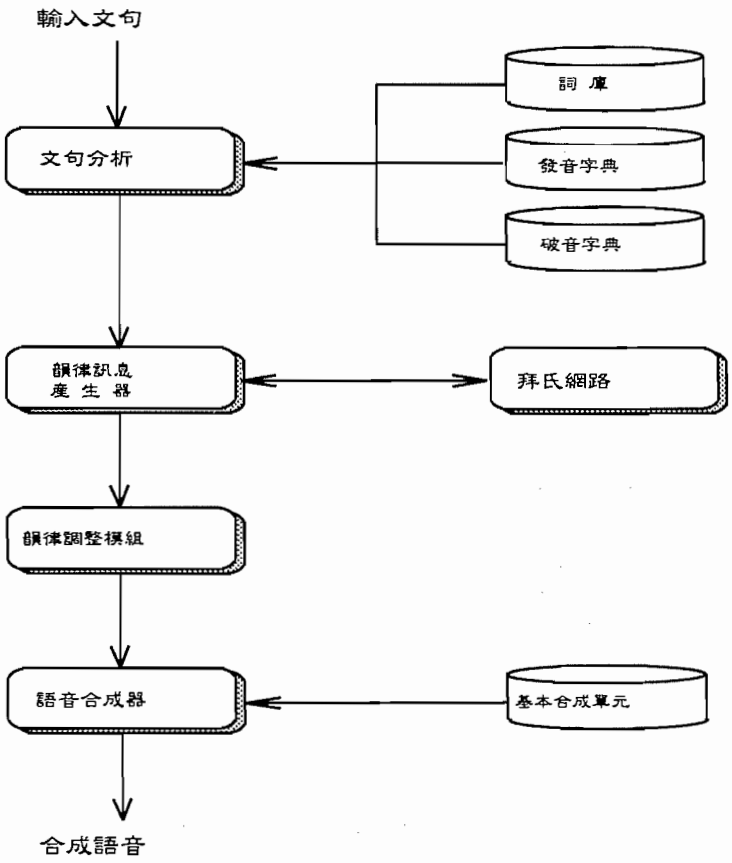
本系統的基本架構如圖(三) 所示，主要流程如下：

1. 文句分析：利用系統詞庫對輸入的文句字串進行斷詞及構詞的工作，找出詞邊界、詞類屬性、以及斷句邊界，並配合發音字典與破音字典，取得各個字相對應的注音符號。

2. 韻律訊息產生器：根據一些簡單規則，以及透過一拜氏網路機率統計模型，產生對於基週軌跡、音強、音長及停頓等韻律特徵的調整訊息參數。

3. 韻律調整模組：依據韻律訊息產生器之結果對語音合成參數進行調整修飾。

4. 語音合成器：利用碼本激發線性預測 (CELP) 語音合成器將調整後的合成參數轉換成語音信號輸出。

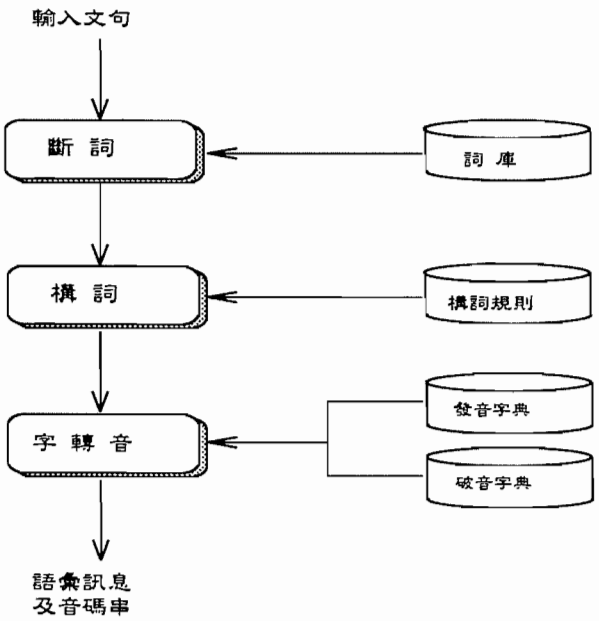


圖（三）文句翻語音系統流程

三、文句分析

整個文句分析的方塊圖如圖（四）所示。第一個步驟是斷詞（Word Identification）。主要有兩種方法，即統計學法與語言學法。統計學法是蒐集大量詞彙加以分析統計，這個方法較為簡單，不太需要語言學上的知識，但語料庫不可能涵蓋所有的詞彙，因此有可能遇到從未出現的語句而造成斷詞上的錯誤。而語言學法則是利用語言學上的知識，以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。兩種方法各有利弊，也都無法完全正確的斷詞。本系統是於事先建立一套詞庫，再配合一些簡單的規則，來處理這一部份的工作。

接下來我們將找出與每個字元相對應的正確發音。首先，我們先建立一個基本的發音字典，裡面記錄了與每個中文字相對應目前使用最普



