國 立 交 通 大 學

電信工程學系

碩士論文

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

Unit Selection for Corpus-based Mandarin TTS System

研 究 生：吳佩穎

指導教授：陳信宏　博士

中華民國九十四年七月

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

Unit Selection for Corpus-based Mandarin TTS System

研 究 生：吳佩穎　　Student : Pei-Ying Wu

指導教授：陳信宏 博士　　Advisor : Dr. Sin-Horng Chen

國立交通大學

電信工程學系

碩士論文

A Thesis

Department of Communication Engineering

College of Electrical Engineering and computer Science

National Chiao Tung University

In Partial Fulfillment of Requirements

for the Degree of

Master of Science

in Electrical Engineering

July 2005

Hsinchu,Taiwan,Republic of China

中華民國九十四年七月

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

研究生：吳佩穎　　指導教授：陳信宏博士

國立交通大學電信工程學系碩士班

中文摘要

文字轉語音系統中所使用的合成單元已經從小量的合成音庫，演變為以大型語料庫為基礎的合成音庫。在本論文中，設計了一套以 Corpus-based 為基礎的中文文句翻語音合成系統，這種作法通常會遇到兩項問題：如何有效率地去大型語料庫中找出所有可能的候選合成單元？如何解決合成單元間相串接時韻律差異的問題？本論文提出以連續相關比對法選取所有可能的候選合成單元，主要是依據在同一句子中其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來作比對，接下來，再利用我們提出的 cost function 選取出具有與合成目標最相近之語音及韻律特徵的合成單元，最後將選出之最佳合成單元作串接輸出成為合成語音。

為了瞭解本套合成系統之語音品質狀況，我們利用主觀式評估方式，進行自然度 MOS 測試，並且進一步對合成語音作結果分析，探討合成語音出現不佳狀況時可能的影響因素。由實驗結果可知，本論文提出之方法，在合成語音的自然度上，會有不錯的表現。

Unit Selection for Corpus-based Mandarin TTS System

Student：Pei-Ying Wu　　Advisor：Dr. Sin-Horng Chen

Institute of Communication Engineering

National Chiao Tung University

Abstract

Synthesis units in Text-to-Speech system have developed from base syllable to waveform units of variable lengths. A set of corpus-based text-to-speech synthesis technologies for Mandarin Chinese usually comprises two main problems to solve： How to find all possible candidates in speech corpora effectively? How to select appropriate synthesis unit to concatenate? Firstly, the thesis presents a continuous-correlative comparison method to solve searching candidates' problem. Secondly, cost function is used to find the appropriate synthesis unit retrieved from the corpus and concatenated to produce the output speech.

Finally, we use a subjective test called Mean Opinion Scores (MOS) to test whether our synthesized speech is natural or not. The assessment indicates our Corpus-based Mandarin TTS System indeed significantly improve the naturalness of synthesized speech quality. Besides, we also analyze synthesized speech to give advices in the future works.

誌謝

我在我的職場狂奔，我和我的理想私奔！大學畢業進入職場，一直以為有了穩定的工作後就不會再重回校園，但是，工作還未到一年，心中對於理想的追求似乎仍聲嘶力竭地吶喊著，身處工作壓力不敢輕舉妄動的我，多虧有了老哥的鼓勵，我才有放手一搏的勇氣。

在這裡要特別感謝陳信宏老師，給了我研究時間上極大的包容，以及對我的耐心指導，讓我無論在生活上或是研究上都獲益良多；謝謝王逸如老師，每週的meeting，是一個很棒的學習機會，可以親自體會要如何呈現及表達。還有，實驗室親愛的夥伴們，很高興能和你們共同生活兩年*^_^*，要大力的謝謝金翰，在電腦及程式上你對於我的指導，我想尊稱你為小老師也不為過吧，讓我從無到有吸收了很多知識；希群，你的感性與知性，我想在不久的將來一定可以成為萬人迷的；柏暄的幽默，總是讓實驗室充滿了歡笑；隆勳與順哥，人家說，認真的女人最美麗，我想你們可以說是認真的男人最帥囉！在兩年研究所生涯中，多謝江振宇學長的指導與國興的協助，以及我工作單位所有的上級長官與同事，謝謝你們的體諒，我才能順利完成我的畢業論文。當然，號稱最美麗的曾琦、最帥的子洋與最疼我的小哈，謝謝你們不時給我壓力上的紓解與心情上的愉悅，沒有你們，我想我的生活將會缺少許多樂趣；還有我親愛的父母親以及在美留學的老哥，僅將此篇論文獻給你們，雖然你們分別遠在屏東與美國，但是你們給予我精神上的鼓勵與支持卻是最令人振奮的，感謝你們，我親愛的家人！

在領到畢業證書的這一刻起，我完成了我研究所的理想，下一步？我想我會繼續在我的職場上努力狂奔吧！

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

隨著科技的蓬勃發展，人類越來越仰賴電腦來處理身邊各項事務，於是乎，電腦科技的發展已從原本的運算能力導向轉變為以溝通與訊息交換為主要研究目標；在這個過程中，早期的研究主要是致力於如何提供最有用，最有價值的資訊，資訊檢索系統、網路搜尋引擎、資料探勘技術應運而生，然而，資訊最終的目的是要提供給使用者，所以人與電腦間的溝通就顯得格外重要。

觀察人類最自然的溝通方式，不外乎聽與說：聽出正確的訊息(辨識)，說出要表達的話(合成)，為了讓這兩種表達方式，也能成為人機間的溝通模式，語音辨識和語音合成技術的研究與發展，扮演了舉足輕重的地位。

最簡單的合成語音，是將預先錄製的有限量字詞音檔儲存在資料庫中，再利用固定的截字句方式，將音檔串接成目標句後播出，但是此類型的播音系統，僅能處理固定的句型與有限的語句。若要處理文字內容不固定，例如一套有聲電子書，要以語音的方式唸出書的內容，或是要以語音方式唸出我們所接收到的e-mail,這樣一種可以將無限制的文句自動轉成語音的合成系統，我們稱為文句翻語音系統(Text-to-Speech system, TTS System)。本論文主要著重在設計一套以語料庫為基礎之中文文句翻語音系統，期能使聲音音質之自然流暢度更為提升。

1.2 研究方向

本論文之研究重點，在於設計一種挑選合成單元的方法。首先，建立合成單元的方式已由獨立錄製單一合成單元，漸漸改為錄製載字句以及錄製一個大型語

音資料庫，載字句錄音法把要錄製的合成單元鑲在一個句子中一起錄音，最後再將它切出來，這種合成單元本身就具有連續音的特性，往往只要微調音韻或是不必調整直接將合成單元串接，就能合成自然的語音。一般而言，合成單元越大，所合成出來的音質及語音自然度也越好。

本論文中，使用大型語料庫，含括較多的韻律及頻譜之串接組合類型，要如何從中挑選出適合的單元始能得到最佳流暢度，便成為本論文最重要的課題。

1.3 章節概要

本論文共分為五章：

第一章 緒論：介紹本論文之研究動機與方向。

第二章 大型語料庫(Corpus)之設計與合成單元之生成：描述大型語料庫之設計與建置，並說明合成單元選用的考量與生成。

第三章 以 Corpus 為基礎的合成系統設計：介紹 Corpus-based 合成系統架構、說明如何從語料庫中挑選合成單元的方法。

第四章 實驗結果與分析：說明本研究之 Corpus-based 中文語音合成系統相關設定、挑選合成單元機制之效能分析，並利用主觀式測試與聽覺測驗，評估合成系統的好壞。

第五章 結論與未來展望。

第二章 大型語料庫設計與合成單元之考量

一套能合出自然流暢語音的 TTS 系統，關鍵在於語音的韻律變化是否自然順暢，尤其當合成單元在相連接處若屬韻律不連續現象，會破壞整體的合成品質[1]，若能減少合成語句中合成單元的相接次數，對於合成品質的提升，有極大的幫助。基於上述理由，近年來以大型語料庫為基礎的 TTS 系統已成為主流；這樣的系統相較於過去我們所發展的合成系統而言，差別即在於語音資料庫的大小。

採用以大型語料庫為基礎的 TTS 系統，首要工作就是建立一個足夠大型的語音資料庫，以供未來合成單元的挑選。在接下來的小節中，將會針對目前所處理的語料及如何建制一個大型語料庫作說明，並介紹要如何求取語料庫中語音參數，為之後我們在合成語音時作準備。

2.1 大型語料庫之設計與建置

一套好的 Corpus-based 語音合成系統，主要關鍵即在於有豐富的合成單元。就中文而言，每個中文字對應一個音節(Syllable)，每個音節有不同的聲調(tone)，也就是中文語音學上的四聲變化，常見的中文字有一萬兩千多個，但因在發音上同音字很多，總共的音節大約是 1300 個[2]，如果不計聲調，共有 411 個不同的基本音節。

然而中文的發音通常是以詞為基礎，音節僅能包含子音和母音相連接的連音變化方式，對於音節和音節之間的音韻變化並無展現，是故，如何使用一個兼具“豐富語音(phonetically rich)”及 “豐富韻律(prosodically rich)”的語料庫，是在進行語音合成之前本論文要提出說明的。

2.1.1 大型語料庫說明

本論文使用的語料庫文字稿(Text)部分來自於中央研究院中文文句結構樹資料庫 1.1 版(Sinica Treebank Version 1.1)，從中央研究院詞庫小組之中央研究院現代漢語語料庫得來，將語料庫內容及構建方法簡述於下，詳細的內容請參考[3]。

中央研究院現代漢語語料庫（簡稱「研究院語料庫」(Sinica Corpus)），是專門針對語言分析而設計的，每個文句都依詞斷開，並標示詞類。在語料的蒐集盡量做到現代漢語分配在不同的主題和語式上，是現代漢語無窮多的語句中一個代表性的樣本，語料庫約五百萬詞，依照各主題其詞類、字數和篇數如表 2-1 所示：

表 2-1：中央研究院現代漢語語料庫文句主題統計表

主題	加總的詞數	加總的字數	篇數
文學	777050	1169801	1385
生活	858750	1398791	2301
社會	1610997	2711720	3246
科學	629838	1054738	994
哲學	439955	673080	695
藝術	474340	781415	518
空白	101394	160306	89
加總結果	4892324	7949851	9228

中文文句結構樹資料庫1.1版，是從上述五百萬詞的中央研究院平衡語料庫求出的。在剖析中文句子方面，是以訊息為本格位語法（Information - based Case Grammar, ICG）的表達模式，此以詞彙為中心，並且配合中心語主導原則

(Head-Driven Principle)的中文剖析系統，對於剖析的句子除了記載語法訊息外，也藉由語意角色的指派標示出詞和詞之間的語意限制關係。關於結構樹的建構方面，首先從具有標記的五百萬詞中，抽取句子，並經由中文剖析系統，電腦自動剖析並且產生結構樹，如此可以盡量維持結構標記的一致性，再利用句結構樹編輯程式並配合標記原則，加以人工修正及檢驗，以維持標記的正確性，並且對於歧義的句法結構形式及詞類標記也有適當的處理。目前本系統使用作為大型語料庫的約有1,900棵中文結構樹，計有8,017個詞，而所選取的文章主題涵蓋政治、旅遊、運動、財經、社會等。

接下來的工作便是錄音。我們請了一位專業的女性廣播人員幫我們錄製語音，為了減少未來在合成單元切割或韻律求取上的錯誤，以得到較佳的合成單元品質，我們在錄製的過程中若有發生口吃、猶豫、或唸錯的情形，我們會請錄音人員再重新錄製該句直到正確為止，錄音軟硬體設備及格式詳如下表：


表：

表 2-2：錄音軟硬體設備規格表

錄音軟體	Cool Edit Pro 直接錄成聲音檔案
麥克風	單一指向性 (uni-directional)
錄音場所	普通房間
錄音情境	依照所選出文稿唸出
取樣頻率(sampling rate)	20 kHz
發音速度	每秒約 4.6 個音節
取樣大小	16 bits (位元)
聲道	單聲(mono)
檔案格式	pcm

2.1.2 合成單元選用的考量

早期的語音合成系統大都採用音素當基本合成單位，在記憶體逐漸便宜及對音質要求下，合成單元也跟著加大，目前外文 TTS 系統大都採用雙音為主要的合成單元[4,5]，中文 TTS 系統則是以單音節當合成單元[6,7]。有研究[8]亦提出，認為中文 TTS 系統也應該使用雙音當合成單元，主要是考量雙音在串接式的 TTS 系統中較能合成富有連音的音，然而使用雙音當合成單元的最大問題是在串接時合成單元間的頻譜不連續的現象，以致於產生不悅耳的語音，雖然音節之間的串接也有這種問題，不過並不明顯。

決定基本合成單元有以下三種考量：

(1)單元內及單元間的連音(Coarticulation)是否足夠。

(2)單元間的頻譜是否連續。

(3)所需的記憶體及錄音的數量是否適當。

在語音合成時，因為相連接部份不連續所造成的修飾，是造成語音品質下降的主因，且若單以單音節為合成單元，雖保留住了大部分子音接母音的音韻與聲調變化，但是會失去字詞或更高層級的變化。於是，為了得到更好的合成語音，有些研究專注於使用不定長度的合成單元[9,10,11]，這種合成單元通常由大量的語音資料庫取得，由於合成單元大，甚至大到整個句子以上，連音資訊多，音韻調整的部分很少，因此合成語音音質及自然度都相當好，所需的代價是錄音時間較長及大量的儲存媒體。

本論文中提出在使用擁有充分的連接變化與韻律訊息的大型資料庫下，除了以單音節為基本合成單位外，亦加入了以二字詞、三字詞、甚至是以一個句子為單位來當做合成單元。

2.2 合成單元參數之求取

建立了大型語料庫後，在語音合成語音階段時，我們需要大型語料庫中語音的資訊去做進一步的處理，其中包括音節的邊界點及韻律參數等相關資訊，這些資訊的精確度，無論是對於日後合成單元的挑選或合成語音品質而言，都有很重要的影響力。

2.2.1 切割資訊的求取與修正

由錄音員錄製的語料是連續語句的聲音語料，在作合成單元的挑選前，需要將最基本的合成單元標記出來，由於語料庫資料量過大，若用人工標記是不切實際的，不僅費時冗長且不一定完全精確，於是我們使用先前學長提出的一套語音自動切割與修正的方法如圖 2-1 所示，詳細內容請參考[12]，在這裡我們只做簡單的說明。

首先，我們使用隱藏馬可夫模型(HMM；Hidden Markov Model)辨認器，先做初步強制切割(Forced alignment)的工作，使用的原始模型由 TCC300 語料庫訓練而成。在參數設定上，我們使用 38 維的參數，其中包含 13 階的梅爾倒頻參數(MFCCs)、delta 梅爾倒頻參數(ΔMFCCs)、delta delta 梅爾倒頻參數(ΔΔMFCCs)、以及能量對數值(log energy) [13]，frame rate 設定為 10ms。在作法上，是對於先前所錄製的聲音語料，根據其文字內容，利用已知基本音節型態(base-syllable type)所對應的隱藏式馬可夫模型，使用 Viterbi 演算法，搜尋每個對應模型狀態外轉的位置，據此找出每個音節的邊界位置。

由於單純使用隱藏式馬可夫模型所切出的邊界位置，並非所有結果都在可以接受的範圍內，因此必須針對某些不適當的邊界作修正。觀察自動切割的結果發現有兩個主要問題存在，一為音節的 boundary 有偏差現象；二為中文音節的子

音、母音交界點位置不正確，因此在學長的研究中，是利用能量(energy)，過零率(zero crossing rate)和 Voicing probability 參數，分別對音節間切割位置及音節內子音母音交界位置作修正。

