

中文連音二字詞之語音合成

Coarticulation of Two-Syllable Words in Mandarin Speech Synthesis

Jun-Wen Hwang , Ming-Shing Yu , Shyh-Yang Hwang and Ming-Jer Wu
(黃志文) (余明興) (黃世陽) (吳明哲)

Department of Applied Mathematics

National Chung-Hsing University

Taichung , 40227, Taiwan

E-mail : MSYU@DRAGON.NCHU.EDU.TW

摘 要

本篇論文最主要在研究語音合成中的連音部份，我們從所錄製好的連音庫中，切取出連音二字詞，然後將之細分成三部份，針對每部份使用單音重新組合而成。在重新組合的過程，利用已知的連續音資訊，從單音中抓取最像連續音的部份來合成連音二字詞，並且模擬它的連音情形，例如考慮音量、基週走勢、音長等，使得日後在合成連續語音時，能達到類似連續語音的自然流暢。

1. 緒論

1.1 連音型態與特性

John R. Deller 等人在所著一書中[1]提到有關連音 (Coarticulation) 的解釋為：『在語音的產生過程中，發音器官在產生一連串所須要的音素時，爲了達到語音的自然流暢性，發音器官的變化是平滑的，而連音正是從這平滑的變化過程中所產生的』。

在中文連音的研究方面，近來有陳志祥[7]在中文連音型態之初步研究一文中指出，連音型態分爲三個基本型態：

(1) 停頓連接

在兩音節間有一段靜音存在，通常發生在詞和詞之間，如（圖 1-a）所示，本圖中爲『辛勞的播種』中的『的播』。

(2) 緊密連接

在兩音節間幾乎沒有停頓，但兩音節間並無重疊的波形存在。通常在詞內這種情形較常發生（圖 1-b），本圖中爲『風吹草動』中的『風吹』。

(3) 重疊連接

在兩音節間，不僅不存在靜音，且其基週呈現出連續變化的情形，兩音節中間會有一段轉換過程中過渡的週期波存在（圖 1-c），本圖中爲『政府官員』中的『官員』。

更詳細的說，停頓連接會發生於呼吸群[2]（李琳山教授於演講時稱之爲韻律段，Prosodic Segment）結束和下一個呼吸群開始之間。而緊密連接和重疊連接則會發生在一個呼吸群之中。至於是緊密連接還是重疊連接，則視此二字詞的結構而定，若後音節的子音是具週期性的子音，如 ㄇ、ㄋ、

2 研究進行方式

2.1 單音庫與連續音庫

爲了研究連續語音中的連音現象，我們請了一位女性錄音員，錄製好單音庫和連續語音庫，其取樣頻率（**Sampling Rate**）均爲 12 kHz。在單音庫方面，因爲我們是利用單音來合成連續音，所以在單音的錄製方面，我們分成前音和後音，其中前音是所錄製的二字詞中的首字，而後音則是其末字。

在連續語音庫方面，我們依據某些報紙文章錄製而成，整個語音檔資料總共錄製有 55,786,883 Bytes，共 13,999 個中文單字，共約一小時又十七餘分鐘，平均每個音長爲 330ms，相當於每秒發 3.03 個音。連續語音庫提供我們在使用單音來合成連續語音時，如何模擬產生連續語音中重疊連接的連音段，進而提高語音合成的自然流暢性。

2.2 動態時間校準演算法之應用

由於我們所錄製的單音，用來合成連續音時，大部份都較真正的連續音爲長，所以我們必須從單音中抽取出真正用來合成連續音的部份，因此我們使用 DTW 來做單音和連續音的比對。首先，以中文單音爲參考樣本（**Reference**），再從連續音庫中，取出相同的音十個做爲測試樣本（**Test**）。又因爲各樣本的長度並不一致，所以我們以音框（**Frame**）爲單位，音框的長度固定爲 20 ms，測試樣本固定爲 50 個音框，參考樣本固定爲 80 個音框，音框和音框之間的重疊部份（**Overlap**），則視樣本的長度而定。也就是說，彈性調整音框重疊部份，使得測試樣本音框數固定成 50 個音框，參考樣本音框數固定成 80 個音框。

在此處的測試樣本音框數 50 及參考樣本音框數 80 的訂定，在測試樣本部份是依據在連續音庫中連續音的長度而定，根據對連續音庫中的個別音節做統計的結果其音長範圍約落在 100 ms 到 500 ms 之間，因音框的長度為 20 ms，所以我們選取音框數為 50，使得對不同長度的測試音（在此處為連續單音），只要調整音框重疊的長度，就可使得不同長度的單音具有相同的音框數。而固定的音框數是爲了使我們在做 DTW 比對時，不同長度的單音，卻仍擁有相同長度的比對路徑，方便觀察比較。