圖（四）文句分析流程

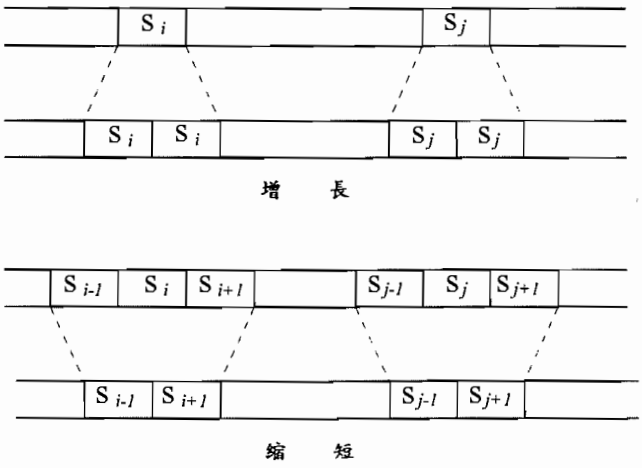
遍的發音；接著，必須對破音字加以處理，我們是對常出現的破音字，蒐集與其相關的破音詞，儲存於系統中的破音詞典中，若輸入的文句中含有破音字時，便將該部分斷詞之結果和詞典中的破音詞相比對，以決定該破音字的正確發音。另外，在我們的說話習慣中，對於某些情況的發音會有固定的變調，如下：三聲字接三聲字時，前一個三聲字會變為二聲；「一」、「不」的變調。

四、韻律調整模組

§4.1 音長、音量及停頓的調整

當音長需要增長或縮短時，主要是對基本合成單元的穩定區段加以調整，才不致使修改過的合成單元與原始合成單元相差太大。各個基本合成單元的穩定區的搜尋，我們已在系統資料的立流程中處理完成，並記錄下來。至於穩定區的找尋方法，說明如下：

首先對於每一個合成單元，以240點為一個音框(frame)，求取各音框的 LPC 係數，並轉換成倒頻譜（Cepstrum）係數。接著利用穩定區內相鄰音框之倒頻譜係數變化極小的性質，依序求出各音框與附近音框的差異性最小的兩個音框，並標示為穩定區，當我們需要調整音長時，便由語音解碼器所產生的激發源脈衝(Excitation Impulse)上，從穩定區增長或縮短，然後合成語音，見圖（五）。



圖（五）增加或刪除激發源脈衝的穩定區以改變音長

基本上音強的調整，是直接由時域上對語音波形乘上一個增益值即可。至於停頓的處理可由斷詞及構詞處理之後，在發音詞邊界處插入所需的停頓，也就是靜音，同時對各個標點符號也做不同的停頓處理，如表（一）所示。