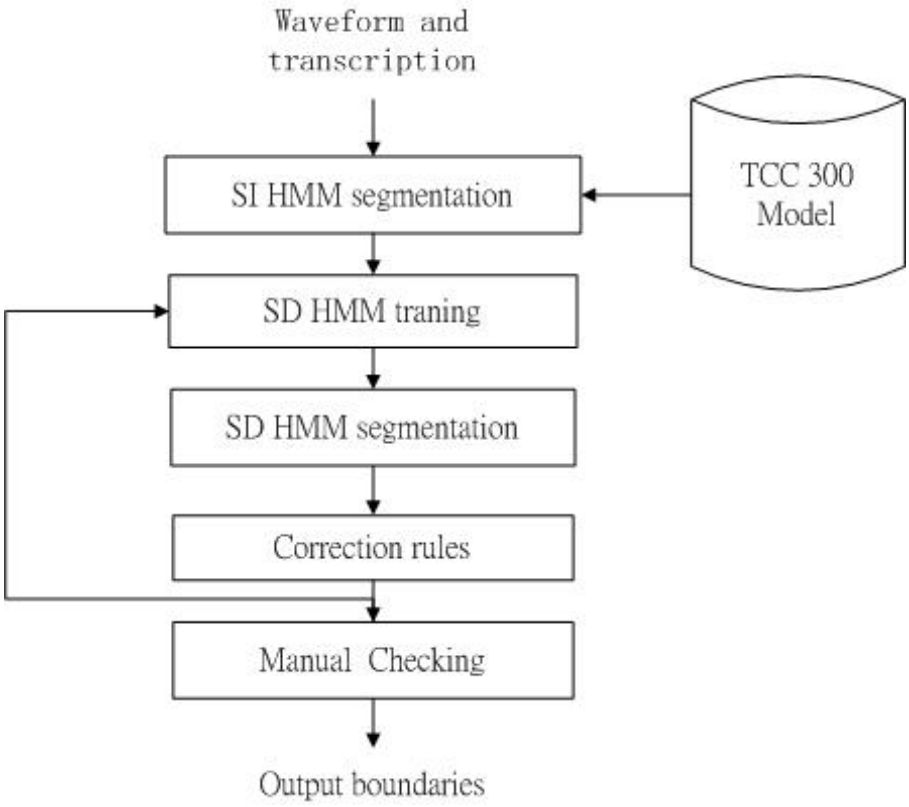


圖 2-1：語音切割處理方塊圖

2.2.2 韻律參數的求取

在確定合成單元的切割資訊後，就可以使用並進一步求出韻律參數。我們使用 Wavesurfer 軟體及 ESPS 演算法，windows size 取 10ms，求出的韻律參數包括：

- 音節編號(Syllable code)
- 音節聲調(Tone)
- 音節在詞中位置

- 詞性類別(Parts of Speech, POS)
- 音節中子音、母音長度及與上一音節停頓長度(Initial，Final and Pause)
- 基頻(Fundamental Frequency)
- 能量(RMS Power)

第三章　以 Corpus 為基礎的國語語音合成系統設計

3.1 Corpus-based 國語語音合成系統架構

近幾年來，國立交通大學電信工程學系語音實驗室致力於國語語音合成系統的發展，已成功合成出相當流利的語音[14]。但是上述系統是以單一音節(Isolated base-syllables)為語音資料庫(Acoustic inventory)，有鑒於近年來語音合成發展日趨進步，若改採大型語料庫，合成出來的語音會較目前系統自然流暢，本論文提出以 Corpus-based 為基礎的語音合成系統，這種作法的主要問題包括：如何設計一個錄音良好的語料庫，如何手動或自動標記切點以及韻律資訊，如何選擇及決定合成單元型態，以及如何挑選每個單元型態的語音段。前面兩項考量在前一章節已經做過討論，接下來的章節，將會提出如何選取最佳合成單元的方法，並實驗分析由該系統合成出的語音音質。

本論文提出的以大型資料庫為基礎的中文語音合成系統基本流程圖如下：

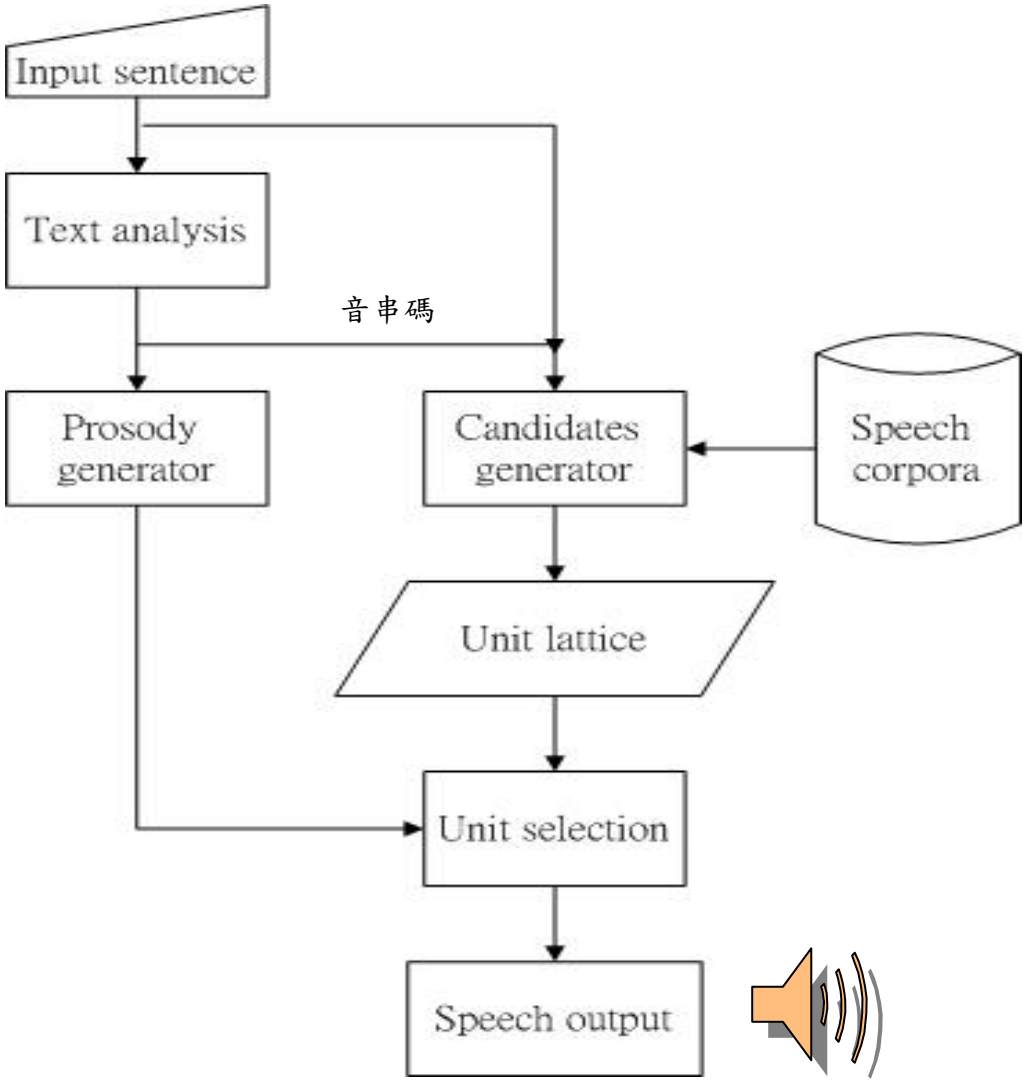


圖 3-1：Corpus-based 中文語音合成系統基本流程圖

它主要包括四個模組：

文字分析：主要目的在對於輸入的文字，分析出正確的音串碼(syllable sequence)、詞串(word sequence)、和詞類串(POS sequence)等語言參數(Linguistic features),因此又稱文字分析器為斷詞器(Tagger)。最基本的文字分析包括字詞分析、字轉音處理以及破音字處理，輔助的部份包含阿拉伯數字、時間表示等處理步驟，更進一步的會加入文句的語法、語意分析與音韻分析，再轉成語言特徵參數後提供給韻律產生器，做進一步的處理。本論文中使用的中文斷詞器斷詞精確率已達 87.5%，召回率 79.3%[15]。

韻律生成：韻律產生器的功能，是由文字分析器(TA)所產生的語言參數syllable sequence、word sequence、POS sequence中抽取出詞長、詞類和音碼作為輸入，輸出每個音節所對應的韻律訊息，包含四個基頻軌跡參數、三個時間長度參數、和一個能量參數。目前韻律產生器是以遞迴式類神經網路(Recursive neural network, RNN)的概念建構而成的[14]，由於類神經網路可模擬人腦學習與記憶的功能，因此在長時間的訓練下可獲得不錯的效果。

候選合成單元(candidates)的產生：根據輸入文章及文字分析器輸出的音串碼，利用語音資料庫中的文字結構樹，挑選出句中所有可能組成目標句的最長詞串，如此便可得到所有組成目標句的可能詞串，亦即候選合成單元，我們稱之為unit。

合成單元挑選：將產生的 unit，構成所有可能組成目標句的 lattice,並依據韻律產生器產生之音韻參數，由所有可能組成目標句的候選組合中選取和目標音韻參數差異最少者，將挑出的合成單元進行串接並輸出合成語音。

3.2 選取候選合成單元分析

在任何對波形有做修飾都會造成合成音質下降的情況之下，大型語料庫所儲存的合成單元，如果可以找到較長詞來當合成單元，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。

合成單元的選取主要是以詞為單位，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。例如：

依 據 行 政 院 主 計 處 的 統 計

就這個句子而言，所有可能衍生出來的組合有很多，譬如：

(1) 依 **據行政院** 主計處 的 統計

(2) 依據 行政院主計 **處的統計**

(3) 依據行政院 主計處 的 統計

(4) 依據 行政院 主計處 的 統計

(5) 依據 行政 **院主計處** 的 統計

……

但是，其中有許多組合是不符合中文音韻的文法結構，例如：「**據行政院**」、「**處的統計**」、「**院主計處**」，除了這個考量外，若真要搜尋所有構成目標句的可能組合，所要耗費的時間跟空間複雜度太過於龐大，因此我們提出了一套選取候選單元的挑選機制。

本套機制主要考慮觀點有下述兩項：第一，依據人類構句的方式，根據中文發音的音韻和斷句，我們可以找到合適的合成單元；由於人類構句的方式，是先將單音節組合成詞，再將多個詞組合成長詞或專有名詞，進一步組合成片語、句子，如下圖所示：

世 界 上 最 耐 渴 的 人 是 澳 洲 中 部 土 人

世界上 最 耐渴 的 人 是 澳洲 中部 土人

世界上 最耐渴 的人 是 澳洲中部 土人

世界上最耐渴的人 是 澳洲中部土人

圖 3-2：人類構句方式模擬示意圖

因此，我們可以依據這個想法，先把不適合的組合性去除，並可根據不同階層的構句組合方式，進行所有階層中合適單元的挑選。首先，我們使用的語料庫是取自中央研究院中文文句結構樹資料庫 1.1 版，每一棵結構樹，是由選出的句

子，經由中文剖析系統，依據中文語法、語意訊息和中心語主導原則，電腦自動剖析並且經人工修正產生的[3]，因此剖析出來的每一個成分，都標記有語意角色、詞、詞組類型等。藉由剖析出來的結構樹，我們可以知道語料庫中所有文字語料的語法結構。以下舉例說明，如圖 3-3 所示：

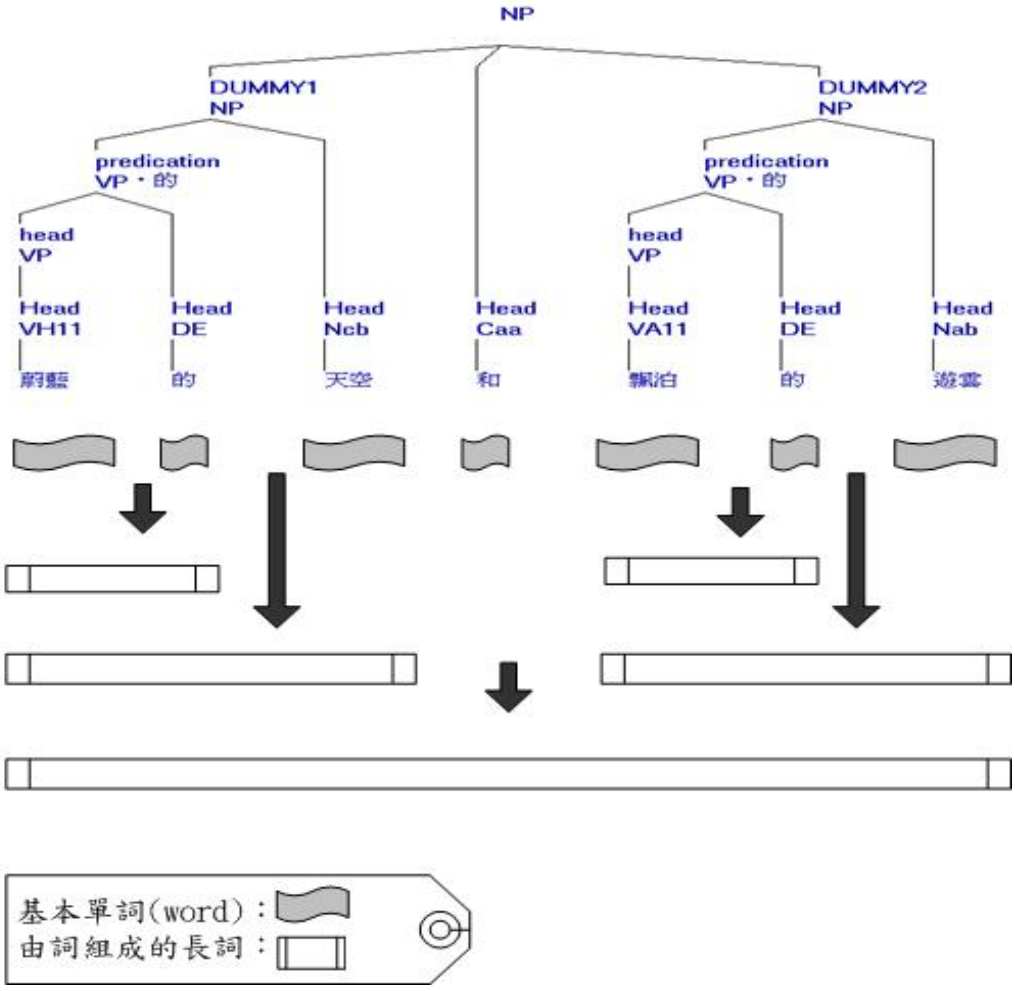


圖 3-3：依據中文結構樹挑選出適合之合成單元範例

我們用一棵結構樹代表一個句子，這棵樹上的每一個終端節點，代表的是一個詞，我們的作法便是將這些詞視為該句人類構句中最適當的基本單詞(word)，接下來再進一步將這些基本單詞組合成長詞或專有名詞，如此便可以移除不適當的長詞組合，並挑選出適合的合成單元；也就是所要選取的合成單元中包括有詞，及由詞組合成的長詞，如此將能減少對語音訊號的修飾，合出來的語音自然較為流暢。

第二，Corpus-based TTS 是基於擁有充分的連接變化與韻律訊息的一套系統，所以能有越大型的資料庫越好，為了能將本套系統之 Corpus 在未來能繼續擴大，我們設計一種有效率儲存語料庫資訊與搜尋候選合成單元的方法。

3.3 候選合成單元的產生

首先，輸入的目標文章可能是由多個句子構成的，要以多大的範圍為限制找出多長的候選合成單元，是一個要考慮的課題。另外，在大型資料庫越大越好的情況下，為使語料能繼續擴充，我們希望能減少儲存語料訊息的空間，並且利用時間複雜度較少的機制，在大型資料庫中將這些可能的候選合單元全部找到。

3.3.1 候選合成單元的長度限制

對於中文發音中韻律與節奏的產生，停頓扮演了一個很重要的角色，它不只可以避免語意上的混淆(syntactic ambiguity)，更可以增加中文句音韻的效果。由於每個中文字是由一個音節所構成，而一個詞，是由一個、兩個或多個中文字所組成的，因此，一個句子便是由一連串的音節所組成，斷句不當會造成意思上的偏差和爭拗，而斷句的位置與長度，便決定了該句的語意及音韻。

在本系統中，輸入的目標文章，具備有標點符號，其中包括有頓號、冒號、逗號、上下引號、句號、問號等...，我們認為，要完整表達文章句意，須以逗號、分號、句號、問號來表示斷句的特性，如此才能把一個句子的完整語意表達清楚，若是將頓號、冒號等停頓標點符號亦視作斷句位置，可能會造成無法表達完整句意的狀況。通常在一個單詞的音節中，不會有停頓或斷句產生，而這也符合了我們在選擇候選合成單元時，並不將無法成為詞的單元考慮進去的初衷；但是，在選擇由單詞組合成的長詞或專有名詞時，是否應該跨越斷句位置，將斷句位置前

的單詞和斷句位置後的單詞組合成一個長詞，並將之視為候選合成單元？在這裡，我們認為，除了聲學上的失真度之外，語意結構上的失真度也該被考量。根據中文語音學的觀點，同一個詞，在不同的語句結構中，它們在聲學參數上的表示會不一樣，舉個例子來說：

(A) 因為女性在社會的地位提高，許多婦女容易面臨工作繁重。

(B) 新研發的產品系列，有助於提高許多行業的工作能力和效率。

在這個例子中，雖然「**提高許多**」皆是由「提高」和「許多」這兩個單詞所構成的長詞，但是就(A)句而言，「提高許多」被斷句所分開，(B)句中這兩個詞則是必須一氣呵成的，所以無論在詞性上，或是語意上，或是能量與音高上，都有著很大的差異。因此，在候選單元的長度限制上，我們以輸入目標文章的停頓位置為逗號、分號、句號、問號時當作分界點，決定斷句的位置跟長度，將輸入目標文章中斷句位置間的句子長度，當作最大的長詞長度，亦即要尋找的目標句；同樣的道理，在大型資料庫中尋找詞與詞的組合當作候選合成單元時，我們也不跨越原本語料庫文字稿(Text)所標記的斷句位置。舉例於下面示意圖；