同理，因爲在單音庫中的單音長度大約落在 280 ms 到 800 ms 之間，大約是連續音的 8/5 倍，所以我們取音框數為 80。對於不同長度的單音利用重疊的長度來調整使其具有相同的參考音框數。

在 DTW 的演算法部份，全域路徑限制（Global Path Constrains）方面，定爲 1:4 大約爲連續音的音長比上單音的音長的最大值。 ϵ 是比對範圍前端和尾端容許的鬆弛值（Relax），在本實驗中我們將 ϵ 的值設成 6，使得前端和尾端的對應彈性較大。在音框的距離量測方面，對每個音框我們使用了 16 階的倒頻譜係數。

2.2.1 DTW 應用於母音部份之合成

首先我們從連續音庫中，切取出各類母音（韻母、複韻母、聲隨韻母）做爲測試樣本，從單音庫中切取相對之母音爲參考樣本，然後求取其 DTW 對應路徑。整個結果如圖 2 所示。

在圖 2 中，我們從所有母音類的 DTW 路徑對應圖中，發現絕大部份的路徑非常接近斜率 8/5 的直線。斜率 8/5 的直線表示在從單音中抽取我們所須要的連續音部份時，只要根據其音長的比例，等比例的從單音中抽取相對應的部份即可藉此合成連續音中之母音部份。

我們在真正利用單音來合成連續音時，是以基週（Pitch）爲合成的單

位，根據上面所發現的結果，我們以音長為比例，等比例的從單音中抓取基週以合成連續音之母音部份。

2.2.2 DTW 應用於子音部份

在上節中我們提到利用音長的比率抓取對應的基週 (Pitch) 合成連續音，但是子音部份並不存在穩定性的基週，所以我們在合成連續音中的子音時必須再特別處理。我們從連續音庫中切取各種子音，各約五十個，在切取這些子音時，必須在尾端保留開始和母音銜接約 2 至 3 個基週。取音框長度為 7 ms，測試樣本數為 50 音框，參考樣本數為 80 音框，進行 DTW 路徑對應。結果發現無氣塞音 (ㄅ、ㄆ、ㄇ)，送氣塞音 (ㄆ、ㄆ、ㄆ) 及無氣塞擦音 (ㄆ、ㄆ、ㄆ) 的路徑對應都落在同一類，非常合乎各類子音的特性，而且這些子音都屬於音長較短的子音，在連續音中的長度和在單音中的長度相差不大，所以我們決定在使用單音合成連續音時，這些子音並不特別處理，直接擷取單音中的子音作為連續音中的子音部份。至於有聲子音 (ㄆ、ㄆ、ㄆ、ㄆ)，因其具有像母音週期性的特徵，所以將這部份當成是母音處理。

最後剩下的為送氣塞擦音 (ㄆ、ㄆ、ㄆ)，及清音 (ㄆ、ㄆ、ㄆ、ㄆ、ㄆ)，因為這些子音的音長較長，會影響連續音合成的結果。我們亦從連續音庫中，對每個子音利用中研院所提供的字轉音介面，各找到約五十個屬於詞內 (含詞尾) 和詞外 (含詞首) 的子音，求其平均長度。整個結果如圖 4 所示，其中單音 (前) 和單音 (後) 分別代表詞首的單音和詞尾的單音。我們從圖中可看出：

1. 除了 ㄆ 和 ㄆ 之外，其餘的六個子音，在連續音中詞首和詞內的長度相差不大，所以在合成子音時，並不考慮是否在詞內的因素。

二字詞，所以這個合成架構的必要條件是：『用來做為模擬的連續語音二字詞必須存在』。一旦從連續音庫取得連音二字詞後，整個利用單音來模擬連音二字詞的步驟如下：

步驟一：找出整個連音二字詞的連音中點

步驟二：利用連音中點切出連音段

步驟三：利用單音合成前段音節

步驟四：合成連音段

步驟五：利用單音合成後段音節

整個流程如圖 6 所示。

3 連音二字詞之合成

3.1 連音中點

在連音中點的求取方面，我們利用兩相連單音和連音二字詞以倒頻譜係數差（Delta - Cepstrum）[3][5]為參數，然後利用 DTW 路徑比對找出連音中點。在此，我們發現使用倒頻譜係數差所找到之連音中點準確性非常高。於是我們使用此方法協助我們決定連音中點。

3.2 決定連音段

連音段（Coarticulation Segment）是指重疊連接時，在兩音節中間所存在的一段轉換過程週期波，基本上它可能是一個存在兩音節中的音素。首先我們必須決定出連音段長度。因為我們在 3.1 節中可以準確的決定連音中點，所以我們取連音中點附近 20ms 為一音框，而且整個連音二字詞

