第三章　中文語音合成的斷詞系統

本文建立的中文語音合成斷詞系統，可略分為斷詞與構詞兩大單元。首先在斷詞單元中，我們嘗試分別利用長詞優先法(Longest Word First)，以及動態規劃演算法(Dynamic Programming)配合詞庫的使用，藉以比較兩演算法間的優異性，並找尋一適合用於中文語音合成系統的斷詞方法。此外，為免收集、整理詞庫所需耗費的大量人力與時間，於斷詞單元的基礎斷詞後，輔以定量複合詞構詞、疊詞及姓名三大構詞，以規則法的建立彌補詞庫的不完整。

本章將分別就斷詞系統中所用的詞庫，斷詞單元中的兩大斷詞方法 Longest Word First 與 Dynamic Programming，以及三大構詞詳細介紹。最後並對姓名資料庫、一般詞資料庫的擴增做一嘗試。

3.1 資料庫介紹

第二章曾提及，在中文語音合成中，斷詞結果的好壞將直接影響韻律節奏的產生，而詞庫在斷詞系統中又佔舉足輕重的角色，詞庫的正確性與涵蓋面，都足以影響斷詞系統的結果與效能。以下將對詞庫與斷詞系統的關連性，以及本文所用到的新竹清華大學多媒體資訊檢索實驗室語文資料庫、中央研究院漢語平衡語料庫一一介紹。

3.1.1 詞庫與斷詞系統的關連

如圖 3-1 所示，斷詞單元中詞庫是整體斷詞系統極具影響力的一角，當輸入的文句進入斷詞單元後，透過各種不同的斷詞演算法與其斷詞所依據的詞庫反覆交相運算，最後輸出基本斷詞結果。由於最簡易的斷詞系統即是將文句與詞庫比對，故此，詞庫的收集、整理若能愈趨正確而豐富，將大幅提升斷詞效能。

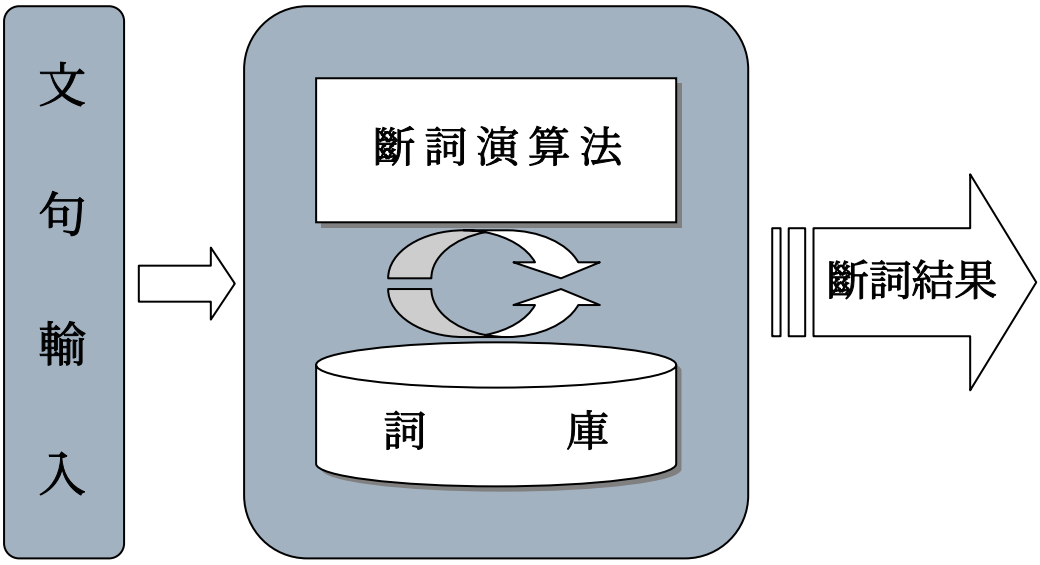

圖 3-1 斷詞單元中詞庫與斷詞演算法示意圖

3.1.2 清大多媒體資訊檢索實驗室語文資料庫(MIR DB)