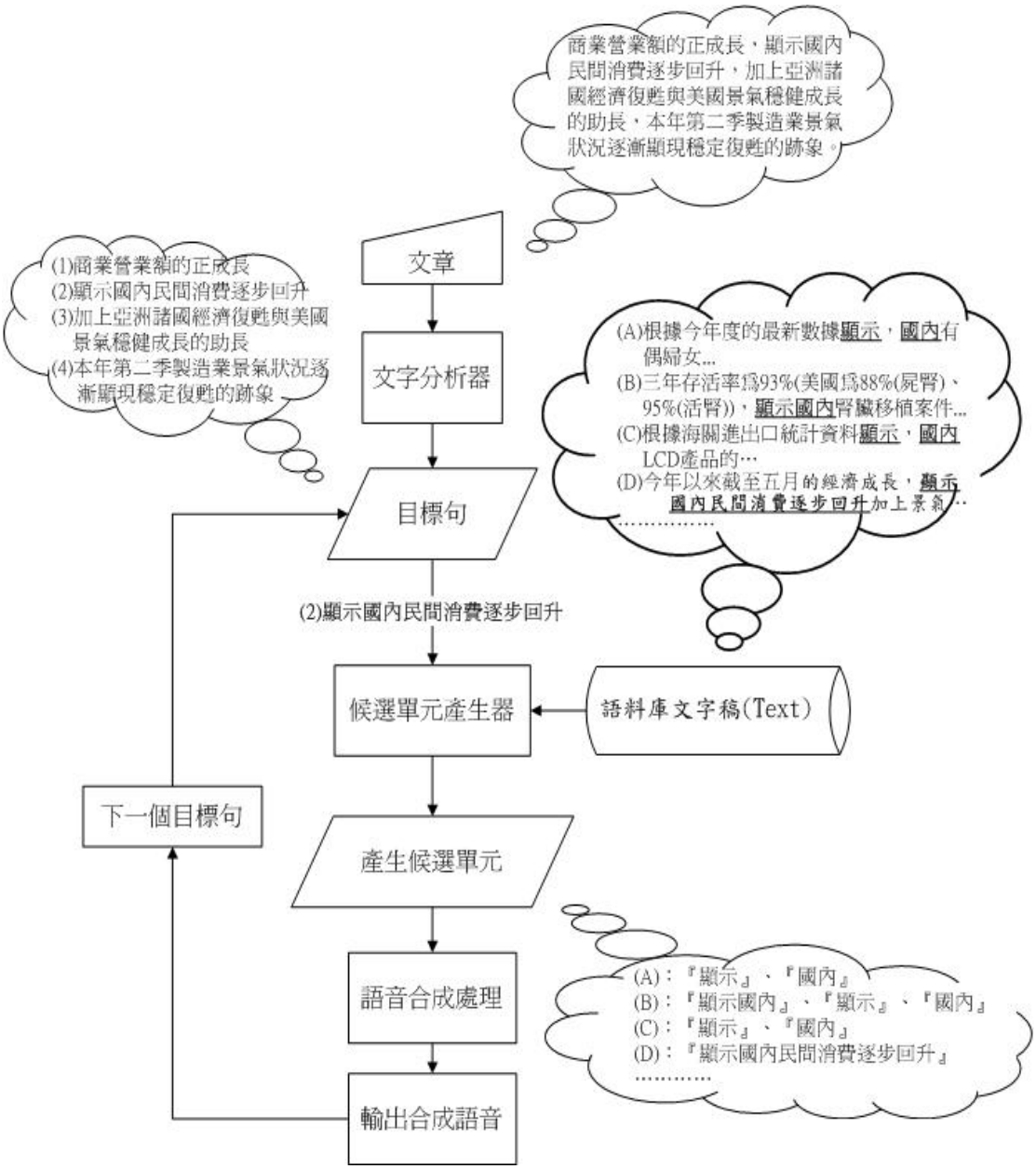


圖 3-4：合成單元選取考量示意圖

3.3.2 構成目標句之選擇流程

確定了目標句後，接下來的工作便是要如何在大型資料庫中找到候選合成單元去構成目標句。一般來說，中文字(character)至少有一萬兩千多個，但是在各種媒體中常見的中文字，只有二至三千個，在本套系統中使用的語料庫，目前

包含有的中文字計有 2659 個，因此，為了要確定有合成單元能構成目標句，我們必須制定一個機制，使得當我們在語料庫中找不到目標句相對應的中文字時，還能構成目標句，進一步才能完成語音合成。

首先，我們知道中文語言學上有四聲變化，因在發音上同音字的很多，所以總共音節大約是 1300 個，如果不計聲調，則會有 411 個不同的基本音節。因此，若目標句出現的中文字太過冷僻或是並不常見，在語料庫中不會出現該中文字的可能性就會極大，此時我們可以利用目標句經由文字分析器產生出來的音節碼，去語料庫中尋找相對應的音節碼，亦即具備有相同聲調，相同基本音節，這種帶有聲調的音節(tonal syllable)，在發音時並不會因為中文字的不同而不同，所以可以將之視為填補目標句的折衷方式；目前我們的語料庫中，帶有聲調的音節共計有 1001 種，所以若以中文總共有音節 1300 種來計算，我們仍有 299 種無法找到。這些不足的地方，我們只好回歸到使用最基本的 411 種單一音節的合成方式，利用 TD-PSOLA 合成器合成找不到的中文字。圖 3-5 為整個構成目標句之選擇流程圖：

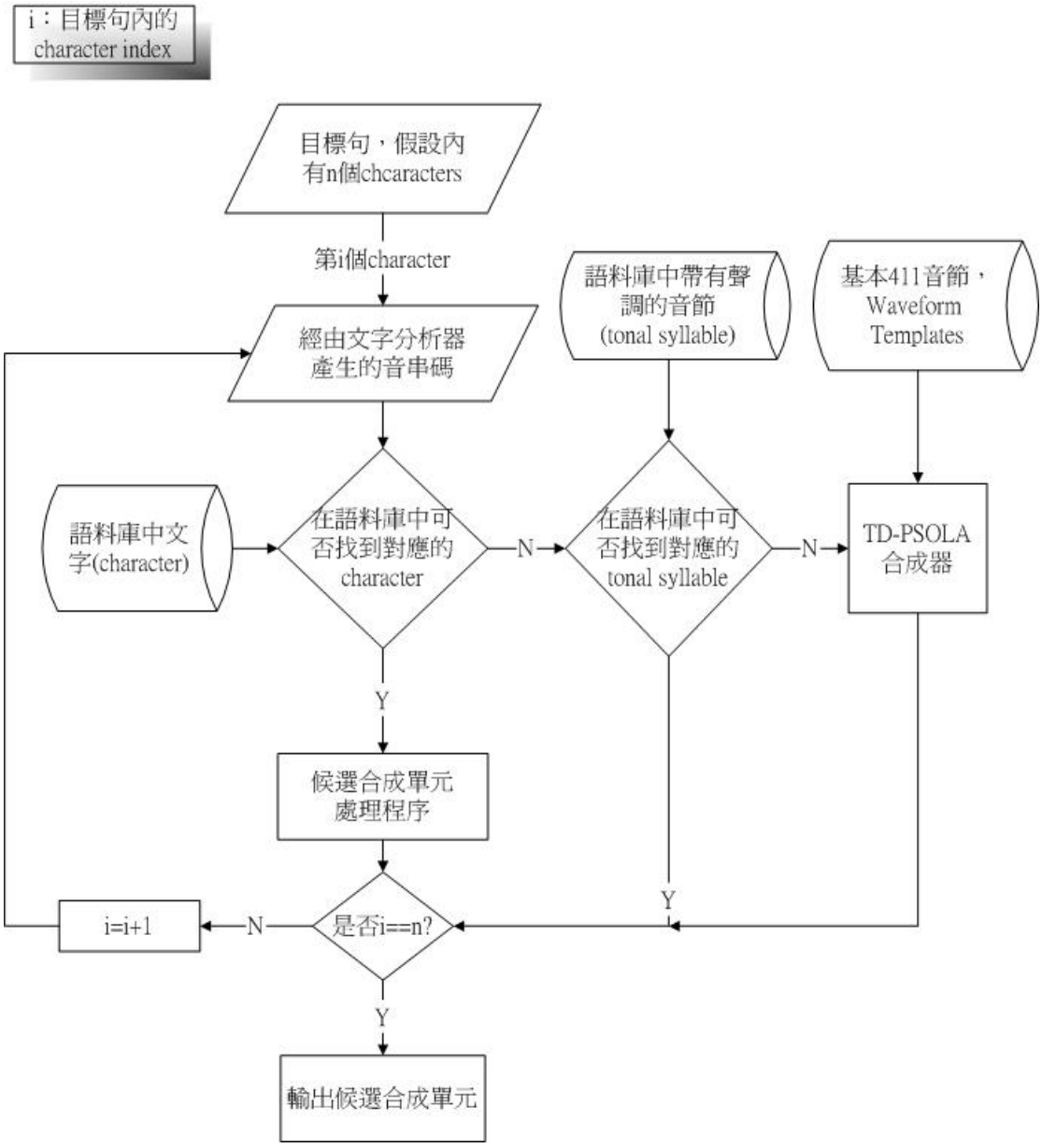


圖 3-5：合成單元選取流程圖

3.3.3 候選合成單元的搜尋機制

在傳統的語音合成系統中，是利用錄製的單一合成音節，以 TD-PSOLA 技術為核心去加以修正樣本音節，藉此可以改善合成語音品質。但是，有經過修飾的波形產生的合成音質畢竟是造成音質下降的主因，因此，現在語音合成系統逐漸由單音節為主的合成架構，轉變成以大型資料庫為主的合成單元架構。一般而

言，採用大型資料庫的系統較單音節為主的合成系統合成品質好，因為它的做法是直接從語料庫擷取所需要的片段來進行串接，並沒有對於聲音做過多的調整，所以原本音質並無遭到破壞，在韻律上的表現當然會較自然。但是，大型語料庫的作法，通常會遇到下述兩項問題：

1. 需要去大型語料庫中做比對：

   如何從大型的語料庫中，快速地去找到最適合片段來做串接，是這方面設計最需要解決的問題。我們在前面已經訂立了要搜尋的目標句，當系統發現在語料庫中有對應的中文字可以構成目標句時，接下來，就是要進入語料庫中做詳細的比對找尋可能的詞片段，並取出當作候選合成單元。然而因為搜尋候選合成單元的比對時間費時，若演算法設計不當，整套系統效能就會低落甚至無法運作，所以發展出一個有效率的演算法來縮短比對時間是相當重要的。

2. 候選合成單元間相串接時韻律差異問題：

   因為候選合成單元是從語料庫各個不同的句子中找到的，所以每個候選合成單元音韻，當然會受它在原句子中前後所相接字的影響。當我們無規則的任意串接它們時，會發生韻律不協調的情形，這會使得聽者明顯感受到語音的不流暢與不舒服。

為解決上述兩點問題，本論文提出連續相關比對法的技巧來降低比對的時間，並在下一節提出一個測量音韻距離的方法，藉此能找到最佳的候選合成單元組合，取出並加以串接使其合成自然度提升。

3.3.3.1 候選合成單元的搜尋比對

首先，大型語料庫的合成系統，必須能在大量的資料中將可能的候選合成單元全部找到，而在實作上，會遇到以下問題：

1. 搜尋目標句所需候選合成單元：

搜尋目標句所需的候選合成單元，除了搜尋的時間問題外，我們希望找到的候選合成單元長度當然是越長越好。當語料庫十分龐大時，我們不可能使用由左往右一一找出最長連續片段的方式，而且當若找到最長片段並取出時，還要從剩下的文句繼續比對語料庫，再找出次長的片段，如此反覆進行，計算量將會非常驚人。在這種情況下，我們可以利用連續相關比對法建立的表格，將所有句子中符合目標句的最長詞串先找出，如此構成這個最長詞串的各個基本單詞因已包含在其中，我們便可以在找到語料庫中所有最長詞串的情形下，相當於已經把語料庫中所有可能構成目標句的詞全部找到。

2. 儲存空間的大小：

目前以大型語料庫為基礎的 TTS 大都是在 PC 上執行，由於硬碟容量日漸增大、語音壓縮的技術也不斷地改良，所以用 PC 實作的 TTS 系統，也會更加的普及與實用。但在運用連續相關比對法時，需要用到大型語料庫的標記資料，因此若大型語料庫增長時，儲存語料庫標記資料的空間亦會加大，在此，我們設計一套能在不浪費空間儲存不必要或重複資訊的情況下，去完成搜尋候選合成單元的比對。

首先，我們建立運用連續相關比對法時所需要的語料庫資料，為了能夠找到符合目標句的中文字，我們建立一個紀錄語料庫中每個中文字在哪一個句子中出現的對照表，稱之為 CLT (Character Location Table)；而且為了能夠找到符合目標句的最長詞串，我們是以語料庫中句子經剖析出來的結構樹為依據，前面已經敘述過，在這棵樹上的每一個終端節點，代表的是一個詞，亦即它具有相當語法資訊，所以我們可以將其詞段及詞段中文字所屬位置紀錄下來，用作尋找最長詞串的依據。下面是語料庫中有的文句，我們先將句子編號並舉例說明：

表 3-1：語料庫中的句子

語料庫中的句子	1.依據 行政院 主計處 的 統計，十月份… 2.非正式 的 統計 顯示，國內 有 … 3.你 如此 不當 的 行為，.. 4.客觀 說 來，其實 原因 很 簡單，… 5…..

根據以上句子，我們將句子中的中文字依照排列順序編號，並將其詞段及詞段中文字所屬位置予以標記。如第一句：依<1,201>據<2,202>行<3,301>政<4,302>院<5,303>主<6,301>計<7,302>處<8,303>的<9,101>統<10,201>計<11,202>…，在這裡<>中前面數字表示該中文字位於句中第幾順位，後面數字表示該中文字在所屬單詞的位置，如「依<1,201>」表示“依”是位於第一句的第一個位置，且屬於“依據”這個二字詞的第一個位置；「處<8,303>」表示“處”是位於第一句的第八個位置，且屬於“主計處”這個三字詞的第三個位置。再如第二句：非<1,301> 正 <2,302> 式 <3,303> 的 <4,101> 統 <5,201> 計 <6,202> 顯 <7,201>) 示<8,202>…，依此類推標記，我們可以將句子編號的資訊和中文字排序及詞資訊整合為<句子編號，中文字排序，中文字所屬詞段位置>，建立以下的中文字位置對照表(CLT)：

表 3-2：中文字位置對照表(CLT)

依	<1,1,201>，…
據	<1,2,202>，…
行	<1,3,301>，<3,7,201>，…
的	<1,9,101>，<2,4,101>，< 3,6,101>，…

統	<1,10,201>，<2,5,201>，…
計	<1,11,202>，<2,6,202>，…
⋮	….