亦以 20ms 爲一音框，每次重疊爲 10ms，然後以連音中點的音框逐次和整個連音二字詞的音框進行比對，經三點平均平滑後，產生如圖 7 的音框距離對應曲線，從圖中可看出連音中點會是一個波谷產生點，於是我們對此曲線微分求連音中點兩端斜率最大的音框，則此兩端所切即爲連音段。

3.3 利用單音合成前後段音節

連音段求出之後，則整個二字連音詞被切成三部份，第一部份即前音節部份，第二部份即連音段，第三部份即後音節部份。因爲我們是利用單音來模擬二字詞連音，單音和前音節的部份音長一定不相同，我們分別使用兩種不同的方法來進行前音節部份的合成

我們將子音和母音分開處理，子音依表二所述，只有那些在單音中較長的子音必須特別處理。而母音部份，在連音二字詞部份，我們取到第二部份，即第一部份加上連音段作爲要利用單音來合成的目標樣本。我們用了兩種不同的方法，都是用來調整單音的長度使其和目標樣本的長度一致。

第一種是我們在 2.2.1 節所提，在母音方面以週期爲單位，以單音基週數和目標樣本的基週數爲比例，取所對應的單音基週來合成目標樣本。而那些子音太長的單音，在子音部份我們亦使用其長度的比例進行長度的調整。

而第二種方法則較爲複雜，將目標樣本和單音以基週爲音框的單位進行 DTW 路徑對應，然後依路徑取出所對應的單音基週合成目標樣本，見圖 8。而子音太長的亦用 DTW 來進行長度上的調整。

3.4 合成連音段

我們在 3.3 節中，做單音節的合成時，前音節的部分有連音段所對應之單音部份，而後音節的部份亦有連音段所對應之單音，照合成的結構看來，多出了一段連音段。而所多出的連音段是爲了要利用圖 9 的方法合成連音段時用來重疊用。也就是將前單音節的連音段和後單音節的連音段重疊合成出連音段。

在合成連音段時，必須使得合成中的週期長度一樣，於是我們使用基週重建的方法來使得週期的長度一致。

3.5 音量及基週走勢

在連音二字詞的合成方面，我們的重點是放在整個合成二字詞是否聽起來順暢自然，而不會有第一個音節尚未結束，第二音節音就已搶先發出之感。

所以連音段即扮演從前一音節過渡到下一音節的角色。雖然音量及基週走勢亦非常重要，但這應該是跟整句話的韻律較相關，我們會在將合成的二字詞放回整句話時，調整其基週走勢及音量。圖 10 是一個連續音和單音的基週走勢比較圖，由圖中可看出，基週的走勢是非常重要的。

4 評估與度量

4.1 實驗評估

最後我們進行整個結果的評估，測試分成兩部份。第一部份是對二字詞的評估。第二部份則是對平衡句的評估。

在二字詞的評估方面，我們利用單音合成從連續音中取得的二字詞，並且分別使用本論文所提及的兩種方法：一種是以音長為比例擷取相對應的基週，另一種則以 DTW 路徑來擷取相對應的基週。在平衡句的評估方面，我們從電信所提之平衡句中，切取一個連音二字詞，然後利用單音合成後，調整其音量及基週走勢，再放回平衡句中。也就是說，我們比較二種句子，一種是其中有某個二字詞是利用單音合成的平衡句，另一種則完全是原音。每一部份皆分別測試其自然度（ Naturalness ）和理解度（ Comprehensibility ）。

在自然度方面，我們從 60 個混著原音和合成音的二字詞中，隨機播放給聽者聽，請他們就所聽到的二字詞評分，分數從 1 到 100 分（ 100 分為最高分）。理解度方面也是從混合著原音和合成音的二字詞中，任意挑選出一句二字詞，然後請他們將所聽到的二字詞寫出。在平衡句方面的測驗亦同於二字詞。受測對象分別為 29 位國中、國小的老師，還有 13 位大學生，總共 42 位。

4.2 實驗結果

在二字詞的自然度方面，受測的二字詞中有原音及合成音，合成音部份又有兩種合成方式，第一種是根據音長的比例，均勻（ Uniform ）的取單音基週部份合成連音二字詞，而另一種則是在取基週時根據 DTW 的對應路徑來合成連音二字詞。在此測試部份，我們並不根據連音二字詞來調整基週走勢和音量大小。

結果顯示出，聽者對原音和合成音的喜好程度差異很小，表三是各種測試音的結果。絕對分數是聽者評分的平均分數，喜好度則是在某位聽者聽起來是三種中平均分數最高的，而厭惡度（ Dislike ），則是平均分數最低的。從表中可看出從連續音中抽取出之連音二字詞，單獨播放時可能比

合成音效果差，因為它本身包含了連續音中的韻律訊息，可能只適合在連續音中。一旦從連續音中獨立出來，就可能產生如表四的結果。

	絕對分數	喜好度	厭惡度
原音	72.5	9.5 %	59.5 %
Uniform	74.3	42.8 %	28.6 %
DTW Path	74.9	47.6 %	12.0 %