本文斷詞所依據的詞庫即本實驗室歷年所收集、整理的語文資料庫（清大多媒體資訊檢索實驗室語文資料庫，簡稱 MIR DB），其中所使用的資訊除 95,276 個詞外，另有其對應詞的注音音標，為整個斷詞系統最後輸出所附訊息，憑供中文語音合成之用。另，為求單音節中文語音合成方法的順利，亦利用本詞庫中的

單字詞相對注音，作為後續標音依據。本詞庫詞數分佈情形如表 3-1 所示。

詞　　長	詞　　數
一字詞	7,682
二字詞	60,107
三字詞	13,921
四字詞	12,310
五字詞	682
六字詞	336
七字詞	154
八字詞	73
九字詞	3
十字詞	8
總　　計	**95,276**

表 3-1 清大多媒體資訊檢索實驗室語文資料庫(MIR DB)詞庫分佈

3.1.3 中研院漢語平衡語料庫(Sinica Corpus 3.0)

中研院詞庫小組於 1997 年 10 月完成一五百萬目詞的語料庫，稱之中央研究院漢語平衡語料庫(Sinica Corpus 3.0)。該語料的整理主題遍及哲學、科學、社會藝術、生活及文學六大主題，語料語式含書面語、演講稿、劇本台詞、會話以及會議記錄五類，語料媒體分佈有報紙、一般雜誌、期刊、教科書、工具書等共十一種。[1]該詞庫亦含有詞頻統計訊息，表 3-5 所示為自平衡語料庫中所抽取出

[1] 黃居仁、陳克建＜中央研究院平衡語料庫的內容與說明（修訂版）＞，中央研究院資訊科學研究所中文詞知識庫小組，1998 年 8 月。Technical Report No. 95-02/98-04。

130,757 個不重複的詞。

主　　題	哲學	科學	社會	藝術	生活	文學	總　計
字數總計	68.53	10.24	276.13	73.22	141.20	127.85	789.27
詞數總計	45.17	67.50	182.03	48.27	93.08	84.28	520.28
百分比	8.68%	12.97%	34.99%	9.28%	17.89%	16.20%	100%
字數／詞數單位：萬							

表 3-2 中研院漢語平衡語料庫主題分佈

語　　式	書 面 語	演 講 稿	劇本台詞	會　　話	會議記錄
字數總計	711.47	10.93	6.45	57.55	2.87
詞數總計	469.00	7.20	4.25	37.94	1.89
百分比	90.14%	1.38%	0.82%	7.29%	0.36%
字數／詞數單位：萬					

表 3-3 中研院漢語平衡語料庫語式分佈

媒　體	報紙	一般雜誌	期刊	教科書	工具書	學術論文	一般圖書	視聽媒體	會話訪談	演說	其他
字數	**246.89**	**230.28**	5.49	32.23	1.06	10.71	66.70	180.20	12.90	2.00	0.81
詞數	**162.57**	**151.80**	3.62	21.25	0.70	7.06	43.96	118.80	8.50	1.32	0.53
百分比	**31.28%**	**29.18%**	**0.70%**	**4.08%**	**0.13%**	**1.36%**	**8.45%**	**22.83%**	**1.63%**	**0.25%**	**0.10%**
字數／詞數單位：萬											

表 3-4 中研院漢語平衡語料庫媒體分佈

詞長	詞數	詞長	詞數
二字詞	66,765	十一字詞	24
三字詞	45,894	十二字詞	18
四字詞	13,243	十三字詞	10
五字詞	2,730	十四字詞	8
六字詞	913	十五字詞	3
七字詞	560	十八字詞	3
八字詞	390	十九字詞	1
九字詞	131	二十二字詞	1
十字詞	62	二十五字詞	1
			總計　130,757 個詞

表 3-5 中研院漢語平衡語料庫(Sinica Corpus 3.0)詞庫分佈

3.2 斷詞單元

斷詞單元為整個斷詞系統的關鍵軸，斷詞單元所輸出的初步斷詞結果，將影響後續構詞的效果。以下分別介紹本文所用的兩個斷詞演算法：長詞優先法(Longest Word First)與動態規劃演算法(Dynamic Programming)。