有了語料庫每個中文字的詳細位置及詞資料後，接下來，就要進行找出最長詞串的比對。要比對出最長詞串，也就是當我們依據目標句找到第一個符合的中文字時，繼續往下比對下一個中文字是否仍然符合目標句要尋找的下一個中文字，意即是否具有連續性，例如，我們的目標句若是「想要知道交通大學的基本介紹」，當我們在建立好的 CLT 中找到「想<6,4,201>，<9,13,202>」，就要去比對在<6,4,201>第六句第四個位置後面，即第五個位置是否接的中文字是「要」，且位於詞段的位置是 202，…依此類推，我們將會發現，若是每找一個中文字就要進入句中往下比對，會有以下的缺點：

1. 儲存資訊重複：

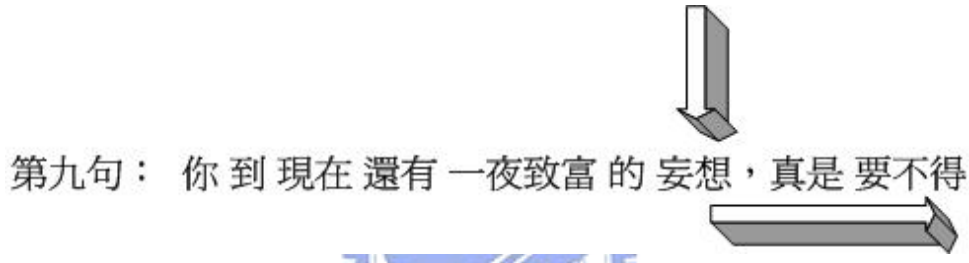
   依據目標句找到第一個符合的中文字時，因為要進入 CLT 中指示的句中位置進行下一個中文字的比對，所以勢必還要再建立一個以句子為單位紀錄中文字及其相關資訊的標記檔，如此紀錄不僅和 CTL 重複，而且語料庫的資料龐大，出現過的中文字勢必還會重複出現很多次，所以在 CLT 中只要紀錄一次出現過的中文字，但在以句子為單位紀錄的檔案中，就要每個中文字都要紀錄，這會儲存很多重複的資訊並且浪費空間。例如語料庫中最常出現的中文字「的」，出現在語料庫中達 1,828 次，若在 CLT 中只紀錄一次，在以句子為單位紀錄的檔案中，就要紀錄 1,828 次，這可是有 1,828 倍之差啊！

2. 比對時間浪費：

   每次找到一個有在 CLT 出現的中文字，只要是不屬於前面找到中文字句中具有連續性的，也就是和前面中文字無相關的部分，就要進入句子中往下

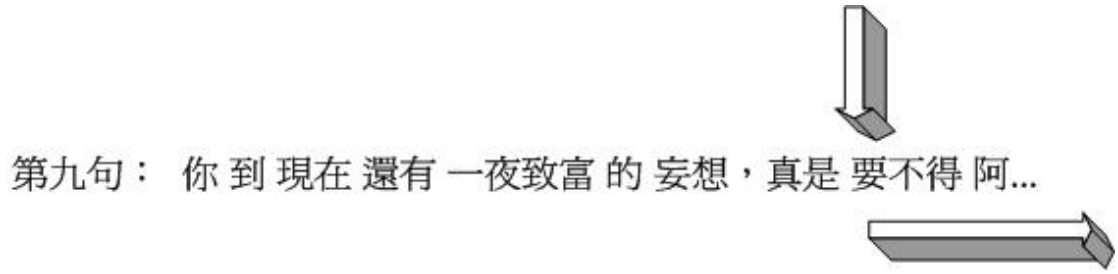
比對，如此會拉長比對的時間。例如目標句為「想要知道交通大學的基本介紹」，我們在 CLT 找到了「想<6,4,201>，<9,13,202>」：

第六句中，因為從「想」開始往下比對後，只有到「要」之後就和目標句文字不符合了，所以本句只搜尋到「要」即停止。



第九句中，因為「想」開始往下比對後，發現並沒有和目標句下一個中文字「要」相同，所以並無法成詞輸出。

接下來，就要繼續尋找目標句的下一個字「要」，隨即在 CLT 找到了「要<6,5,202>，<9,17,101>，….」，我們會發現，第二個「要」出現的位置是在第九句，且對於「要」而言，雖然是第一次進入第九句，但是，到目前為止，第九句已經重複進入了兩次了,可想而知，若對於整個目標句而言，重複進入同一句去搜尋的機會是很大的。



基於上述兩項缺點，我們將改善到只單純利用 CLT，設計一套有效率的比對

法，因為採用的方法，是基於上述觀念，亦即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。

3.3.3.2 連續相關比對法

當語音系統的目標句輸入候選合成單元產生器後，候選合成單元產生器就必須立刻從語料庫中找到相同的文字片段，並加以取出。而在片段需在適當的位置開始及結束才符合稱之為詞的前提下，找到的文字片段的原則是每個片段的字數越多越好。為了避免計算量太過龐大，我們可以利用連續相關比對法建立的表格，將所有句子中符合目標句的最長詞串找出，如此構成這個最長詞串的各個基本單詞亦相對地被找出，我們稱上述表格為 Word Sequence Candidate Table (WSCT)。利用此表，我們就可以知道語料庫中所有可能構成目標句的詞有哪些，舉例如下表所示：

目標句：想要知道交通大學的基本介紹

語料庫中出現的相關句子，有標示底線部份為和目標句有部分詞相同之處：

表 3-3：候選合成單元範例

語料庫中出現的相關句子	1. 我 不 大 想要 結婚 2. 交通 大學 是 一所 國立 大學 3. 應徵 面試 需要 個人 的 基本 資料 4. 透過 林先生 的 介紹 5.上 高速公路 前 我 想要 知道 交通 狀況 6.…

利用連續相關比對法最後得到的 WSCT 顯示如下：

表 3-4：連續相關比對法之 WSCT 範例

句子#	在目標句起始位置	在句中起始位置	詞串組合 pattern
1	1	4	2
2	5	1	2,2
2	7	10	2
3	9	9	1,2
4	12	7	2
5	1	8	2,2,2

上列表中，第一行表示在語料庫中的第 1 句，「我 不 大 想要 結婚」中的第 4 個字開始，有比對到從目標句「想要知道交通大學的基本介紹」的第 1 個字開始往後的一個詞，這個詞是屬於 2 字詞。再繼續往下看到表格中紀錄在第五句有發現詞串的地方，表格內容表示在語料庫中的第 5 句，「上 高速公路 前 我 想要 知道 交通 狀況」中的第 8 個字開始，有比對到從目標句「想要知道交通大學的基本介紹」的第 1 個字開始往後的詞串是由三個詞組合的，這三個詞分別為 2 字詞+2 字詞+2 字詞，所以根據這個紀錄，我們可以知道在語料庫第 5 句中，除了有「想要 知道 交通」這個最長詞串，也知道有「想要 知道」、「知道 交通」這個次長詞串，以及「想要」、「知道」、「交通」這三個基本詞。所以根據此一表格，我們就可以很快的找到語料庫中符合目標句的所有詞串了。

當然，在我們比對的過程中，並不是一開始就可以馬上找到符合目標句的最長詞串紀錄在 WSCT 中。首先，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，在比對的過程中，我們使用 CLT 資訊，依據目標句中文字的次序處理，WT 的格式及比對規則如下：

表 3-5：連續相關比對法之 WT 範例

句子#	前一個 char 在目標句中位置”	在目標句起始位置	在原句中起始位置	已完成詞串組合 pattern
5	4	1	8	2,2

若以前面例子語料庫的第五句來做舉例說明，目標句是「想要知道交通大學的基本介紹」，第五句是「上 高速公路 前 我 想要 知道 交通 狀況」，在上面的 WT 中，“前一個 char 在目標句中位置”表示前一個比對程序處理到目標句的第幾個中文字，這裡是指在語料庫中第五句中，目前已經比對到符合目標句的第 4 個字「道」，再繼續往表右邊看，便可以知道在第五句的「道」為止，其實是從第 5 句的第 8 個字「想」之後，就有符合目標句第 1 個字開始往下比對相同中文字進一步組成的詞串，是由 2 字詞+2 字詞組合而成，亦即從第 5 句第 8 字之後有「想要 + 知道」這個詞串。

接下來整個運用 CLT 搜尋最長詞串的比對規則，處理程序如下：

1. 從目標句的第一個中文字開始，因為我們所要尋找的是詞串，所以必須要先從中文字所屬詞段位置的詞首開始。首先，將在 CLT 中紀錄目標句第一個字中的所有屬於詞首(即中文字所屬詞段位置是 x01)的資訊加入 WT。
2. 從目標句的第二個字到目標句的最後一個字，分為兩種情形：
   (1) 和目標句上一個中文字有連續相關性的：

   當我們完成了目標句的第一個中文字的 WT，繼續往下一個中文字比對時，我們亦先在 CLT 中找到有關該中文字的<句子編號，中文字排序，中文字所屬詞段位置>，利用這三個資訊，以及由 WT 中“原句中起始位置”和 “已完成的詞串組合 pattern”，可以知道在原句中，之前已經比對到該句中的哪一個字位置。首先，我們將 CLT 找到的資訊和 WT 比對

是否和前一個中文字隸屬於同一個句子，若在同一個句子中，隨即可再繼續比對在該句中的中文字排列是否和 WT 中紀錄的有連續性。又因為所有的詞串要第一次置入 WT 時，都是以詞首為考量，所以若是在同一句中中文字排列是有連續性的，那麼，中文字在詞中的位置亦具備有連續相關性。要注意的是，判斷是否成詞的依據是在於 CLT 當中紀錄的“中文字所屬詞段位置”資訊，當這一個詞段比對到詞末時(中文字所屬詞段位置是 a0a)時，才能判定其為理想的詞，也才能夠加入 WT 中“已完成詞串組合 pattern”這個項目的內容，其餘若只比對到詞段中間，都不能成詞輸出。例如「想」位於「想要 知道...」這句的第一個字，若在語料庫第五句中找到「想」，並屬於詞首(中文字所屬詞段位置是 201)，當我們找下一個字「要」(中文字所屬詞段位置是 202)，亦在第五句時，若檢驗其對於「想」具備中文字排列連續性，就不用再去比對中文字在詞中的位置是否連續(201-202)，而且其所屬詞段位置是 202，表示整個詞已經比對完畢，所以可以納入 WT 中“已完成詞串組合 pattern”，此例為 2。如此依據 CLT 與 WT 相互比對，並且不斷更新 WT 內容，一直到目標句的最後一個字。

(2) 和目標句上一個中文字沒有連續相關性的：

依據在 CLT 中找到有關該中文字的 <句子編號，中文字排序，中文字所屬詞段位置>，除了有(1)的情形外，還有找到的資訊是完全新出現的句子或是和 WT 相互比對後沒有連續相關性，遇到這種情形，我們只要將其視為如同一個新的開始即可，亦即將這些是屬於詞首(即中文字所屬詞段位置是 x01)的，新加入 WT 中，當作尋找最長詞串的起頭。

3. 在比對過程中經過處理程序 2 的連續相關比對後，若遇到不能再繼續往下比對如(1)的情形，或是並非如(2)所規定的情形時，要隨時將 WT 中無法更新的內容且其中具有成詞部分的資料輸出至 WSCT，其餘不符合的資料予以放棄。

如此程序依序往下比對至目標句的最後一個字，整個比對過程就可以結束，我們便可以將 WT 中“已完成詞串組合 pattern”這個項目中有成詞的部分取出，也就是說我們已經在語料庫中找到了符合目標句的所有詞串，最後再將 WT 中最長詞串及其相關資訊填入 WSCT 中，連續相關比對法示意如圖 3-6 所示。

由示意圖可知，當我們依照連續相關比對法進行比對時，比對時間是和 WT 和 CLT 的資料多寡成正比。如果從示意圖的例子來看，此時 WT 有 36 個候選單元，CLT 在 43 個位置找到目標中文字，因此，此例中要比對的次數為 36x43=1548。也就是說，如果有一個中文字在語料庫中出現的次數相當頻繁，如此在 CLT 中紀錄它出現位置的次數就會是相當的龐大，那就意味著，當我們在做連續相關比對時，所需要的比對次數增多，比對的時間就會相對的被拉長，這並不是我們所樂見的，例如語料庫中最常出現的中文字「的」，出現在語料庫中達 1,828 次，若在 CLT 中紀錄其全部的位置，就要紀錄 1,828 個，而它在比對時更是會耗費 nx1,828 的時間(n：前一個中文字前在 WT 列出的候選單元次數)，針對這個問題，我們在上述的連續比對流程中加入了一些技巧來減少遇到此類情形的比對時間。

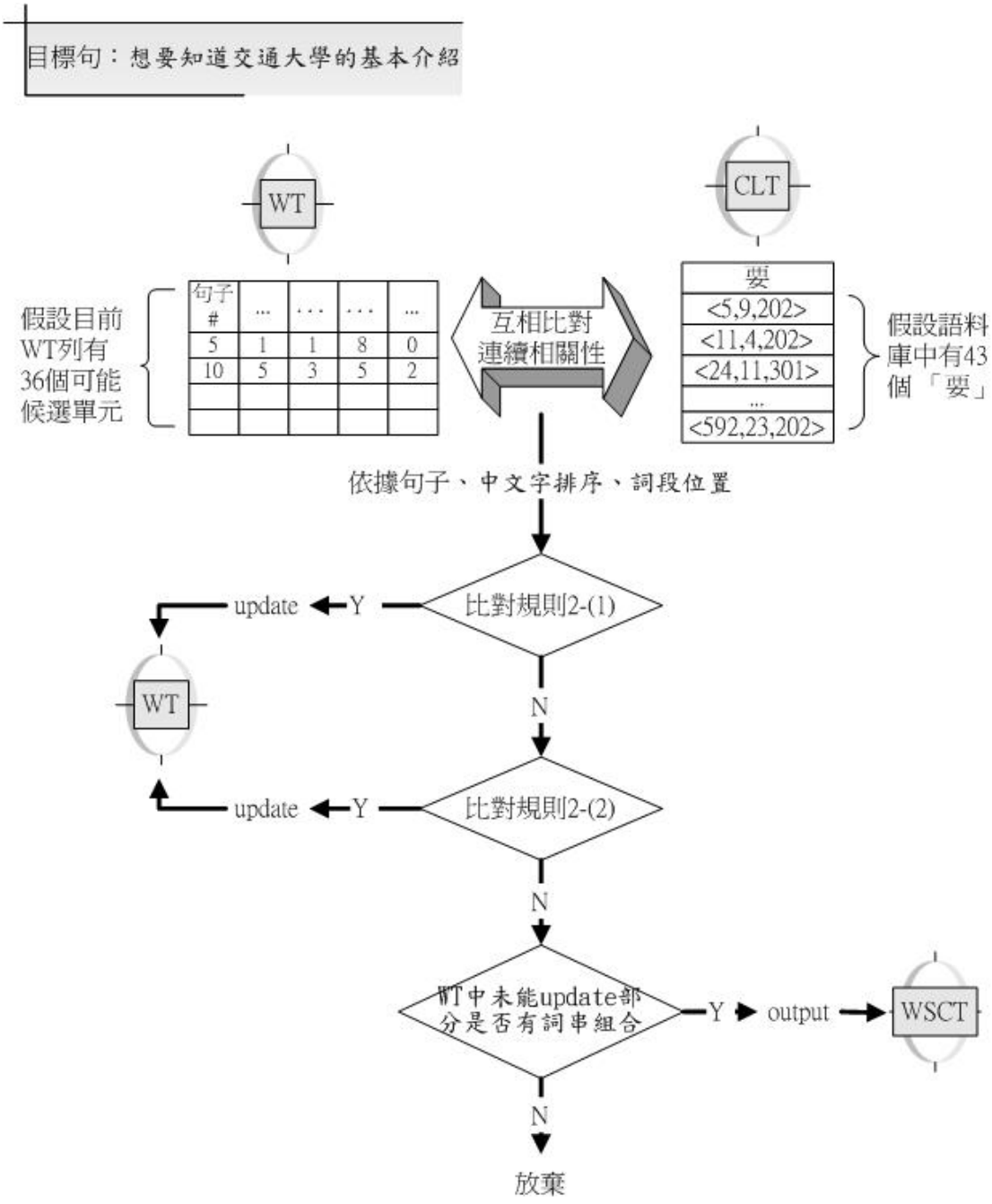


圖 3-6：連續相關比對法示意圖

3.3.3.3 建立常用字額外比對

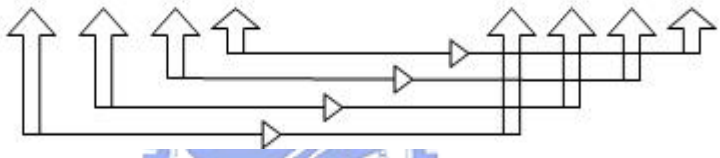
在前面我們提出了對於在語料庫出現次數頻繁的中文字，會導致比對時間的拉長，所以我們必須將這些字找出並做一些特殊的比對處理。首先，我們發現這

種出現次數頻繁的字，通常是在一個句子中重複出現的連接詞或是一字詞，例如「的」、「是」、「在」、「有」...，在本系統使用的語料庫中，我們統計了出現次數最頻繁的前十個，當作要做特殊比對的常用字，這些字計有「的」、「是」、「在」、「一」、「人」、「有」、「不」、「我」、「以」、「他」。

接下來，我們建立一個以句子為單位的表，稱之為 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，FCLT 表如下：

表 3-6：連續相關比對法之 FCLT 範例

句子#	在句子中出現規定的常用字	常用字在句中出現的排序位置
7	他、的、有、的	1、9、12、15



若語料庫中第七句是：他 覺得 女性 不孕症 的 人數 有 增漲 的 趨勢，上表表示句中「他」、「的」、「有」是屬於我們定義的常用字並且亦為一字詞，因此我們把本來要紀錄在 CLT 中的「他」、「的」、「有」轉而紀錄在上述 FCLT 中，而其餘句中的中文字資訊依舊紀錄在 CLT 中，並且因為我們定義這十個常用字是要當其屬於一字詞時才將它歸類在 FCLT，所以我們並不需要在 FCLT 中紀錄其所屬詞段位置。以上表為例，第七句單就中文字看來，雖然有「他」、「不」、「的」、「人」、「有」五個字屬於常用字，但是因為「不」屬於「不孕症」這個詞的詞首，「人」屬於「人數」這個詞的詞首，皆非屬一字詞，故將其歸屬於 CLT 中；而其它的「他」、「的」、「有」皆屬於一字詞，故將其紀錄在 FCLT 中。