標點符號	，	；	。	？	！
停頓時間	480ms	500ms	520ms	540ms	560ms

表（一）標點符號與停頓時間之關係

在執行時，依據以下的原則作整體的音長、音量及停頓的調整：

（１）模擬人的換氣動作，大約每六個字為一循環，遇到詞則提前或延後換氣。換氣前音量逐漸降低，換氣時略微停頓，換氣後音量較為提高。

（２）句尾要拉長音長，減小音量。

（３）遇到常用的介詞時，例如的、是、著、到、和、得等字，縮短音長、降低音量，並拉長前一個字的音長。

（３）詞之前、句尾亦加入適當之停頓。

（４）將文句分為直述句、疑問句、驚嘆句、命令句，分別調整其音長及音量整體走勢。

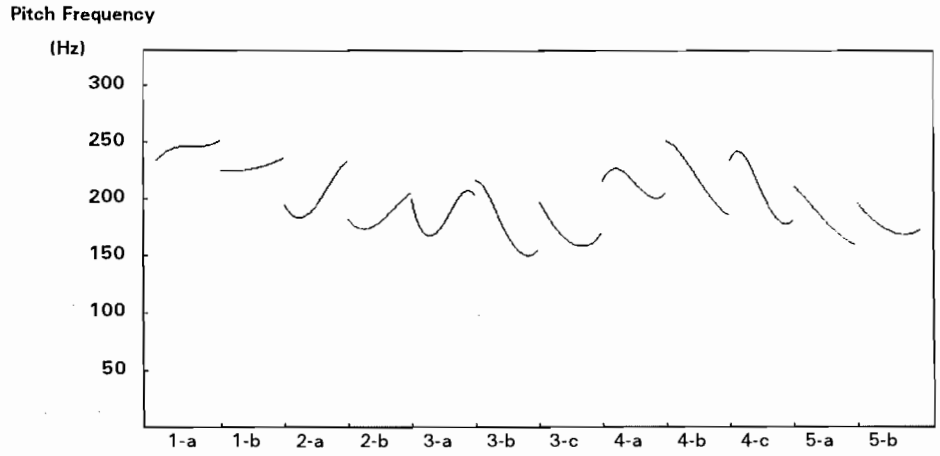
§４.２ 音高的調整

基本上每個中文單音都具有四聲變化以及輕聲，但是在組成句子時，隨著連接狀況的不同，單音的聲調會有許多的變化，僅用四聲及輕聲來表示，是不夠完整的。因此，我們根據分析統計之後，取了12組的基週軌跡樣本（pitch contour pattern）來代表原來的五聲變化，其中一聲、二聲及輕聲分別包含二類基週軌跡樣本，三聲與四聲則分別包括了三類，參考圖（六）。

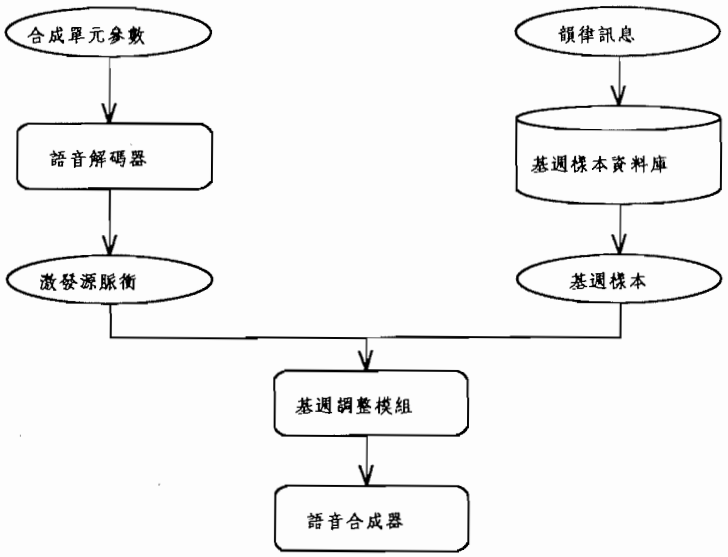
對於一個單音的基週軌跡，可利用正交化多項式展開法(Orthnormal Ploynomial Expansion)〔10〕將其轉換為一組四維的向量值 $\vec{a}=(a_0,a_1,a_2,a_3)$，來表示相似的曲線。因此，對於每個單音的基週軌跡，僅需用一個四維的向量來表示，即 $\vec{a}=(a_0,a_1,a_2,a_3)$。