3.2.1 長詞優先法

3.1.1 節介紹詞庫與斷詞系統的關連時曾提及，最簡單的斷詞方法即是將文句與詞庫作一比對，而此類作法稱之「結構性斷詞法」，或「機械斷詞法」。[2]亦即斷詞系統依一定法則將文句與詞庫中的詞進行匹配，比對成功即視為斷詞輸出結果，如此反覆直至輸入的文句完全匹配結束。而其中，長詞優先法(Longest Word First)即屬此類斷詞演算法。

本文所採取的長詞優先斷詞法為正向長詞優先法，亦即，將輸入文句由頭至尾分別與詞庫進行比對，如圖 3-2 所示。至於長詞優先法中「最長詞」字數的認定，為適合不同中文語音合成方法使用，討論如下：

（一）單音節合成的中文語音合成系統

單音節合成方法為中文語音合成的常見方式，利用中文中常見的 411 個單音節作為語音合成的基礎。作為單音節中文語音合成用的斷詞系統用時，我們可依人的發音習慣將「最長詞」的字數設為 5~6 字，如此將大幅降低比對時間，增加斷詞系統的效能。

（二）大量語料庫合成的中文語音合成系統

在中文語音合成中，另一常見的合成方法即為利用大量語料的語音合成。不同於單音節語音合成，大量語料合成由於藉由自然人的發音，因此在自然度上將較單音節語音合成來得自然。[3]然而，由於大量語

[2] 朱怡霖，＜中文斷詞與專有名詞辨識之研究＞，國立臺灣大學資訊工程學研究所 90 學年碩士論文。Page. 16。

[3] 謝明峰，＜使用大量語料庫的中文語音合成系統實作＞，國立清華大學資訊工程學系 92 學年碩士論文。

料庫的語音合成所收集、整理的語料庫來源往往多元化，非同一人、同時間、情境下所錄製，故此時的斷詞系統「最長詞」應依情形設定在語料庫中的最長詞字數，盡可能避免因連接不同語料所造成的不連續性。

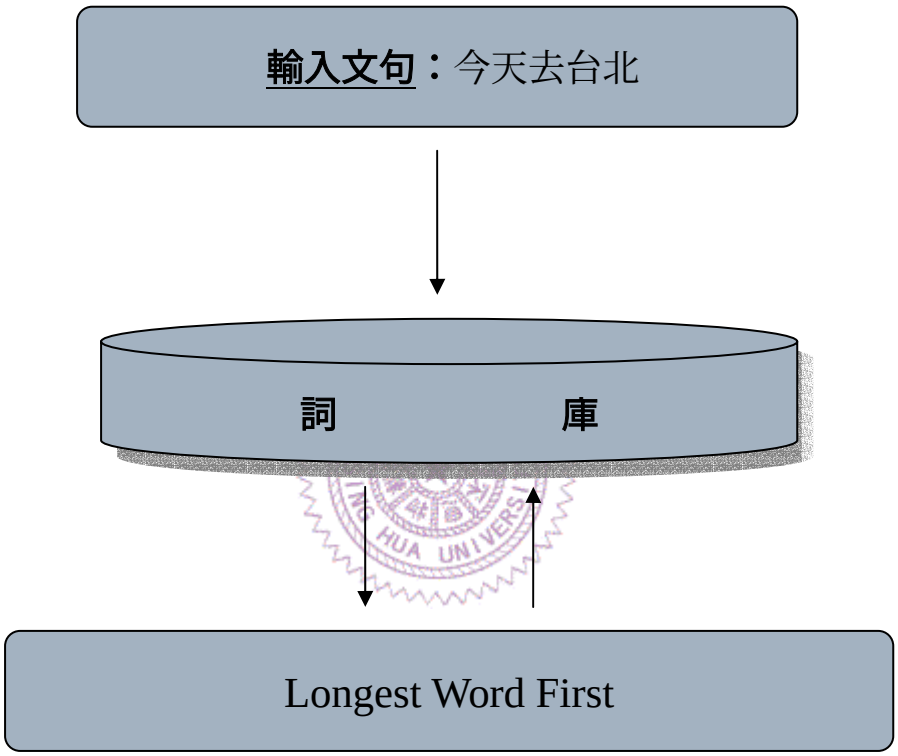

圖 3-2 長詞優先法示意圖

3.2.2 動態規劃演算法

動態規劃演算法(Dynamic Programming)在訊號處理中常被作為一種相似度比對的方式，如圖 3-3 所示，首先，我們將輸入文句分別攤於 *t*、*r* 軸，其次於 *t*、*r* 軸所構成的平面中列出可能的詞組，期間我們可發現當 *t*、*r* 相等時，為斷詞所最不期望出現的單字詞。其次我們就可能出現的詞組於斷詞所憑藉的詞庫相互比對、給分，分數定義如下：