建立 FCLT 的作法有以下幾個優點：

1. 減少常用字在 CLT 的儲存空間：

原本要分別紀錄 <句子編號，中文字排序，中文字所屬詞段位置> 等常用字資訊在 CLT 的空間，現在只要以句子為單位紀錄紀錄位置即可，當語料庫很龐大，且其中常用字出現次數很頻繁時，這可減少一些 CLT 的儲存空間。

2. 減少比對的時間：

建立了 FCLT 後，要如何利用此表去連續比對相關性呢？因為 FCLT 都是以一字詞作為考量，且有紀錄字在句中的位置，所以我們可以利用 FCLT 和 WT 去作連續相關檢驗，檢驗情形分為下述兩種：

(1) 依據 WT 內容的往後檢驗：

當比對程序進行到的中文字屬於常用字時，我們可以根據 WT 已經記載的資訊，直接依照 WT 有候選單元記載的句子，去 FCLT 中的該句往後比對是否有目標常用字緊接其後，舉例如圖 3-7。從圖 3-7 所舉例子看來，只需要比對 WT 中原有的候選單元個數 24 次，但若沒有 FCLT，單純依照以往只有 CLT 的做法，則會有 24x1,828=43,872(1,828 是「的」出現在語料庫中的次數)的比對次數。

(2) 依據 CLT 內容的往前檢驗：

另外，當比對程序跨過常用字並進行到連續下一個要比對的中文字時，當其進入比對程序屬於 2-(2)階段，就會有新的詞首候選單元填入 WT 中，此時，便會有一種情形產生，即在要填入的詞首候選詞前面會有出現是跨過常用字的可能，所以在填入詞首的同時，要同時往前檢驗候選單元在原本句中位置前的中文字是否屬於常用字，再將最後檢驗結果填入 WT 中，舉例如圖 3-8。

由上述(1)(2)可知，CLT 再加上 FCLT 的檢驗方法，會比原先只利用 CLT 的比對方法節省掉一些比對的時間。

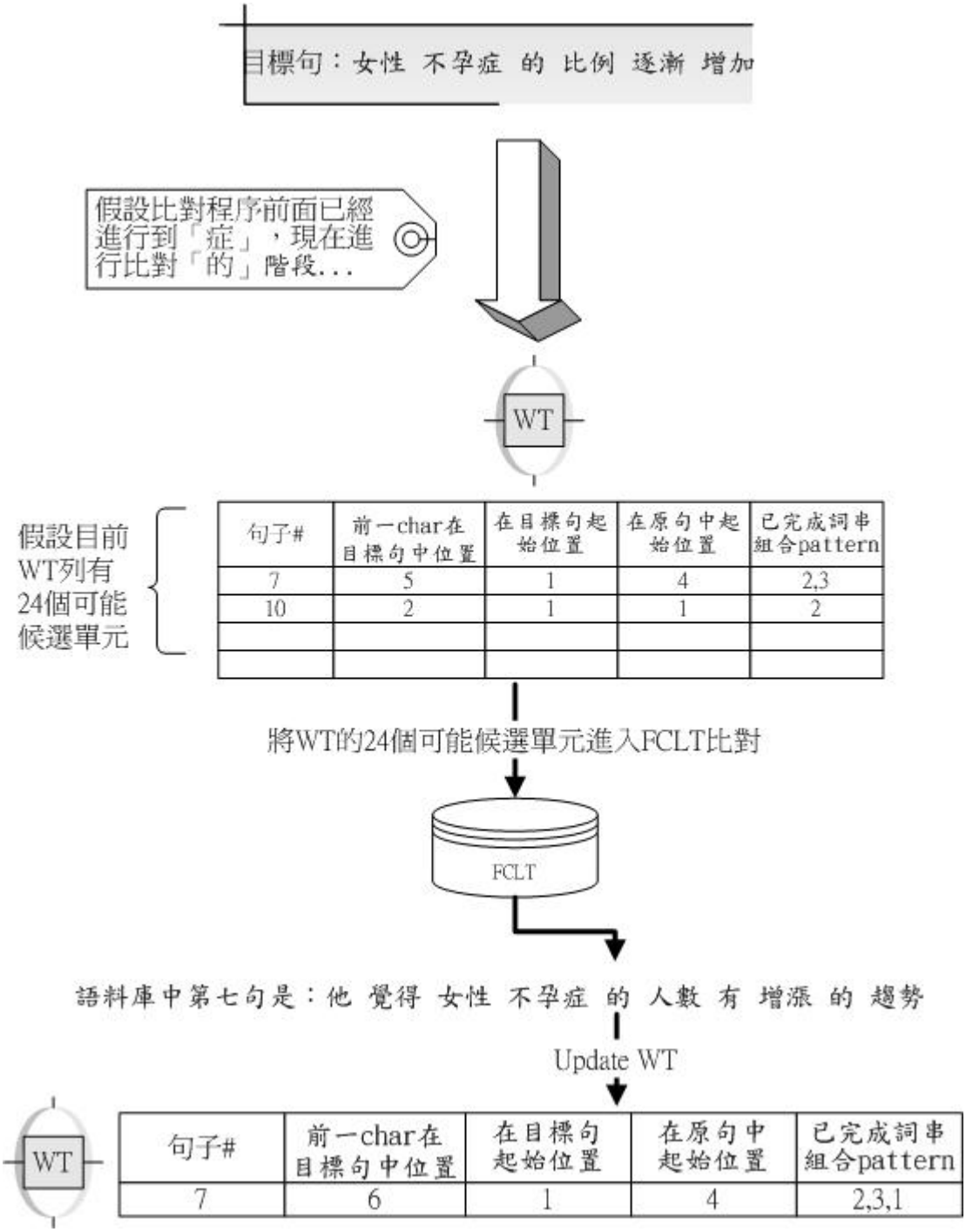


圖 3-7：使用 FCLT 往後檢驗作搜尋之範例

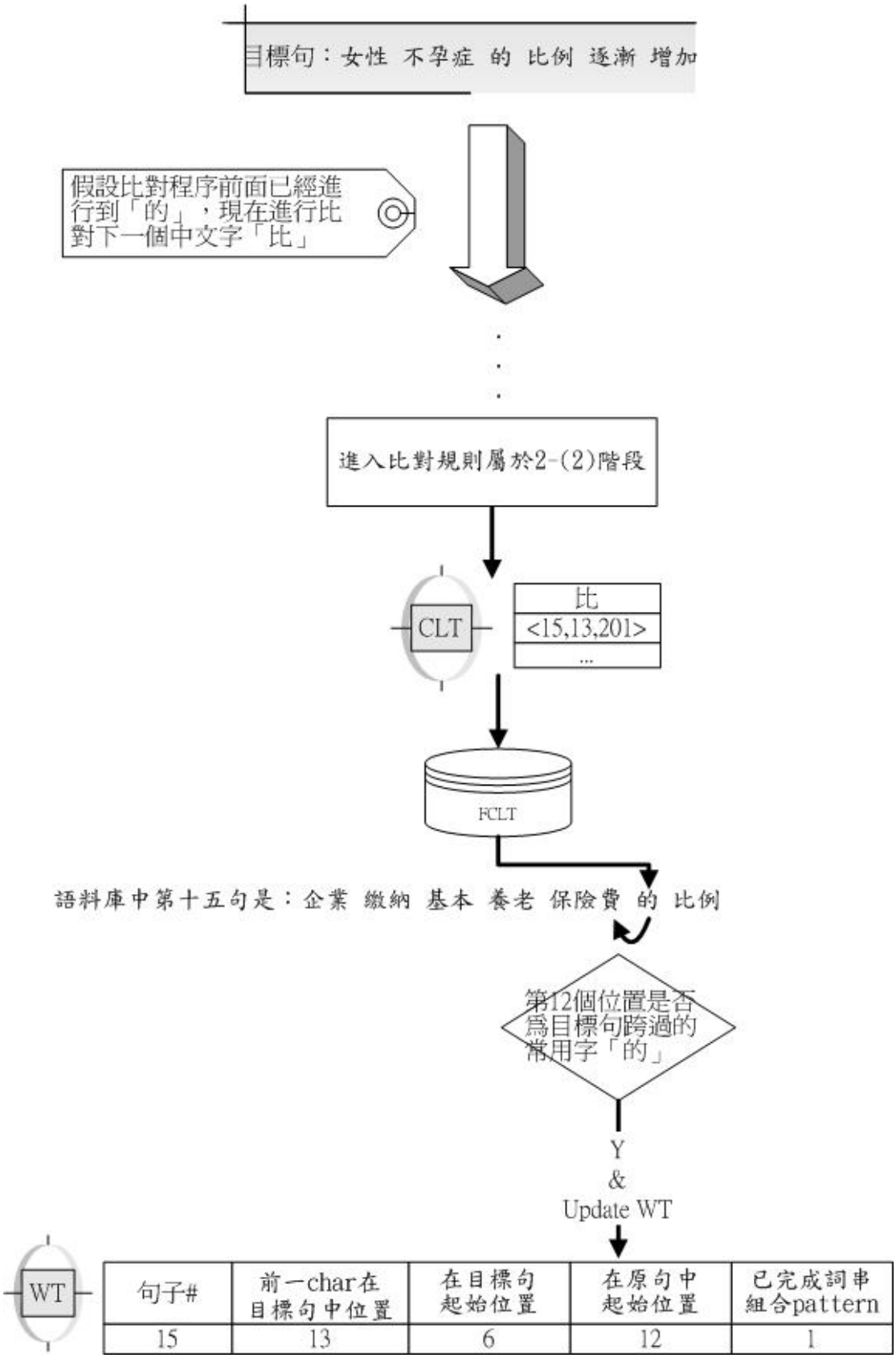


圖 3-8：使用 FCLT 往前檢驗作搜尋之範例

3.3.3.4 Unit Lattice 的形成

利用連續相關比對法，我們已經將所有句子中符合目標句的最長詞串找出，如此構成這個最長詞串的各個基本單詞因已包含在其中，我們便可以在找到語料庫中所有最長詞串的情形下，相當於已經把語料庫中所有可能構成目標句的詞全部找到，亦即候選合成單元已經找出。若將這些候選合成單元以 units 稱之，則利用所有找出的 units 來構成目標句，可以建立出一個 Unit Lattice，這個 Lattice 蘊含了所有可能由候選合成單元構成目標句的排列組合，舉例如下圖所示。下一步，就是要探討如何在這個 Lattice 中，找出最佳的組合方式，來解決候選合成單元內音韻與候選合成單元間相串接時韻律差異的問題。

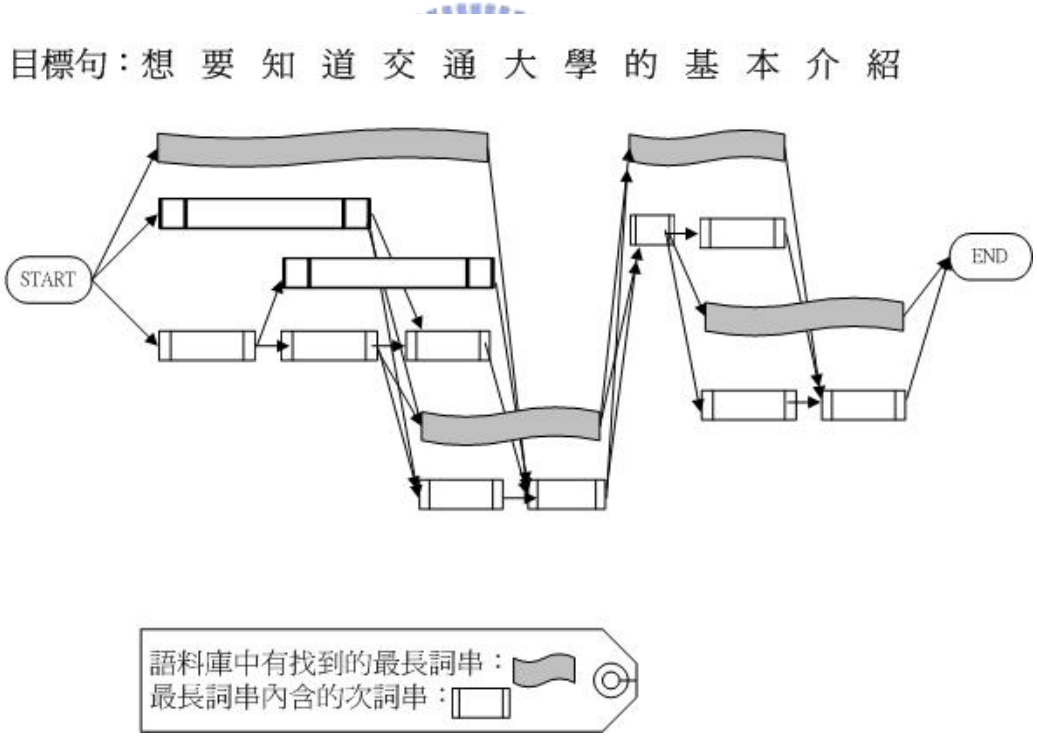


圖 3-9：語料庫搜尋產生候選詞 lattice 之範例

3.4 合成單元的選取

在大型語料庫中搜尋到候選合成單元 unit 後，我們並非直接從中任意選取來做串接的動作，因為這樣的串接會造成韻律不協調的情形。造成韻律不協調的原因包括有，第一：找出來候選合成單元內的韻律和由韻律產生器產生的目標音韻可能會有落差，第二：串接前後候選合成單元間會有韻律不連續的現象。所以我們必須在所有的候選合成單元中找到可以減少這兩個問題的最佳組合，以期能合出流暢自然的語音。

3.4.1 文獻回顧

本節研究的一個重點是，設計一種挑選合成單元的方法，以選出最適當的合成單元來進行串接合成自然流暢的語音。使用大型的語音資料庫可以含括較多的韻律及頻譜串接組合類型，再從其中挑選出適合的單元以滿足平滑轉移的目標。從大型語料庫中挑選合成單元的方法，國外研究者發表的論文如 Wang[16]、Hunt [17]、Chu[18]、Toda[19]；而在國內方面，陳昭宏[7] 、周福強[20]也分別在他們的博士論文中提出有關於選取合成單元的方法。從這些研究中，我們可以將常見的單元挑選方法歸納為兩類，一類是以 cost function 作為選取候選單元的依據，前述文獻採用此法的有[Hunt, et al., 1996]、[Chu, et al., 2002]、[Toda, et al., 2002] 、[陳昭宏,1997]，另一種方法則是採用決策樹(Decision Trees) 來選擇候選單元，如[Wang, et al., 1993]、[周福強，1999]。接著，我們就分別簡單說明此兩類的方法。

(1)依據 cost function 挑選：

此類方法主要是去計算語料庫中的候選單元，當它被選用時所導入的 cost (通常為誤差)數值，調整權重值，採用動態規劃演算法來挑選出一個句子整體上

最佳的路徑，使得挑出的合成單元序列具有最小的累積 cost 值。在[Toda, et al., 2002]的研究中，將誤差細分為四個方面---聲韻環境變動誤差(cost on substitution of phonetic environments)、頻譜連續性誤差(cost on spectral discontinuity)、韻律誤差(prosody cost)、頻譜中心誤差(cost on spectral centroid)等，前兩者是屬於串接上的誤差，後兩者則是屬於候選單元與目標合成語句的頻譜與韻律相似度之誤差。另外的幾篇論文提到的誤差量測基準以及對應之誤差函式的定義，雖然有各自的見解，但主要的想法均可大致對應到上述四項誤差成分之中。

(2) 依據決策樹挑選：

第二類的合成單元挑選方法是利用決策樹的方法，在語音合成領域中，決策樹常被用來做語音單元分群或是用以產生韻律參數，決策樹可分為兩種,分別為分類樹(classification trees)以及迴歸樹(regression trees)。分類樹的目的在於區分成一離散的類別數值，而迴歸樹的目的則在分析已得到連續型參數值，至於在挑選合成單元的應用上，決策樹主要是依據語言參數(linguistic features)來挑選合成單元，其中語言參數包括韻律參數、詞邊界、呼吸邊界等，因此較近似於迴歸樹的應用方式。實作上的一種做法是，使用一組遞迴式分割演算法(recursive partitioning algorithm)，建構決策樹時，對每個節點測試分裂的條件，去計算每一個節點分裂所對應的誤差值，可以是計算聲學或韻律方面的誤差，並據以選擇適合的分裂條件來最小化誤差值，然後分裂成兩個子樹，之後再繼續遞迴上述分裂流程並產生出整棵決策樹。最後，依據每個節點上的分裂條件至樹葉中搜尋出最佳的合成單元。