接著再討論有關音高的調整，處理流程請參考圖（七），首先由韻律訊息產生器產生音高調整訊息，接著由資料庫中找出相對應的基週樣本，然後透過基週調整模組，對原合成單元的激發源脈衝，作基週上的調整。其中，基週調整的演算法是利用基週

同步重疊相加合成法(pitch synchronous overlap and add, PSOLA)[11])。



圖（六）十二組基週軌跡樣本(pitch contour pattern)



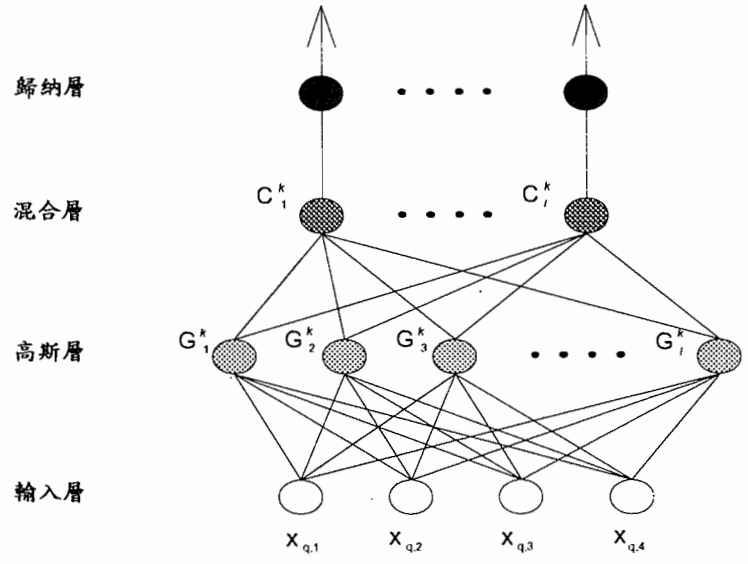
圖（七）音高週期調整流程

五、韻律訊息產生器

韻律變化的相關訊息，包括音高、音長、音量及停頓等調整參數，在過去的文句翻語音系統中，通常是採用條列式的法則

(Rule-based) 來決定韻律的調整參數，必須藉由人工去分析大量的文句與語音資料，這是一件極費時且困難的工作，另一方面，語音韻律的變化是非常繁多的，我們無法用少量的規則，即可涵蓋所有可能的變化，若使用大量規則，會造成處理速度變慢，且常會有使用規則混淆（ambiguous)的情況發生。因此我們採用機率統計觀念與分類法則，從大量的語料庫中找出一些具代表性的參考樣本，以描述文句與韻律變化之間的關係。在此我們採用拜式網路 (Bayesian Network)來建立參考樣本。

拜氏網路是一個以拜氏定理(Bayes' Theorem)為理論基礎的網路模型，其架構可分為四層：輸入層 (Input slab)，高斯層(Gaussian slab)，混合層 (Mixture slab) 及歸納層(Aposteriori slab)。圖（八）是拜氏網路的架構圖。



圖（八） 第 k 聲調的拜氏網路架構

基本上，拜式網路如同是一個拜式分類器再配合混合高斯機率分佈(Gaussian Distribution) 的觀念。首先對輸入的特徵向量分類，分類結果的每一類便代表一個高斯分佈，其平均值和變異數可經計算而

得，至於加權值則是根據某一特徵向量分佈在各類中的個數而定。拜式網路的主要目的是在求得某一輸入向量 X 在歸納層的某一類別 C_i 中出現的機率 $P(C_i|X)$，根據拜式定理 (Bayesian Theorem):