（一）自訂分數

在 3.1 章節介紹斷詞資料庫時曾提及清大多媒體資訊檢索實驗室語文資料庫(MIR DB)並無含有詞頻資訊，因此我們採取自訂分數的方式實驗，定義如下[4]：

length(words) == 1 *score = 1;*

2 <= length(words) <= 6 *score = 3;*

length(words) > 6 *score = 2;*

（二）詞頻給分

在 Dynamic Programming 另一分數定義中，我們利用了中研院漢語平衡語料庫(Sinica Corpus 3.0)中的詞頻訊息，定義如下：

score = length(words) x *frequency(word);*

[4] 如 3.2.1 節所述，我們可依系統需求自訂最長詞的字數，本文以六字為例。

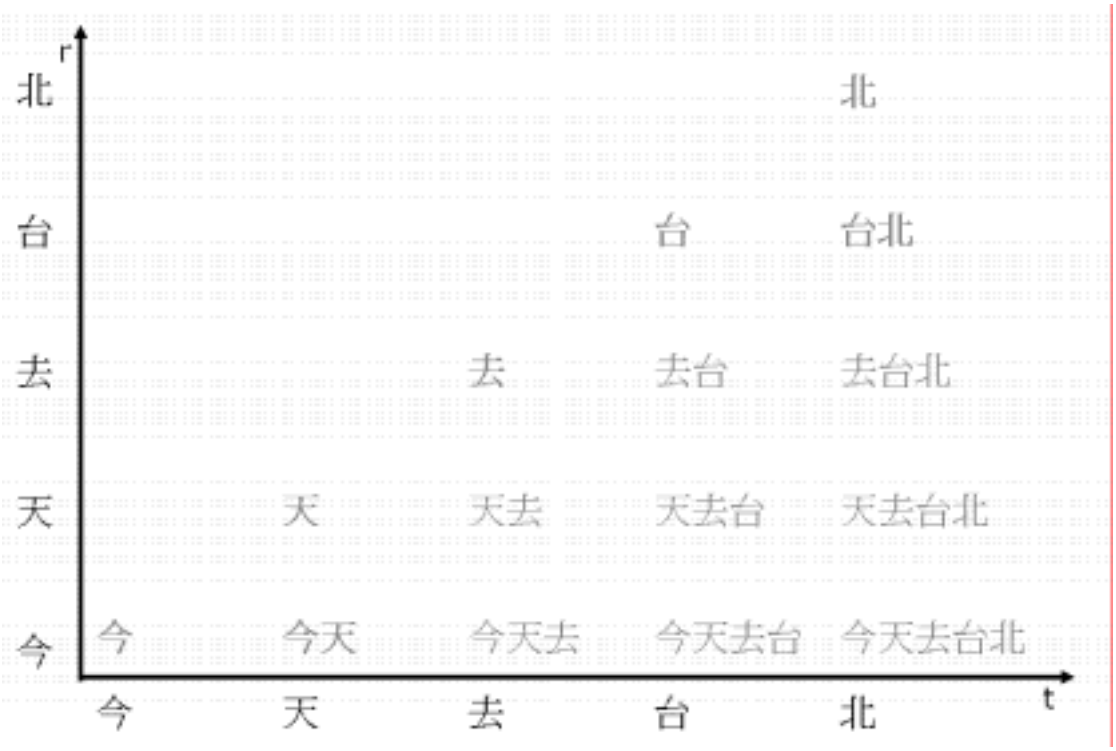

圖 3-3 動態規劃演算法示意圖一

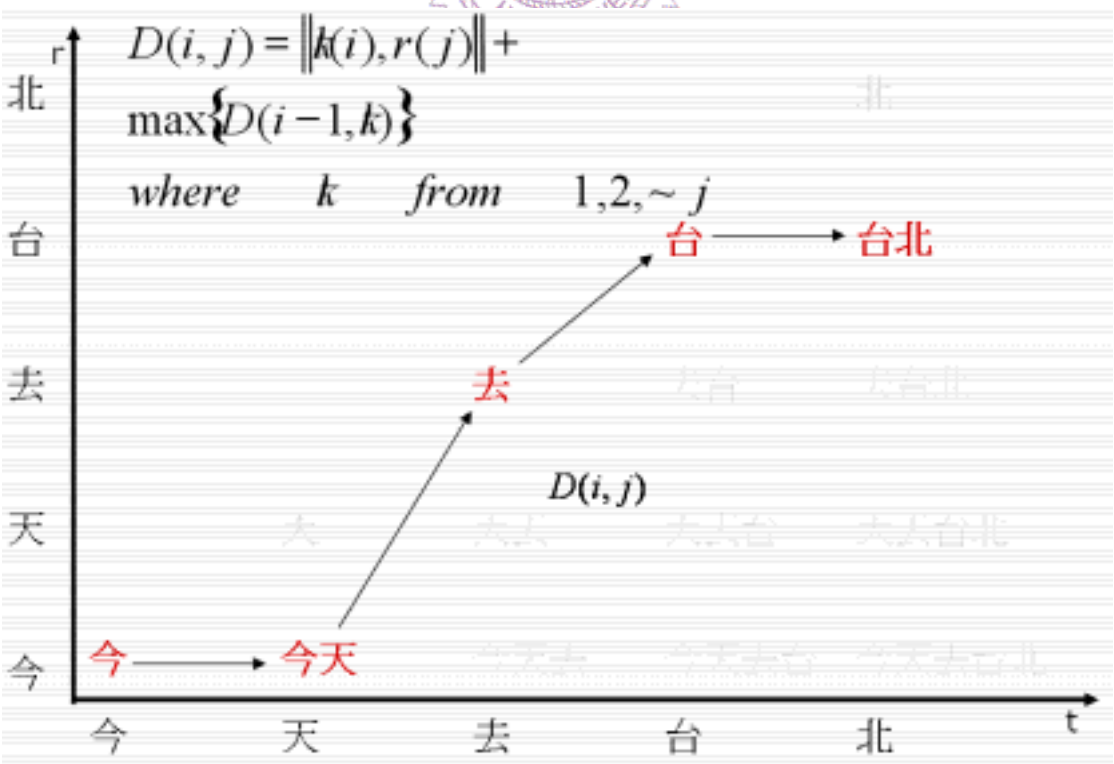

圖 3-4 動態規劃演算法示意圖二