本論文提出要從所有的候選合成單元中挑選適當合成單元的作法，主要是使用 cost function，將於下一節作詳細說明。

3.4.2 挑選合成單元的方法

兩種型態的失真(distortion)在候選合成單元串接時會被考量，一種是在候選合成單元內部的韻律參數差異(intra-unit cost)，一種是候選合成單元間的連接差異(inter-unit cost)。

3.4.2.1 Intra-unit cost

合成單元除了考慮本身的發音與韻律特徵外，也考慮由其前後發音與韻律特徵所共同定義的單元(contextual unit)。因為候選合成單元是從語料庫各個不同的句子中找到的，所以每個候選合成單元的音韻，當然會受他在原句子中前面後面所相接字的影響。所以挑選出來的合成單元，必須具有與合成目標有最相近的語音及韻律特徵，若在發音及韻律上都符合合成目標的需求，則在合成時對訊號的修飾就可以盡量避免，合成出來的語音自然會較流暢。

在量測候選合成單元內部的韻律參數差異時有三項主要考量因素被使用，包括有詞串字數、語意參數、以及韻律參數。

$$\hat{S}_i = L_1^i L_2^i L_3^i ... L_n^i \tag{3-1}$$

上式表示第 i 種可能的組合組成目標句，而這個組合是由 L_1^i 、 L_2^i 、L_3^i …等 n 個 units 組成的 。對於每個組成目標句的 unit，我們分別將其 Intra-unit cost 表示為：

$$C_{intra}(L_k^i) = \frac{w_l D_l(N_k^i)}{Q_l} + \frac{w_s D_s(S_k^i)}{Q_s} + D_p(P_k^i, P_t\ ) \tag{3-2}$$

其中各項變數定義為

N_k^i ：第 i 種可能組合中第 k 個 unit 內的詞串字數

S_k^i：第 i 種可能組合中第 k 個 unit 內的語意參數

P_k^i：第 i 種可能組合中第 k 個 unit 內的韻律參數

P_t：由韻律產生器產生和 unit 相對應文字的韻律參數

intra-unit cost 各項，分別表示為：

◆ $D_l(N_k^i)$：第 k 個 unit 內的詞串字數在整句目標句中所佔字數的比重

$$D_l(N_k^i) = (\frac{N_k^i}{N_{total}})^2 \qquad (3\text{-}3)$$

其中 N_{total} 表示目標句的總字數。

◆ $D_s(S_k^i)$：第 k 個 unit 內的語意分數

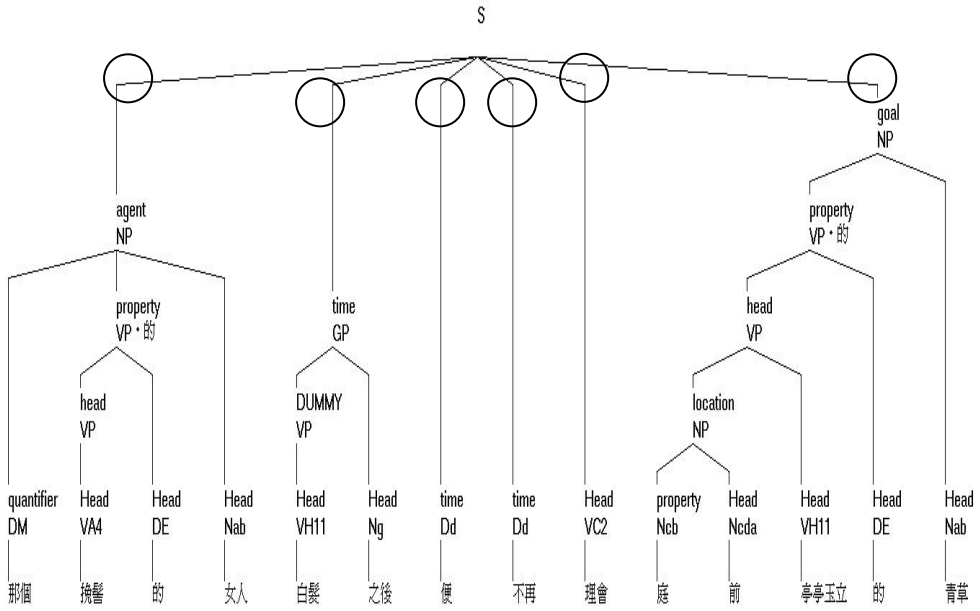


圖 3-10：語料庫之句子的語法樹範例

在語料庫中，每一個句子都依據句法結構被剖析成結構樹，我們可以觀察到，在結構樹的最高層分枝處，可以說是在目標句中最佳的停頓呼吸處，而且以最高層往下看的分支，也會發覺在最左邊的分枝是屬於較佳的開始點，而分枝的

最右邊是屬於較佳的結束點。例如上例中第一個最高層所分出的次子句是「那個挽髻的女人」，而他的最佳開始點是位於分枝的最左邊，也就是「那個」，最佳的結束點就是「女人」；也就是說，當我們只取出「的女人」時，其所代表的語氣停頓並不完整，我們便要給予較少的分數。

◆ $D_p(P_k^i, P_t)$：第 *k* 個 unit 與欲合成目標的韻律差異

在計算 Intra-unit cost 時，我們利用在 3.1 中提到的韻律產生器，來產生輸入文句的韻律參數，包含有四個基頻軌跡參數、三個時間長度參數、和一個能量參數，這就是我們的目標韻律，其中計算第 k 個 unit 與欲合成目標的韻律差異表示如下：

$$D_p(P_k^i, P_t) = -\frac{1}{m}\sum_{j=1}^{m}\left(\frac{w_{f_0} d_{f_0}^j}{Q_{f_0}} + \frac{w_d d_d^j}{Q_d} + \frac{w_e d_e^j}{Q_e}\right) \quad (3\text{-}4)$$

m 表示在第 *k* 個 unit 中含有幾個中文字，而其中

$$d_{f_0}^j = \left(\frac{\bar{F}_j}{\bar{F}_{target}} - 1\right)^2 \quad (3\text{-}5)$$

$\bar{F}_j$ 表示在第 *k* 個 unit 中的第 *j* 個字的平均基週(pitch mean)，$\bar{F}_{target}$ 則是由韻律產生器相對於同一個中文字所求出的平均基週。

$$d_d^j = \left(\frac{U_j}{U_{target}} - 1\right)^2 \quad (3\text{-}6)$$

訂定一音長(duration)失真度量測，所謂音長包括有聲母加上韻母的長度，其中U_j表示第 *k* 個 unit 中的第 *j* 個字的音長，U_{target} 表示目標音長。

$$d_e^j = \left(\frac{E_j}{E_{target}} - 1\right)^2 \quad (3\text{-}7)$$

求取每一個中文字的最大能量，定義能量失真度，為第 *k* 個 unit 中的第 *j* 個字的

最大能量 E_j 與目標能量 E_{target} 間的差距。

- $w_l, w_s, w_{f_0}, w_d, w_e$ ：相對各項變數的權重
- $Q_l, Q_s, Q_{f_0}, Q_d, Q_e$ ：相對各項變數的正規化(Normalize)值

因為每項變數各具有一定的範圍與數值，而且相差頗多，正規化的目的是將其作等比例的放大或縮小，使其都落於同一個範圍中，如此在調整權重時，方能依據變數參數在挑選合成單元時佔有的重要性來作適當的比重規劃。

綜合以上詞串字數、語意參數、以及韻律參數的差距，取其正規化並給予各項適當的權重，其總合即表示為候選合成單元內部的韻律失真度。

3.4.2.2 Inter-unit cost

挑選出來的合成單元，除了候選合成單元內部必須具有與合成目標最相近的語音及韻律特徵之外，還必須要能在串接後保持整句音韻流暢自然，因此我們亦要注意候選合成單元間的連接差異。首先，要使得候選合成單元間能夠在連接後音韻流暢自然，就必須在候選合成單元的連接處基頻軌跡與能量軌跡都能平滑(Smooth)轉移，這部份我們發現，當我們在使用由韻律產生器產生出來的韻律參數時，其實韻律產生器在以遞迴式類神經網路(Recursive neural network, RNN)的概念訓練時，產生的韻律訊息就是平滑轉移的訊息了，所以當我們在計算韻律失真度時，就已經將平滑轉移考量進去了。除此之外，串接時候選合成單元間是否有連音(coarticulation)效應，亦是造成候選合成單元間是否平滑轉移的重要因素。

對於合成語音而言，因為句子是由不同的候選合成單元組合而成，所以若是取出的候選合成單元有連音效應，會在候選合成單元接合處聽到明顯的雜音，而且在連音處聲音也會有不平滑相接的感覺，因此，對於這種具有連音效應的候選合成單元我們並不希望將他視為最優先的合成單元。

$$C_{inter}(L_k^i) = D_c(O_l^k, O_r^k) \quad (3\text{-}8)$$

上式表示在第 i 種可能組合中第 k 個 unit 時的連接差異，我們用第 k 個 unit 在原本屬於的語料庫句子中，是否具有連音效應來當作是否適宜作合成語音候選合成單元前後連接的依據。對於一個候選合成單元而言， 連音效應會對其有兩種影響：一是候選合成單元的左邊連接失真 O_l^k ，一是候選合成單元的右邊連接失真 O_r^k 。候選合成單元的左邊連接失真，是指在候選合成單元的聲母一開始和它前一個字的韻母結束有連音效應，而候選合成單元的右邊連接失真，則是指候選合成單元的韻母結束與連續相接其後的字聲母開端有連音效應。

本系統中，判斷連音效應的方法為利用前後中文字的交界處的能量下降趨勢 (energy dip)，我們訂出一個門檻值，當交界處能量並沒有降至前後相接中文字五個音框 (5 個 frame，亦即 50ms)之平均能量的 1/10 倍時，我們可以將他視為具有連音效應。如下為 energy dip 很小，具有連音效應的範例圖：

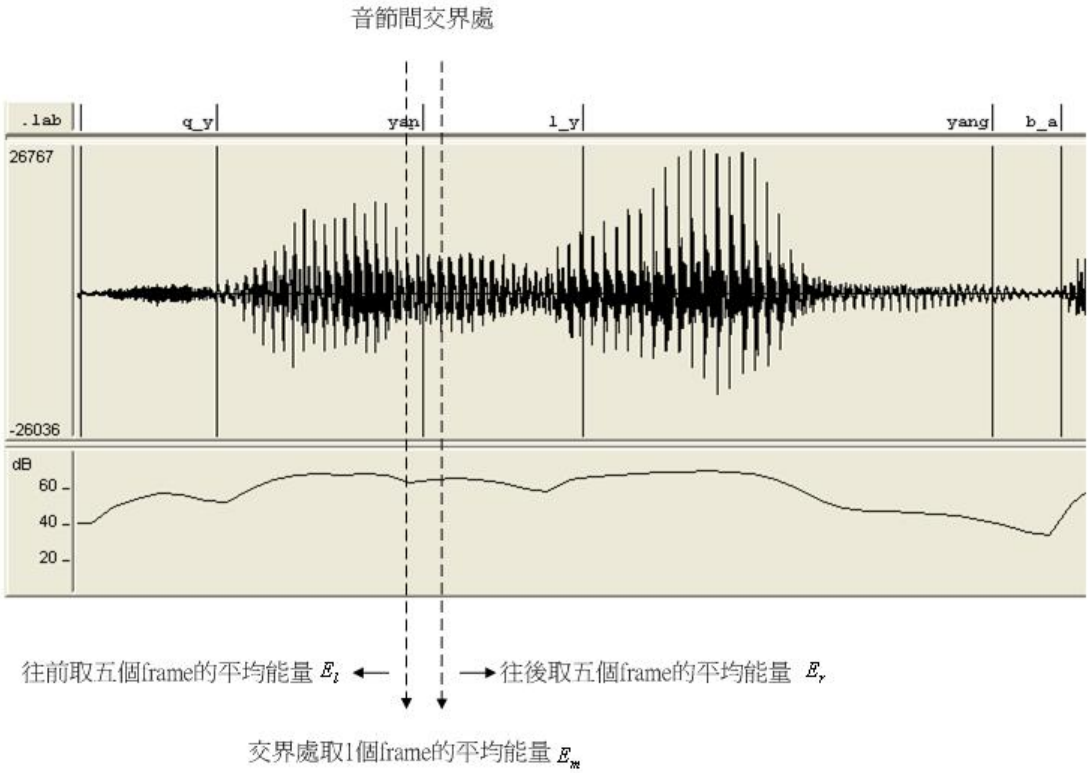


圖 3-11：具有連音效應的範例圖

$$EnergDip = (\frac{E_l + E_r}{2} - E_m) \tag{3-9}$$

最後，我們想要從所有的候選合成單元中挑選出最佳的合成單元組合成目標句，就要總合上述所定義的候選合成單元內部韻律參數差異(intra-unit cost)，和候選合成單元間的連接差異(inter-unit cost)，加以一個權重，並利用 Viterbi Search 來找出差異最小、分數最高的最佳路徑，以取得最佳的合成單元來作串接。

$$C(\hat{S}_i) = w_{intra} \sum_{k=1}^{n} C_{intra}(L_k^i) + (1 - w_{intra}) \sum_{k=1}^{n} C_{inter}(L_k^i) \tag{3-10}$$

3.4.3 連音效應的修正

根據語言學的特徵，我們將音節之間的連音效應分成鬆散連接(loose concatenation)、緊密連接(tight concatenation)以及重疊連接(overlapped concatenation)，在這三種連音型態中，我們很難將緊密連接或是重疊連接的語音乾淨的切出音節，因此，在以單一音節為語音資料庫之合成系統中，通常會選擇以鬆散連接的語音作為合成單元[7]；可是，本研究中以大型資料庫為基礎的中文語音合成系統，在長詞優先的原則下，選出的合成單元可能會有連音效應的情形出現。因為當我們依據選取最佳合成單元公式來選取合成單元時，若有候選合成單元具備相當長度，並且 Intra-unit cost 相較於 Inter-unit cost 高出很多，相較之下，雖然該候選合成單元在原句中和前一個音節有著連音效應，但亦有可能會被選出為最佳的合成單元。此時，在串接處就會有不平滑的雜音出現，為了要解決這個問題，我們對有連音效應的部分作 fade in，如下圖所示：

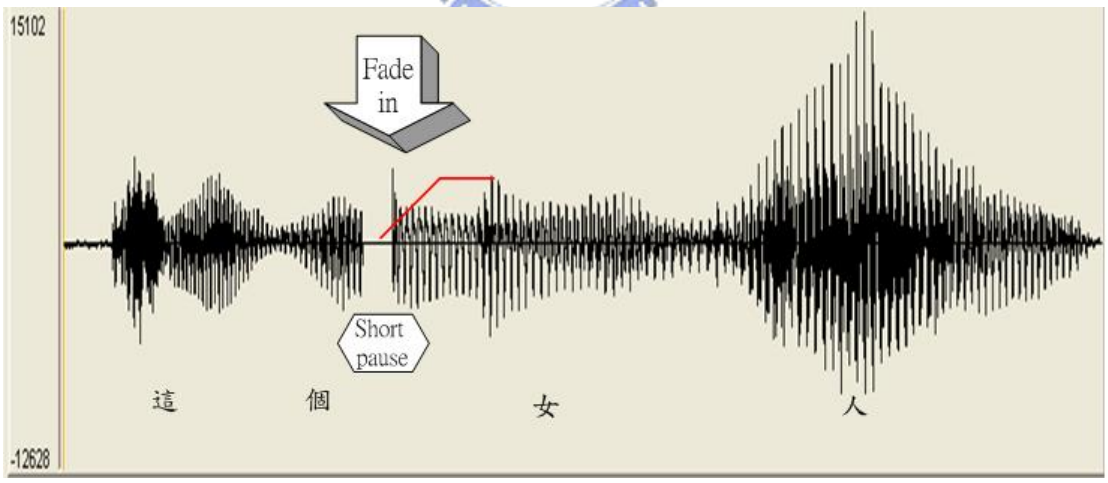


圖 3-12：對有連音效應的部分作 fade in 之範例

經過上述修正後，在連音效應的雜音部分，聽覺上有著明顯的改善，如下圖：

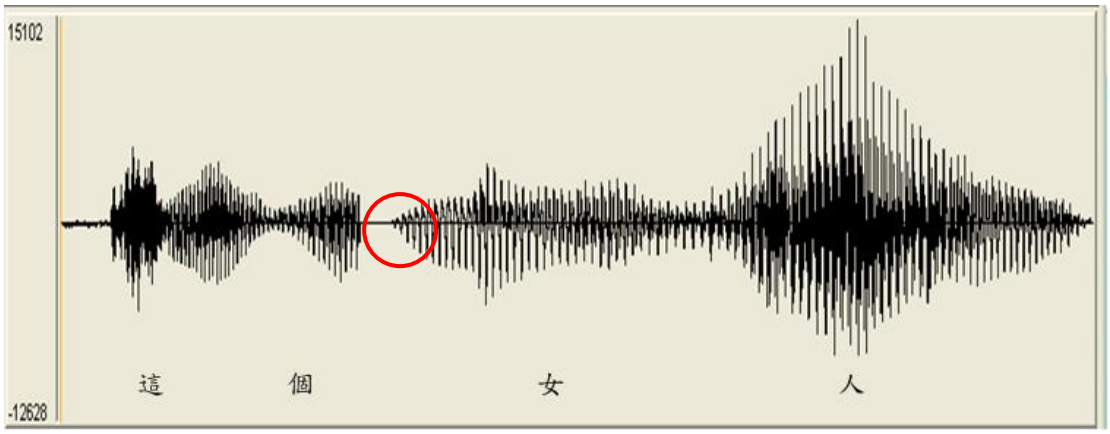


圖 3-13：對有連音效應的部分作 fade in 後之波型

第四章 實驗結果與分析

說明本論文 Corpus-based 中文語音合成系統相關設定，並對提出之挑選合成單元機制作效能分析，最後，利用主觀式測試與聽覺測驗，評估合成系統的好壞。

4.1 系統設定

在本套語音合成系統中，我們使用 Pentium PC，作業系統在 Windows XP 的環境下操作，輸入的文句為 big5 形式的中文字，輸出的合成語音取樣頻率為 16 kHz 的 pcm 檔案格式。我們的大型資料庫總共有 629 篇短文，內含有 2,523 個目標句，共有 8,017 個詞，51,682 個音節，語音檔案大約為 417 Mbytes，其標記基本音節的切割位置檔案大小大約有 4.75 Mbytes。在挑選合成單元時，所需要的資訊及其檔案大小包括有：