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$

其中P(X): 向量 X 出現的機率；$P(C_i)$: 屬於類別 C_i 的機率；$P(X|C_i)$: 在類別 C_i 中，屬於向量 X 的條件機率；混合層的輸出是一個混合高斯機率，它是取高斯層中的各個分佈機率乘上個別的加權值當作輸入，而決定輸出的法則一般有兩種，一是 GAM ，取機率密度總和為輸出，另一為 PGAM，則是取最大的機率密度為輸出。高斯層是根據對輸入特徵向量分類後的結果而定的。

在基週訊息產生模組中，我們對於不同的聲調(tone)分別建立一個拜氏網路，其輸出是對於某一個欲處理的字，決定應選取那一類的基週軌跡樣本以供調整。其輸入則是考慮該字在文句中相關的特徵參數，共有四項，第一項參數是相連的前一字的基週軌跡，以一組經正交化函數轉換後的向量$\vec{a}=(a_0,a_1,a_2,a_3)$表示，第二項參數是考慮下一個字的基週軌跡，但由於下一字的基週軌跡尚未計算，因此取下一字聲調相對應的代表性基週軌跡，最後兩項參數則是考慮該字在文句中的位置及在詞中的位置。下面將描述我們所用的拜氏網路架構及機率估算流程，參考圖 (八)，首先定義下列變數：

k：第 k 聲調(tone)；

N：第 k 聲調的基週軌跡類別總數；

C_i^k：第 k 聲調的第 i 類基週軌跡樣本；

M：高斯層中的子類別個數；

G_j^k：高斯層中的第 j 個子類別；

Xq：第 q 組輸入特徵向量；

$x_{q,l}$：Xq 中的第 l 項參數；

w_{ij}：連接混合層中 Node i 與高斯層中 Node j 的加權值；

對於一輸入向量 Xq ，其對應高斯層中子類別 G_j^k 的機率 $p(X_q|G_j^k)$，我們用一個高斯機率密度函數來計算，如下：

$$p(X_q|G_j^k) = N\left[X_q, \mu_j^k, \sigma_j^k\right] = \frac{1}{(2\pi)^{D/2}\left(\prod_{d=1}^{D}\sigma_{j,d}^{k\ 2}\right)^{1/2}} \exp\left(-\sum_{d=1}^{D}\frac{(x_{q,d}-\mu_{j,d}^k)^2}{2\sigma_{j,d}^{k\ 2}}\right)$$

其中 D 表輸入特徵向量的維數當高斯層子類別的條件機率求出之後，再乘上高斯層與混合層之間的加權值，即可在混合層得到混合高斯機率，計算公式如下：

$$p(X_q|C_i^k) = \sum_{j=1}^{M} p(X_q|G_j^k)\, w_{ij}$$
$$= \sum_{j=1}^{M} N\left[X_q, \mu_j^k, \sigma_j^k\right] w_{ij}$$

其中 w_{ij} 為類別 C_i^k 中的向量被分到子類別 G_j^k 的個數除以訓練時類別 C_i^k 中的向量個數所得之值。然後根據拜式定理，可由下面關係式求得歸納機率：

$$p(C_i^k|X_q) = \frac{p(X_q|C_i^k)p(C_i^k)}{p(X_q)}$$

因為對所有的類別而言，輸入向量的機率 p(Xq) 是相同的，所以上式可化簡為：

$$p(C_i^k|X_q) = p(X_q|C_i^k)p(C_i^k)$$

上式中的 $p(C_i^k)$ 等於類別 C_i^k 中訓練樣本的個數除以全部訓練樣本總數所得之值。最後，找出具有大輸出機率值的類別 C_i^k，即是我們欲選取的基週軌跡類別。