在進行初步給分後，我們依照如下原則將所有可能詞組的分數作一累加。最後將 *t(i)*上最高分者連接起來，如圖 3-4 所示，當路徑走勢非水平時，即為單字詞，若走勢為水平則是一連續詞。

$$D(i, j) = ||k(i), r(j)|| + Max\{D(i-1, k)\}$$

$$Where \quad k \quad from \quad 1, 2, \ldots, j$$

3.3 構詞單元

斷詞單元憑藉斷詞演算法與詞庫的相互比對進行斷詞，然而，要將所有可能的詞組收錄詞庫中，非但耗費大量人力與時間，龐大的詞庫亦將影響斷詞效率。以下將分別介紹定量複合詞、疊詞及姓名構詞規則，以補斷詞單元中詞庫不足之憾。

3.3.1 定量複合詞構詞

在中文文法構詞中，定詞與量詞常被用來修飾名詞，例如「一本書」、「三張椅子」、「兩檔子事」，定詞「一」、「三」、「兩」與量詞「本」、「張」及「檔子」，分別修飾名詞「書」、「椅子」和「事」。此之吾人稱為定量複合詞。

定量複合詞的組合無限，詞庫無法一一收集，但由於定量複合詞的構詞具有規則性，本文參照中央研究院中文詞知識庫小組陳克健、黃居仁先生等人所訂 34 條定量複合詞規則加以刪增、修改。定詞 67 個，量詞部分共 678 個，如附