表三：二字詞的自然度結果

	正確率
原音	86.3%
Uniform	79.1%

表四：二字詞的正確率

在二字詞的理解度方面，其結果如表四所示，我們的評量計算方式是先求出原音二字詞的正確率為 86.3 %，再求出合成二字詞的正確率為 79.1 %，最後假設原音二字詞的正確率為 100 %，則相對的合成二字詞的正確率為 91.7 %。

就表三的結果而言，使用 DTW 路徑對應比 Uniform 對應有較佳的效果，但是 DTW Path 比 Uniform 耗費較大的處理過程，基於語音合成即時的考量，我們在使用於整句平衡句的合成時，只使用 Uniform 的方式。

在平衡句的測試方面，自然度的結果如表五所示，合成音和原音相對的自然度為 96.1 %，在這個測試中，從受測者的分數中可看出，百分之九十以上的聽者偏好原音。在理解度的測試方面，其結果如表六所示，原音

的理解度為 94.8 %，而合成音方面其理解度為 83.3 %，其相對之理解度可達到 87.8 %。

	絕對分數
原音	86.7
合成音	83.3

表五：平衡句的自然度結果

	正確率
原音	94.8%
Uniform	83.3%

表六：平衡句的正確率

4.3 實驗結果討論

從上一節的實驗結果中可看出，在二字詞的測試方面，我們得到一個不錯的結果，但是放回平衡句時，效果不是很好，我們觀察的結果，可能有二個問題存在：

- 1.調整基週走勢的方法
- 2.所使用單音和連續音的差異

在調整基週走勢的方法方面，我們只有試過一般的線性和基週重建，可以試試其它的調整方法，如 Pitch Synchronous Overlap and Add (PSOLA) 演算法[4][6]，許多文獻提及有不錯的效果。

在本篇論文中一直嘗試使用單音來模擬合成二字詞，但是單音和連續

音除了音長、音量的差異外，還有許多差異：第一，單音在錄製時期完全不含韻律訊息。第二，有些會產生變音、藕合效應的連續音（例如：『神機妙算』中的『機』受『妙』的影響而變成『ㄐ一ㄥ』），並不存在我們的單音庫中。因此，我們可嘗試從連續音庫來建單音庫，由經由適當的分類來建立單音庫，則這些單音可能已具備一些韻律的訊息，應該比我們現在用來合成的單音，更接近連續音。

5 結論

我們從事語音合成的目標在完成一套、能夠連續發音的中文文句翻語音系統，儘管有許多合成系統被提出，但是很少有文獻提出有關連音的探討。本篇論文從模擬連續語音中的連音二字詞開始，首先利用 DTW 對應找出單音中的子音和母音與連續音中的子音和母音是跟其音長呈等比例關係。在確定連音中點方面，我們使用倒頻譜係數差來改善原 DTW 的缺點，進而確定連音中點。再從已知之連音中點，找出連音段長度，到整個連音二字詞的完成。

整個實驗的結果說明，在二字詞的合成方面，我們從一個較長的句子中所切取之二字詞，反而不會比合成的二字詞好聽，根據此項訊息告訴我們，長度不同的句子應有不同的韻律訊息。可知韻律訊息在語音合成中之重要性。

我們希望能朝向自然的語音合成之路邁進，在做連音段的合成時可以在沒有連續音的資訊下，完成連音段的合成，這其實和我們的發音器官的變化性，還有聲道的頻譜特性有很大的相關性，我們一直很缺乏這方面的知識[1]，所以值得在此方面投入研究，以期獲得更佳的連音方法。

參考文獻

- [1] John R. Deller, Jr. John G. Proakis, and John H. L. Hansen, "Discrete-Time Processing of Speech Signals", MACMILLAN 1993。
- [2] 呂士南, 周同春, 初敏和陸亞民, "漢語合成系統中音高音長規則", 第三屆全國人機語音通訊學術會議論文集。中國, 1994。
- [3] 丁培毅, 張保忠和黃英峰, "以分段量測的動態時間歸正法則改善混淆群集的辨認", 電信研究季刊, 第19卷第三期, 1989。
- [4] Pei-yih Ting, Chun-Yu Tsai and Chi-Shi Liu, "The Post-Processing Stages of a Mandarin Waveform Synthesizer", in Proc. of 1994 International Computer Symposium Taiwan, 1994, p1262-p1266。
- [5] Carl D. Mitchell, Mary P. Harper and Leah H. Jamieson, "Using Explicit Segmentation to Improve HMM Phone Recognition", in Proc. of ICASSP 1995。
- [6] Eric Moulines and Francis Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", Speech Communication, September, 1990, p454-p467。
- [7] 陳志祥, "國語連續語音連音型態之初步研究", 中興大學應數研究所碩士論文, 1995。