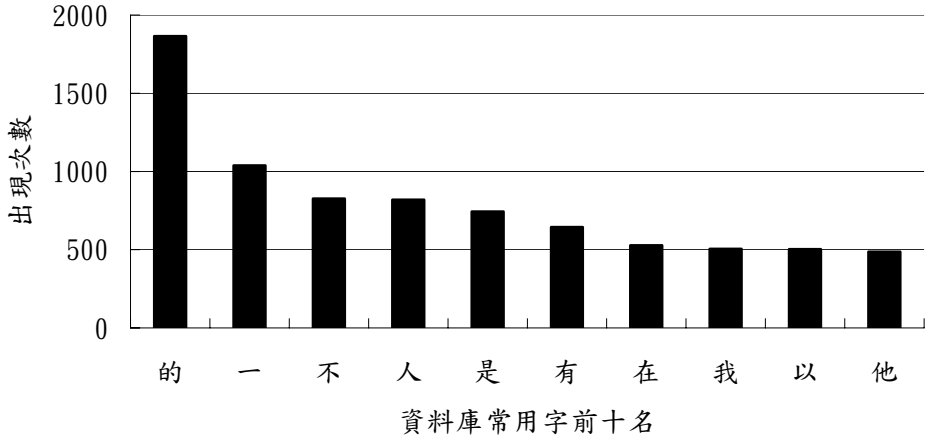
- Tonal syllable table：177 Kbytes。
- CLT：403 Kbytes。
- FCLT：64 Kbytes。
- 記載資料庫中文結構樹語意訊息及語音的韻律參數：2,770 Kbytes。

4.2 連續相關比對法實驗測試

為了要確認連續相關比對法，是否真能有效率的找尋我們所需要的候選合成單元，我們另外錄製語料庫外的文句，並利用本系統的搜尋模組作時間測試。首先，我們統計在資料庫中的十個常用字，出現次數如表 4-1 所示，其中屬於一字

詞的部分紀錄在 FCLT 中作額外比對；

表 4-1：前十名常用字出現次數統計表



除此之外，全部紀錄在 CLT 中，統計共有 2,659 個中文字，我們必須取出 CLT 中所紀錄的資訊，以作前後中文字的連續相關比對，出現次數之相異字數累積百分比如下圖：

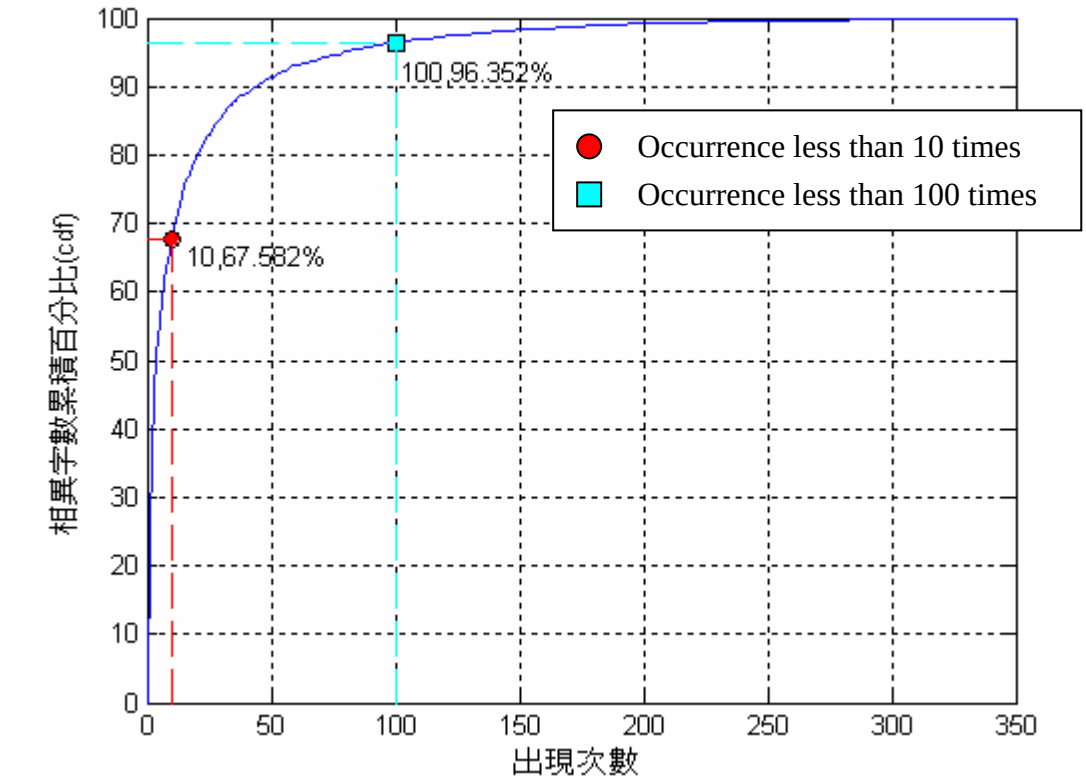


圖 4-1：CLT 中中文字出現次數分布圖

由圖中我們可以發現，依據連續相關比對法中所需的比對時間和中文字出現次數成正比的關係看來，若已經去除掉 FCLT 中記載的資訊，在 CLT 中約有 96% 中文字出現次數少於 100 次，在資料庫中更有超過 67%的中文字出現次數是少於 10 次，這除了說明這些中文字都是很少使用的，而且這也表示了我們在連續相關比對法中所需要的比對並不會耗費太多時間。

現在我們用錄製語料庫外的任意 150 個句子測試比對時間，其中共有 1,198 個目標句，11,663 個中文字，9,606 個詞；平均一個句子有 8 個目標句，每個目標句有 10 個中文字，8 個詞，實驗結果平均一個目標句的執行時間為 0.0126 秒，這個結果顯示了一個不錯的搜尋效能，因此，我們亦可實現一個效能不錯的即時語音合成系統。

4.3 正規化參數與權重值的設定

為了要決定選取合成單元公式中的最佳權重值，首先，我們需要經過正規化的步驟。在本論文中，我們使用上一節提到另外錄製的 150 個句子來作各項參數的範圍預測，為了要分別取得各項參數的約略範圍，在只保留該項參數的情況下，其餘參數先予以忽略。例如，在預測 d_d^j 第 j 個字的音長失真度時，我們給予它權重為 1，其餘權重皆設為 0，因此整個選取合成單元公式就變成只有考慮 d_d^j 的因素，整個公式亦變為只剩下

$$C(\hat{S}_i) = 1 \times \sum_{k=1}^{n} \left(-\frac{1}{m} \sum_{j=1}^{m} \frac{1 \times d_d^j}{Q_d} \right) \qquad (4\text{-}1)$$

由 1198 個目標句去作統計，橫軸表示音長失真度的大小，縱軸表示出現在對應失真度的目標句共有多少，由此可以知道 d_d^j 大約的分佈範圍如下：

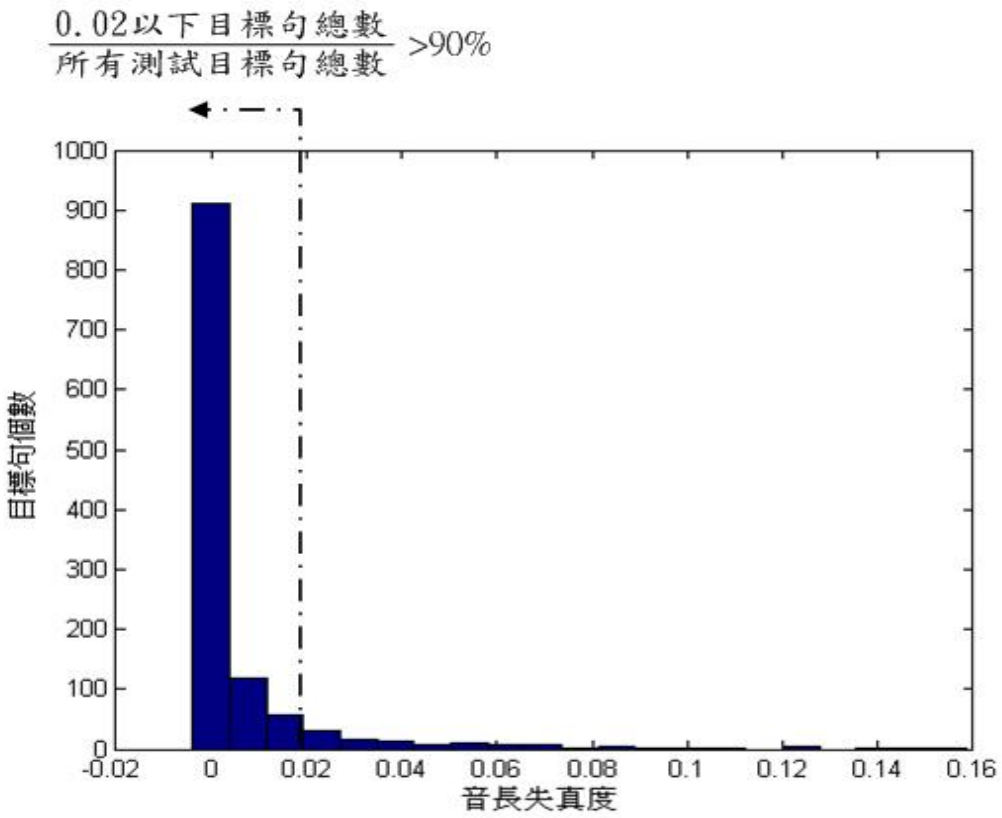


圖 4-2：音長失真度分布範圍圖

觀上圖，我們可以知道分布範圍為(0,0.16)，但是注意到其實大部分的目標句音長失真度相當集中，且在 0.02 以上的值出現機率幾乎很少，所以我們定其多數範圍為落在總數的前 90%以上處，即分布範圍為(0,0.02)，經過正規化的步驟後，大部分的音長失真度範圍可界於(0,1)之間，因此對於 d_d^j 而言，它的正規化參數值，Q_d 定為 0.02。其餘的正規化參數，則依此方式找出前 90%的範圍大小後，賦予各失真參數的正規化值。

在經過正規化使其各範圍皆落於(0,1)之後，我們將依據每個參數對於挑選合成單元的貢獻度來給予權重。因為本系統是以大型資料庫為基礎的中文語音合成系統，所以在選出來的合成單元越長越好的情況下，我們給予 $D_l(N_k^i)$，詞串字數這項參數的權重最重；其次，在長詞優先原則下，合成單元間相串接時韻律是否協調也是另一項重點，其中韻律參數差異由平均基週、音長、能量失真程度依

序降低其重要性。因此我們給予相對各項參數的權重分別為 $w_l = 20$ ，

$w_{f_0} = 4$ ， $w_d = 3$ ， $w_e = 2$ ， $w_s = 1$ ， $w_{intra} = 0.5$ 。

4.4 主觀式評估比較

本實驗之合成語音評量利用主觀式評估法(Subjective Test)，對於合成系統做進一步的評估。

本研究採用平均鑑定分數(Mean Opinion Scores,MOS)作為評估之標準[21]，這種評估方式將合成語音輸出的自然度分為優良(Excellent)，良好(Good)，尚可(Fair)，差(Poor)，極差(Unsatisfactory)五個等級，分別給予 5 至 1 不等的分數。測試人員在聽過合成語音後，以所感覺到的自然度評分。

測試是由合成系統根據使用單音節的語音資料庫，與以大型語料庫為基礎的合成系統，合成相同的中文句，做對照實驗。在此實驗中，合成二十個句子，由十五位測試人員(五位女性，十位男性)，聆聽並根據自己所感受的語音自然度打分數，最後取一個平均。

實驗中，比較兩套系統(A)、(B)，在合成語音自然度上的差異。

(A) 系統是利用單一音節為語音資料庫之合成系統。

(B) 系統為本研究之系統，是利用大型語料庫為基礎的合成系統。

表 4-2：合成語句 MOS 值比較表

使用合成器	平均MOS
(A) TD-PSOLA合成器	2.1
(B) Corpus-Based合成器	3.7

由上表結果可瞭解，利用本研究所提出的方法，進行合成單元的挑選，在自然度的表現上，相較於利用單音節的方式所合成的語音，有相當大的改進。

4.5 實驗結果分析

利用本研究中之合成系統，合出中文文句之語音，仍有些許的瑕疵，我們可以發現合成語音在聽覺上造成不舒適感的主要原因如下：

1. 韻律產生器的影響：

(1) 韻律不協調：

合成語音會有斷斷續續之感覺，主要原因可能是串接合成單元時，是由找到的長詞或未找到長詞之單字相串接而成，於是當有找到長詞時，其合成單元內的停頓(short pause)是依據原句韻律決定；但是，當未找到長詞而要由單字去作填補，或是每個長詞合成單元相串接時，語氣停頓就是依據由韻律產生器產生的韻律訊息來作決定，如下圖：

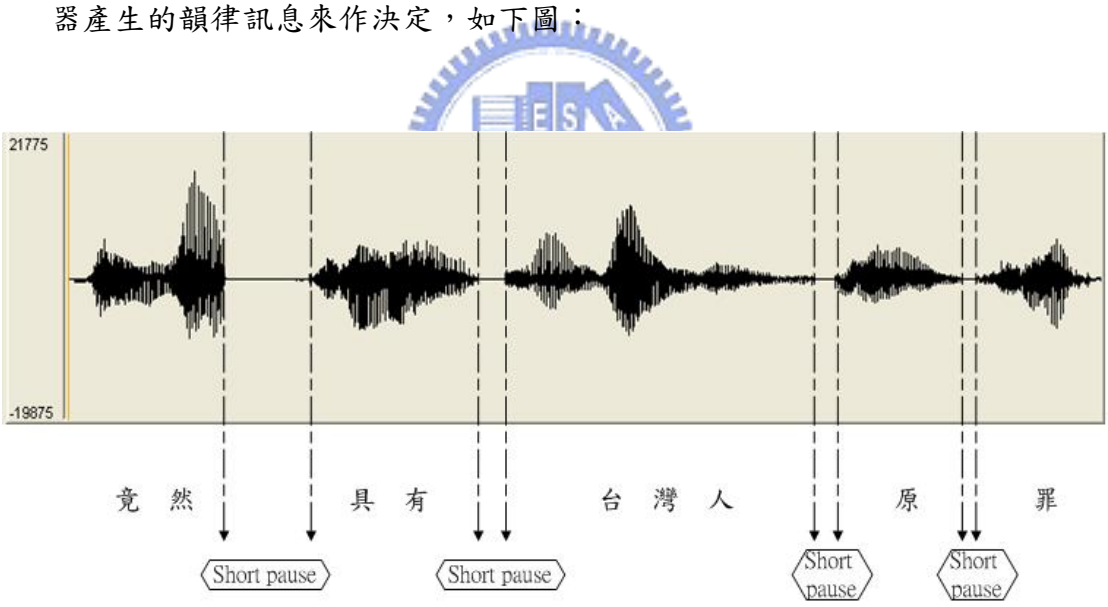


圖 4-3：合成語句韻律不協調範例圖

由上圖可以發現，當我們挑選出適當合成單元而從原句取出時，因為挑選的合成單元在原句的音韻上音節間並無停頓，或是有停頓但並非符合韻律產生器產生的韻律停頓，因此，當取出合成單元要作單元間相接時，單元間會給予韻律產生器產生的韻律停頓，在有時符合韻律產生器產生的韻律