錄一所詳列。

定量複合詞的構詞規則能較有效解決收集、整理無限度的定量複合詞組合，使得詞庫中不再需要大量的定量複合詞詞組，表 3-6 和表 3-7 則是我們根據定量複合詞的構詞規則，將 3.1.2 與 3.1.3 節所提兩斷詞依據詞庫做一刪減修正，是為後續斷詞所實際使用的斷詞詞庫。

詞　　長	詞　　數
一字詞	0
二字詞	59,429
三字詞	13,800
四字詞	12,298
五字詞	678
六字詞	336
七字詞	154
八字詞	73
九字詞	3
十字詞	8
總　　計	**86,779**

表 3-6 修正後的清大多媒體資訊檢索實驗室語音資料庫(MIR DB)詞庫分佈

詞長	詞數	詞長	詞數
二字詞	65,961	十一字詞	13
三字詞	44,412	十二字詞	5
四字詞	11,771	十三字詞	3
五字詞	1,408	十四字詞	1
六字詞	509	十五字詞	2
七字詞	236	十八字詞	3
八字詞	209	十九字詞	1
九字詞	68	二十二字詞	1
十字詞	36	二十五字詞	1
			總計　124,640 個詞

表 3-7 修正後的中研院漢語平衡語料庫(Sinica Corpus 3.0)詞庫分佈

3.3.2 疊詞構詞

中文中常見許多疊詞，如高高興興、快快樂樂、輕飄飄、慢慢地，諸如此類的中文詞我們能有效地歸納其構詞規則，藉以協助斷詞系統對此類詞組的斷詞。本文所採疊詞構詞如下：

（一）AABB。如：

高高興興、快快樂樂

（二）ABAB。如：

活動活動、快活快活

（三）ABB。如：

輕飄飄、慢吞吞、輕悄悄

（四）AA、AAAA。如：

慢慢、————

（五）AA 地、AAAA 地。如：

輕輕地、慢慢慢慢地

（六）{C}又{C}。如：

一天又一天、一次又一次

3.3.3 姓名構詞

一直以來專有名詞辨識在中文斷詞系統中是個棘手的問題，其中姓名辨識是為典型。以往姓名辨識常利用龐大姓名資料庫的統計為依據，建立姓名辨識模組，然而姓名資料庫的建立與取得不易，龐大的資料庫對斷詞效能亦有影響。衡量斷詞系統效能以及語音合成用途的容錯性，本文採簡易的姓名構詞方式，亦即利用斷詞系統最後所剩無多的連續單字詞，搭配中國百家姓作為姓名構詞的依據。在百家姓中，吾人將原有 444 個單字姓及 60 個複姓，篩選常用 214 個單姓、30 個複姓，如附錄二。

3.4 詞庫擴增

3.1 章節中曾提及詞庫在斷詞系統中又有舉足輕重的角色，因最簡易的中文斷詞系統即將輸入文句與詞庫一一比對，故詞庫的正確性與涵蓋面，都將影響斷詞系統的結果與效能。而除原有詞庫的收集整理外，以下我們將嘗試利用簡易的 Mutual Information 方式建立詞庫，其中以專有名詞的姓名資料庫為主[5]。

3.4.1 新聞資料庫簡介

在詞庫擴增所依據的資料庫方面，我們收集了民國九十二年七月二十日起，至九十三年六月五日止，共 169,747 篇電子新聞，其中包含中時電子報、台灣新生報、自由電子報、中央社及新浪網新聞五大報，含括焦點新聞、政治新聞、社會綜合、國際新聞、大陸新聞、地方新聞、財經產業、股市理財、醫藥生活、影視娛樂、運動天地、藝文新聞、論壇、旅遊新聞、生活新聞共十五類主題。

3.4.2 姓名及一般詞資料庫的增加

Mutual Information 在自然語言處理中常用來處理未知詞的辨識，而本文所採之簡易 Mutual Information，即藉由統計與門檻值的設定，求得 N-gram 的詞，最後輔以人工方式修正。在姓名資料庫方面，我們又輔以 3.3.3 節所列常用 244 個百家姓，共得 1,659 筆的三、四字的姓名資料庫，以及 123,906 筆一般詞資料庫(N-gram, N = 2 ~ 10)。

[5] 本文實驗性之簡易 Mutual Information 對一般詞資料庫的增加易產生許多不適宜的詞，故增加的一般詞資料庫暫不納入斷詞系統中。