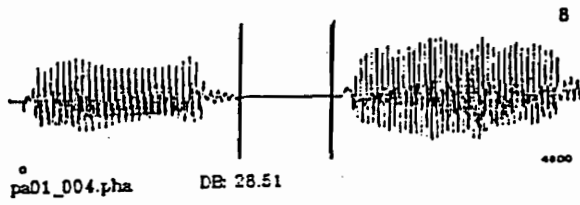


圖 1-a 停頓連接

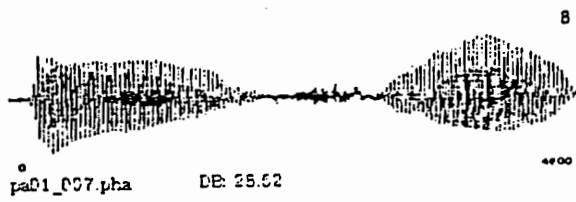


圖 1-b 緊密連接

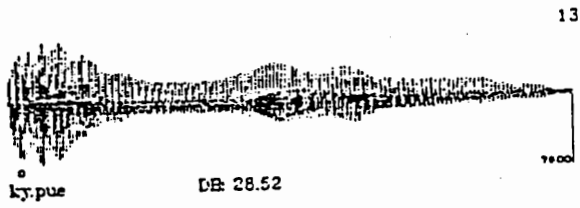


圖 1-c 重疊連接

圖 1 停頓連接、緊密連接和重疊連接之圖例