停頓而有時卻不符合的情況下，會造成整句合成語音音韻節奏較不順暢的感覺。改進方法可能要重新訓練韻律產生器或是未來設計的韻律產生器可以產生出以詞為單位的韻律停頓，並和選出的合成單元作對應，才能改善韻律不協調的現象。

(2) 斷詞錯誤：

輸入韻律產生器的斷詞若不適當，所產生的韻律訊息及韻律停頓會連帶受到影響而發生錯誤，如此會使得當我們挑選出適當合成單元後，若有找不到的對應長詞，便要以帶有聲調的音節來填補目標句，且在合成單元間就會要加入韻律產生器所產生的錯誤韻律停頓，如此依據韻律訊息所挑選出的合成單元不僅韻律訊息有誤差，在語流上也會和真正的的語意有所出入。如下舉例為要合成「貨幣基金會年會」，斷詞應為「貨幣」+「基金會」+「年會」，但是韻律訊息產生器卻將它斷詞為「貨幣」+「基金」+「會」+「年會」，所造成的結果如圖：

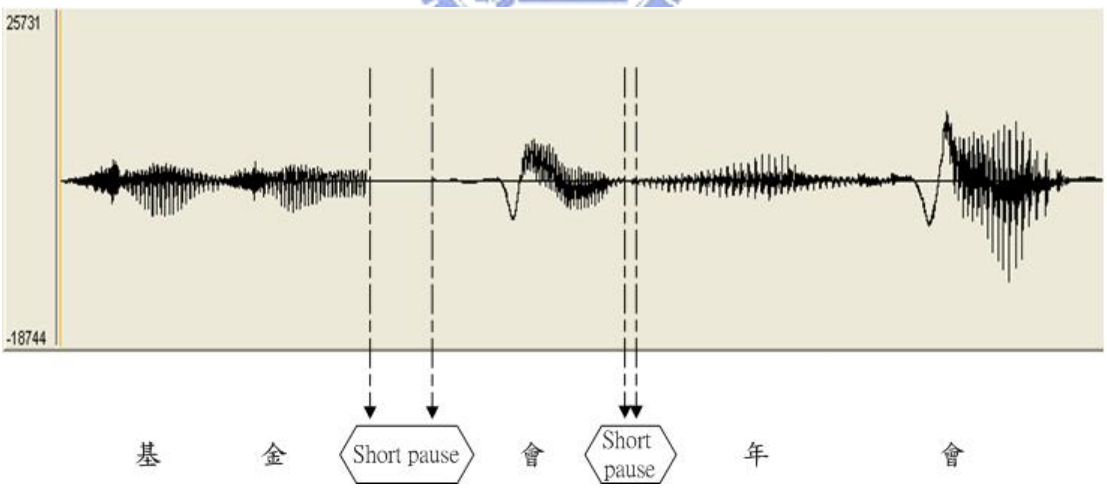


圖 4-4：合成語句韻律斷詞錯誤範例圖

由上圖我們可以明顯發現，原本應該是「基金會」和「年會」之間的韻律停頓較長，現在卻發生了「基金」和「會」之間的韻律停頓比「基金會」和「年會」之間的韻律停頓還要長的現象。

2. 切割位置不完整的影響：

切割位置的不完整尚包括有前面不完整或後面不完整兩類，實驗結果發現，當合成單元受到連音效應的影響使切割位置不容易準確或是切割位置錯誤，導致切割位置會在該音節尚未開始前或是在前一音節尚未結束時，這屬於切割位置前面不完整，相對的，對於前一音節而言，則屬於切割位置及早結束，即後面不完整類型。針對切割位置前面不完整方面，人耳聽覺上可以感受到其存在，如上一章最後一節說明，若是受連音效應的影響我們可以利用 Fade in 將聽覺上的不舒適感減到最低，但是若為切割位置錯誤，則是屬於調整切割位置不完美的問題；而在切割位置後面不完整方面，發現人耳對此並未如切割位置前面不完整時敏感，情況較顯著者，會有結束急促之感但並不會有雜音出現所造成的聽覺不舒適現象，但若情況非屬太過嚴重人耳甚至對其沒有異常的感覺。如下所示為原始音檔「個」和「問」之間切割位置錯誤，但是當取出切割位置後面不完整的詞「這個」當作合成單元時，人耳並未能明顯感覺出來：

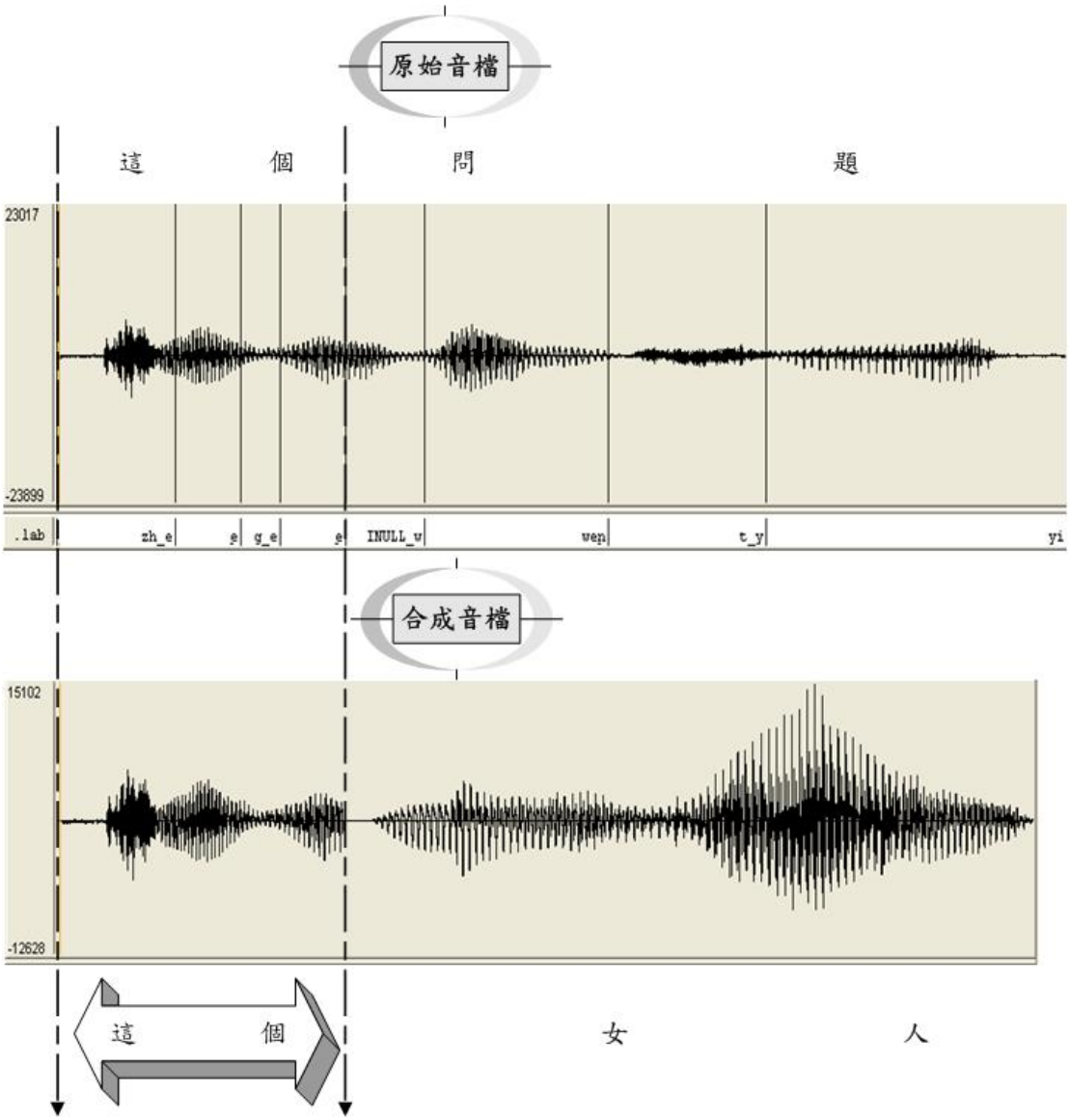


圖 4-5：合成語句切割位置不完整範例圖

3. 摩擦類子音能量過大的問題：

在合成語音的過程中，發現到有些以摩擦類起頭的音節，子音能量太強，使的合成到該段語音時會摩擦音能量會太過明顯，影響到整個合成語音給予人的舒適感，要解決這個問題，未來我們可用壓縮該類子音能量的方式作為改善，但是，最根本的解決辦法，是在錄製語料庫時，針對這個部分要特別要求音檔的品質，如此在有較佳品質的語料庫下，合成的聲音自然會更加流暢。

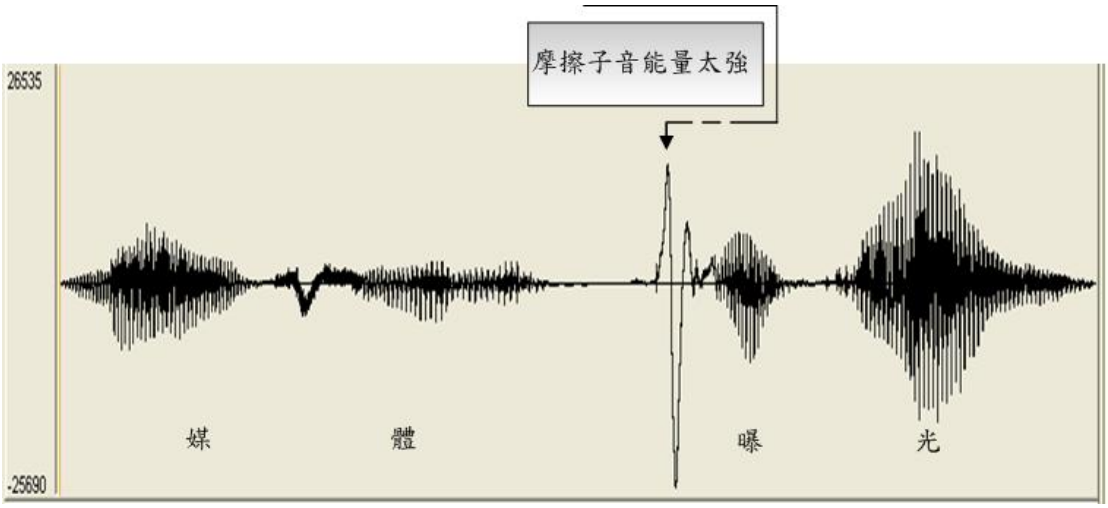


圖 4-6：合成語句摩擦類子音能量過大範例圖

4. 資料庫涵蓋涵蓋範圍的影響：

我們可以知道，在本研究使用挑選合成單元的公式中，較重要的依據為詞長及韻律參數，由實驗結果發現，當可以找到詞當作合成單元時，大部分所合出的語音是如預期所推論會較為自然流暢的；不過，仍然可以發現，有一些雖有找到對應長詞，但是因為在我們的資料庫中，這個詞的所涵蓋的韻律資訊並沒有很多，所以造成最後選取出來當作合成單元的韻律參數雖然是所有合成單元衡量之下最好的，可是其實和真正的目標韻律參數仍是有相當差距的，或者是導致最後並無法選出較長的詞而改由較接近目標韻律參數之短詞或帶有聲調的音節來作為最後要串接的合成單元；除此之外，我們亦發現當未找到詞當作合成單元而必須使用帶有聲調的音節來組成目標句時，若該帶有聲調的音節很接近所需要的目標韻律參數，雖然它的詞長只有單一音節，但是卻亦能合出不錯的語音。由上所述，我們可以知道，資料庫中韻律參數涵蓋範圍佔了決定語音良莠相當重要的一環，本研究雖然已可合出相當自然流利的語音，但若能繼續增加幾個常用詞及豐富其韻律參數，可以減少因選用韻律參數不適當所造成合成語音不順暢的現象。

第五章　結論與未來展望

本文提出以語料庫為基礎之中文文句翻語音系統中，合成單元選取之方法，使用電腦從一個大的語音資料庫中自動地挑選出一組合成單元，由實驗結果也證實，由大型語料庫中挑選合成單元合成的語音，相較於單一音節語料庫的合成語音，明顯自然流暢許多。

然而，未來除了要針對實驗結果分析中所提出的問題作進一步的加強外，對於以語料庫為基礎之中文文句翻語音系統中，如何進一步改善語音單元連接處不平滑的現象也是需要的；再者，因為本論文中提出的語音合成技術，不因語言不同而有所改變，所以未來可朝向建立一套整合國、台、客語的翻語音系統(multilingual system)邁進。另一方面，目前評估一個合成單元的好壞，需要經由人聽以辨識好壞，並無科學方法可用，這是全自動化建立合成單元的最大障礙，因此在未來的研究課題中，更精確的分析影響合成單元好壞的因素以及發展一個全自動化的科學評估方法是有必要的。

參考文獻

【1】 V. Kraft, “Does the Resulting Speech Quality Improvement Make a Sophisticated Concatenation of Time-Domain Synthesis Units Worthwhile?” Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, NY, pp65-68.

【2】 王小川教授，”語音信號處理”

【3】 陳鳳儀，蔡碧芳，陳克健，黃居仁,“中文句結構樹資料庫(Sinica Treebank)的構建”，中央研究院資訊所、中央研究院研究所。

【4】 Klatt, D. H. (1987) Review of text-to-speech conversion for English. J. Acoust. Soc. Amer, 82(3), pp.737-793.

【5】 Hamon, C., E. Moulines, and F. Charpentier (1989), “A diphone synthesis based on time-domain prosodic modifications of speech” in Proc. ICASSP, pp.238-241.

【6】 Chen, S.H., S.H. Hwang and Y. R. Wang(1998), “An RNN-based prosodic information Synthesizer for Mandarin text-to-speech,” IEEE Trans. On Speech and Audio Processing, Vol. 6, NO. 3, pp.226-239.

【7】 Chen, J. H. (1998) A Study on Synthesis Unit Selection and Prosodic Information Generation in a Chinese Text-to-Speech. Ph.D. Dissertation. National Cheng Kung University, Tainan, Taiwan, R.O.C.

【8】 Shih, C. L.and R. Sproat (1996), “Issues in text-to-speech conversion for Mandarin” in Computational Linguistics and Chinese Language Processing, vol. 1, Aug. 1996, pp.37-86.

【9】 Iwahashi, N. and Y. Sagisaka (1995), “Spech segment network approach for optimization of synthesis unit set,” Computer Speech and Language,

pp.335-352.

【10】Chiou, H. B., H. C. Wang, and Y. C. Chang (1991), “Synthesis of Mandarin speech based on hybrid concatenation,” Computer Processing of Chinese and Oriental Languages, Vol. 5, No. 3/4, pp. 217-231.

【11】Chou, F. C. and C. Y. Tseng (1998),”Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties” in Pro. ICASSP, pp.893-896.

【12】林立峰，”中文 TTS 系統與音合成之改進”，國立交通大學碩士論文，民國九十三年六月。

【13】The HTK Book (for HTK Version 3.2)

【14】魯弘茂，”中文語音合成技術之實作與分析”，國立交通大學碩士論文，民國九十一年六月。

【15】江振宇，”中文斷詞器之改進”，國立交通大學碩士論文，民國九十三年六月。

【16】W.J. Wang, W.N. Campbell, N. Iwahashi, and Y. Sagisaka, “Tree-based unit selection in speech synthesis,” in Proc. Of the Int’l Conf. on Aoustics, Speech, and Signal Processing, Vol. II, pp.191-194, 1993.

【17】A.J. Hunt and A.W. Black, “Unit selection in a concatenative speech synthesis system using a larger speech databse,” in Proc. ICASSP, Atlanta, 373-376, 1996.

【18】H. Peng, Y. Zhao, and M. Chu, “Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation,” in Proc. ICSLP, Denver, USA, 2002.

【19】T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, “Unit Selection Algorithm for Japanese Speech Synthesis Based on Both Phoneme Unit and Diphone Unit,” in Proc. of IEEE-ICASSP 2002, pp.465-468, May 2002.

【20】Chou, F. C., C. Y. Tseng, and L. S. Lee, “A Set of Corpus-Based Text-to-Speech Synthesis Technologies for Mandarin Chinese” in Pro. ICASSP, Vol. 10, pp.481-494, 2002.

【21】Min Chu and Hu Peng, “An Objective Measure for Estimating MOS of Synthesized Speech” in EuroSpeech 2001.