六、實驗結果與結論

§6.1 實驗過程與結果

我們分別就可辨度(intelligibility)與自然度(naturalness)這兩方面評估系統的效能。實驗對象共 20 人，過程說明如下：在硬體方面包括了 PC/AT 486個人電腦、8-bit 聲霸卡以及外接喇叭，軟體方面則分為系統主程式 (250k bytes) 、詞庫檔 (360k bytes) 與合成單元資料庫 (939k bytes)。在可辨度的評估方面，我們以本系統對輸入文句做即時處理，立即輸出合成語音。測試者必須在語音輸出之後，將所聽到的語音以聽寫方式寫出，最後統計正確文句與被測試者所寫下的結果之間的差異，以作為評估可辨度的方式。在自然度方面，則是將事先準備的測試文件先由原先錄製音檔的該名女性以自然方式唸出，當作滿分標準，再由未經過韻律處理的系統播放，最後由本系統播放同一份文件，測試者則根據個人觀點對此系統的接受程度進行評分，評分的等級分為「優」、「佳」、「可」及「劣」四種。實驗結果顯示在表（二）及 表（三）。在可辨度方面，使用12組基週軌跡樣本可以達到平均 96.6%的可辨度，，未使用12組基週軌跡樣本則可達到平均97.1%的可辨度。我們發現兩者的可辨度相差無幾。在自然度方面，使用12組基週軌跡樣本的系統，在等級「可」以上的佔了 84.4%，未使用使用12組基週軌跡樣本的系統，在等級「可」以上的僅佔 79.6%。仔細比較各表中自然度的等級所佔的比例，我們可以發現本系統所輸出的語音已讓人感覺有不錯的效果。

測試種類	數　量	可辨度
單音	1410	92.8%
二字詞	200	95.4%
三字詞	200	98.7%
四字詞	200	99.1%
句子	100	97.3%
平　　均		96.6%

(a)可辨度方面實驗結果

測試種類	數量	自　然　度			
		優	佳	可	劣
二字詞	100	45%	34%	8%	13%
三字詞	100	32%	37%	16%	15%
四字詞	100	30%	44%	9%	17%
句子	100	25%	36%	17%	21%
短文	5	20%	52%	16%	12%
平　　均		84.4%			15.6%

(b)自然度方面實驗結果

表（二）使用12組基週軌跡樣本的實驗結果

測試種類	數　量	可辨度
單音	1410	94.4%
二字詞	200	95.4%
三字詞	200	98.9%
四字詞	200	99.1%
句子	100	97.6%
平　　均		97.1%

(a)可辨度方面實驗結果

測試種類	數量	自　然　度			
		優	佳	可	劣
二字詞	100	39%	30%	15%	16%
三字詞	100	31%	34%	15%	20%
四字詞	100	26%	33%	19%	22%
句子	100	20%	33%	22%	25%
短文	5	19%	35%	28%	18%
平　　均		79.6%			20.4 %

(b)自然度方面實驗結果

表（三）未使用12組基週軌跡樣本的實驗結果

§6.2 結論

在本論文中，對於韻律訊息的產生及調整的研究，確實對合成語音的自然度及流利度有明顯的提昇，也可達到即時處理的要求，但是要讓一般大眾能夠廣泛接受，成為一套實用價值高的文句翻語音系統，仍有一些需要改進的地方。

首先，若能夠增加斷詞及構詞的正確性，有助於音韻調整訊息的產生，對於破音字的處理，如何能找出正確的讀音，也是目前需要改善的地方。接著，在語音合成器方面，由於為了降低語音資料所佔的記憶體空間，所以利用語音編碼技術將資料加以壓縮，但是解碼還原之後的語音與原音比較仍有些許誤差，而在經過韻律調整模組中，PSOLA 演算法的處理後，由於為了將兩段波形能夠平順的連接在一起，部分運算難免會造成波形頻譜上的失眞，導致雜訊產生。因此，我們在時域上加上一個後置低通濾波器(low-pass filter)，其主要的功能是將共振峰波谷內的雜訊降低，因為人類的聽覺遮蔽效應，使得語音中共振峰附近的雜訊會被遮蔽，而波谷處的雜訊便相對地特別明顯。另外，在我們的錄音過程中，是以一種類似標準化的方式來錄製每一個基本合成單音，這種方式雖然能使合成出來的語音，各個單音都能較為清楚，但是，它卻缺少了連續語音中，字與字之間該有的連音(coarticulation)資訊，所以，未來我們應找出適當的方法來彌補這一個缺失，以提高合成語音的流利度。另外，目前系統所使用供作分析的語料庫，似乎仍不夠完備，在分析並學習文句與語音韻律變化之關係上，應再收集更多自然語音資料，藉由更大量的資料訓練，將來在處理各方面的應用時，才能產生更客觀且完整的韻律訊息。