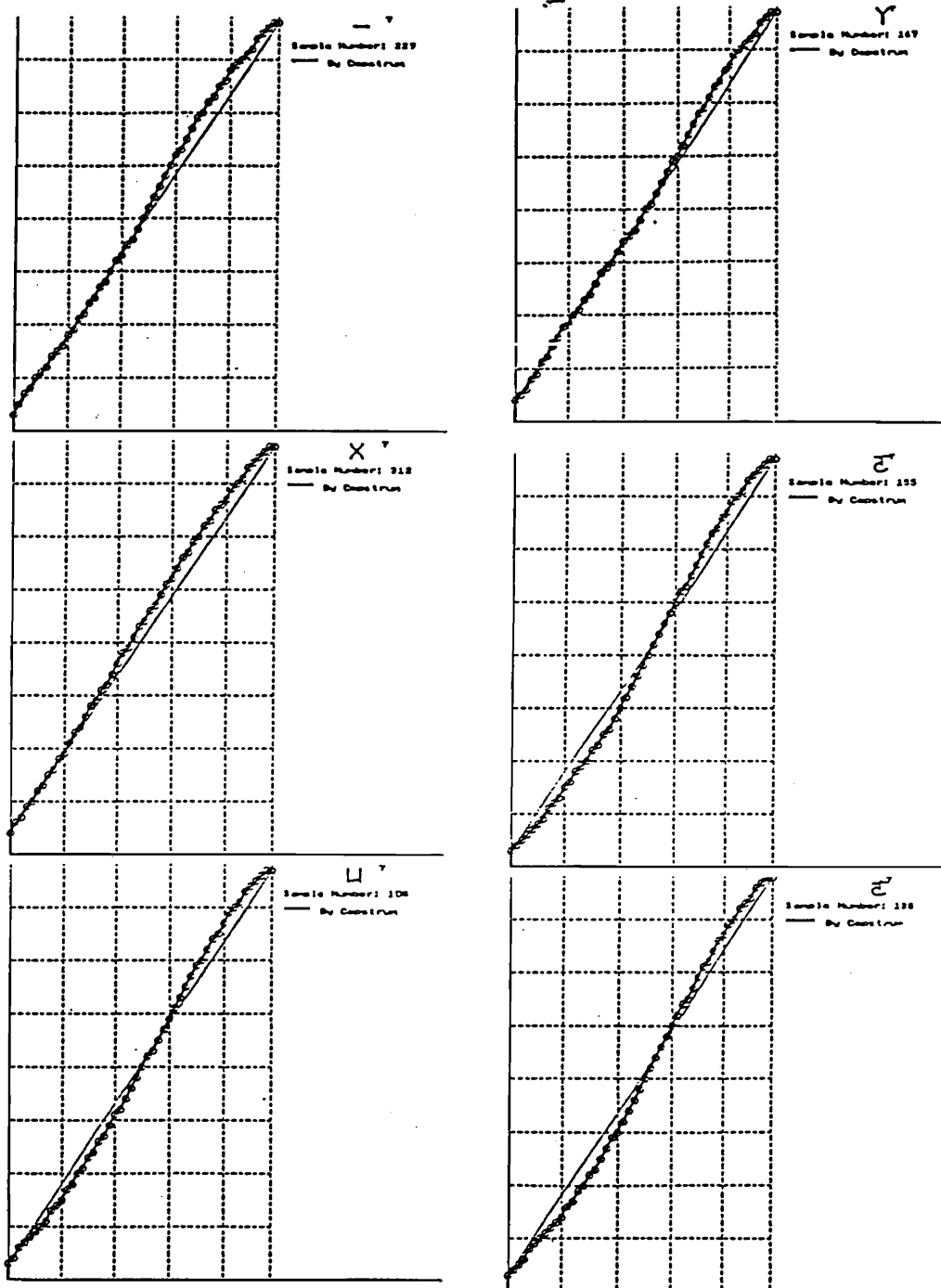
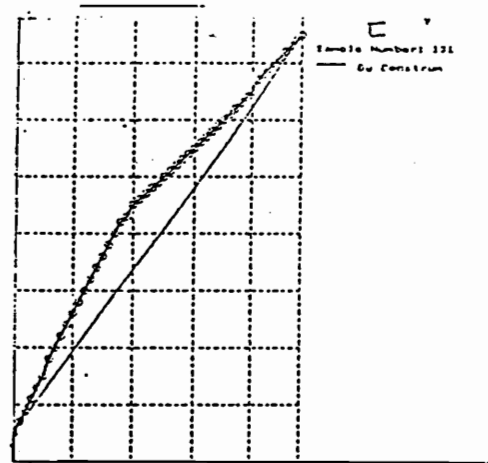
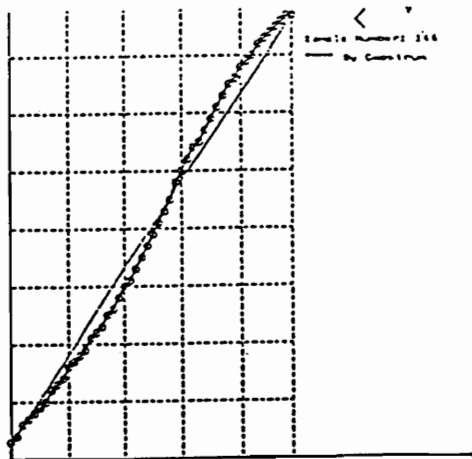
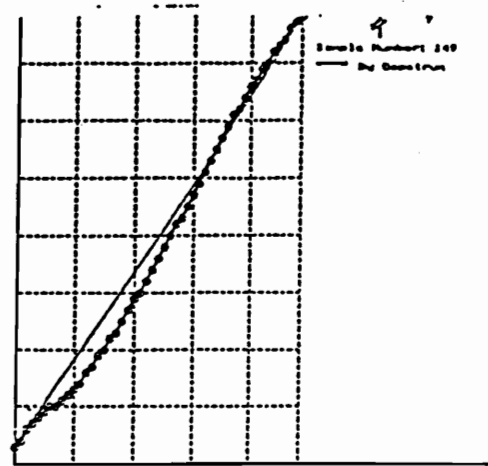
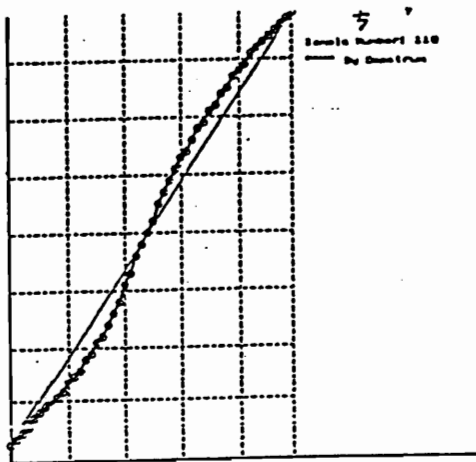