參考文獻

〔1〕Lin-shan Lee and Chiu-yu Tseng "Mandarin Speech Input/Output Techniques for Chinese Computers - The Stateof the Art," Proc. Natl. Sci. Counc. ROC(A), Vol. 11, No. 4, pp. 273-290, 1987.

〔2〕歐陽明、李琳山，「一套中文的文句國語音系統」，台大電機研究所碩士論文，1985年6月

〔3〕劉繼謚等，「以線性預測編碼為合成器的中文文句翻語音系統」，電信研究季刊，第19卷第3期，民國78年9月

〔4〕朱國華等，「一套以多脈衝激發語音編碼器為架構之即時中文文句翻語音系統」，電信研究季刊，第21卷第4期，民國80年12月

〔5〕Ngor-chi chan and Chorkin Chan, "Prosodic Rules for Connected Mandarin Systhesis," Journal of Information Science and Engineering 8, pp.261-281, 1992.

〔6〕John Choi, Hsiao-wuen Hon, Jean-luc Lebrun, Sun-pin Lee, Gareth Loudon , Viet-Hoang Phan, and Yoganathan S., "Yanhui, a Software Based High Performance Mandarin Text-to-Speech System," Proceedings of ROCLING VII, pp.35-50, 1994.

〔7〕李俊曉等，「中文文句翻語音系統的語言處理模組」，電信研究季刊，第21卷第4期，民國80年12月

〔8〕Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 bit/second Code Excited Linear Prediction (CELP), National Communications

System, Office of Technology and Standards, Washington, DC 20305-2010, 14 February 1991.

[9]Lin-shan Lee, Chiu-yu Tseng, and Ching-jiang Hsieh, "Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System," IEEE Trans. on Speech And Audio Processing. Vol. 1, No. 3, July, 1993.

[10]S. H. Chen and Y. R. Wang, "Vector quantization of pitch information in Mandarin speech," IEEE Trans. Commun., Vol. 38, pp. 1317-1320, Sept. 1990.

[11]F. J. Charpentier and M. G. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation," ICASSP, pp.2015-2020, 1986.

[12]Christian HAMON, Eric MOULINES, Francis CHARPENTIER, "A diphone synthesis based on time-domain prosodic modifications of speech," ICASSP, pp.238-241, 1989.

[13]L. S. Lee, C. Y. Tseng and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," IEEE Trans. Acoust, Speech, Signal Processing, Vol.37, No.9, pp.1309-1319, Sept. 1989.

[14]Michael S. Scordilis and John N. Gowdy, "Neural network based generation of fundamental frequency contours," ICASSP, pp.219-222, 1989.

[15]S. H. Chen, S. H. Hwang, and C. Y. Tsai, "A first study on neural net based gendration of prosodic and spectral

information for mandarin text-to-speech," ICASSP, pp.45-48, 1992.

〔16〕Jhing-Fa Wang, Chung-Hsien wu, Shih-Hung Chang, and Jau-Yien Lee, "A Hierarchical Neural Network Model Based on A C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition," IEEE tras. Signal Processing, Vol 39, No. 9, pp.2141-45, 1991.

〔17〕Michael S. Scordilis and John N. Gowdy, "Neural network control for a cascade/parallel formant synthesizer," ICASSP, pp.297-300, 1990.

〔18〕Yoshinori SAGISAKA, "On the prediction of global F0 shape for Japanese text-to-speech," ICASSP, pp.325-328, 1990.

〔19〕Shaw-hwa Hwang and Sin-hrong Chen, "Neural network synthesiser of pause duration for mandarin text-to-speech," ELECTRONICS LETTERS, Vol.28, pp.720-721, April, 1992.