圖 2 母音類 DTW 路徑對應圖



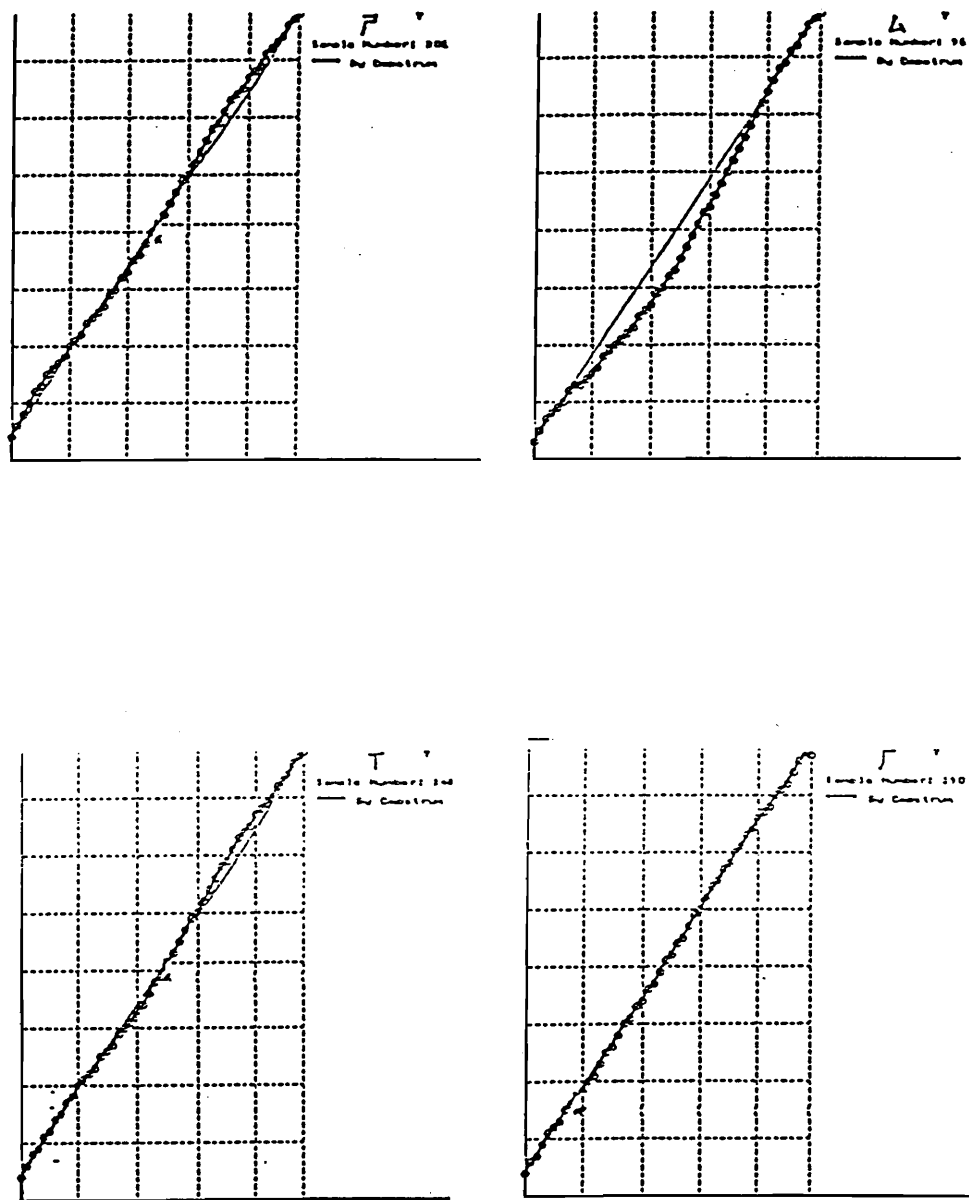


圖3 子音ㄈ、ㄇ、ㄊ、ㄍ之DTW
路徑對應圖

子音	詞首(含落單)	詞內(含詞尾)	單音(前)	單音(後)
ㄅ	1185.15	1364.19	1991.08	2103.13
ㄆ	1238.82	1250.57	1920.19	1947.56
ㄇ	1327.41	1261.3	2126.43	2114.36
ㄏ	405.28	256.51	781.31	815.52
ㄏ	1451.23	1521.33	2675.54	2680.7
ㄆ	1637.35	1473.1	2778.71	2547.13
ㄊ	1472.75	1288.62	2839	2697
ㄍ	1036.97		1895.96	1700.47

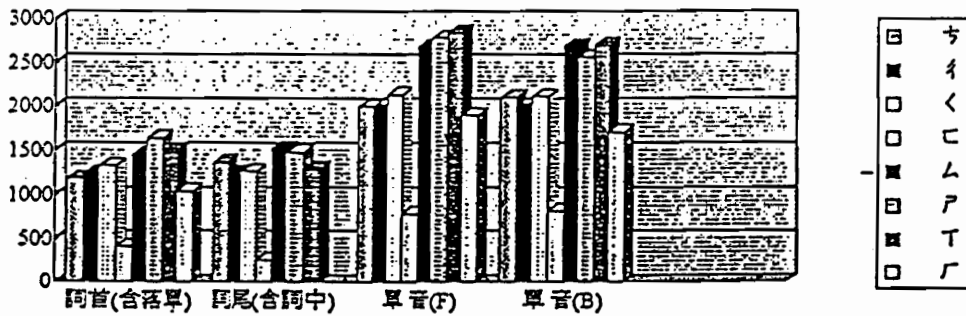
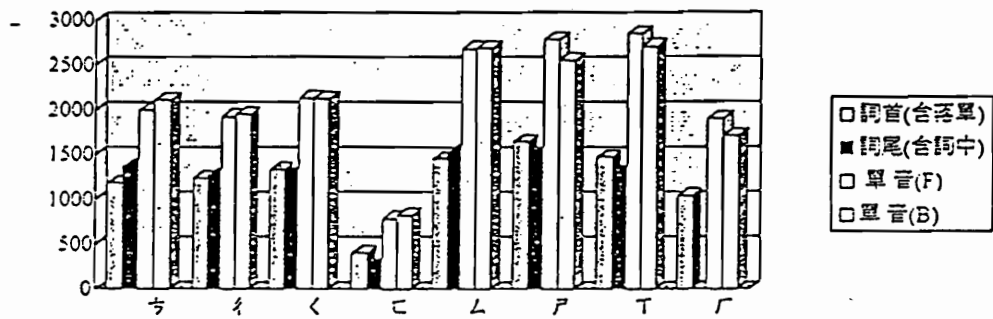


圖4 子音音長統計圖表



「南韓」的連音情形



「聯合」的連音情形

圖5 子音r在詞內產生連音的情形，“|”表示人工切音時所做的標記。

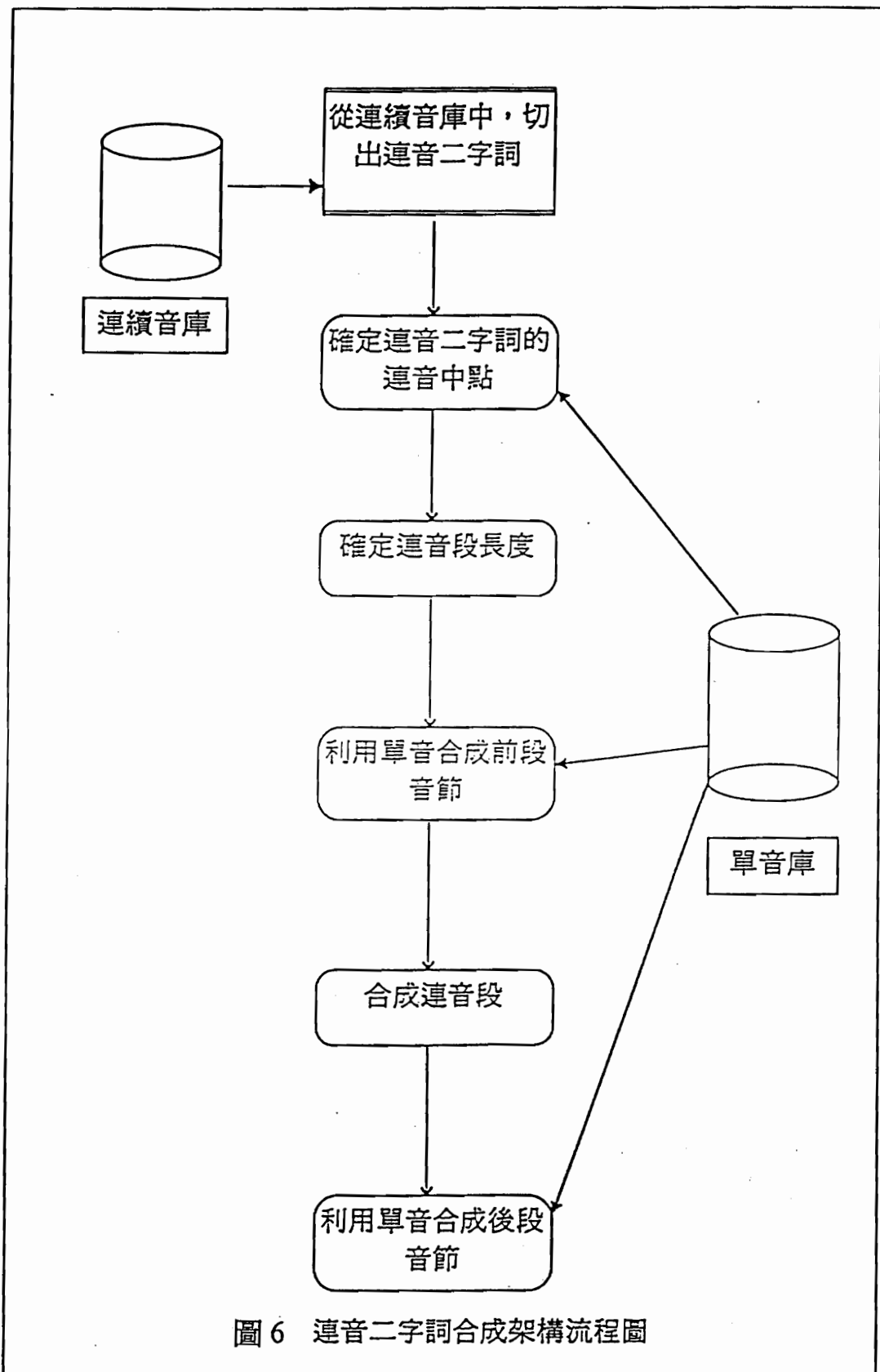




圖 7-a 連音二字詞『孤立』

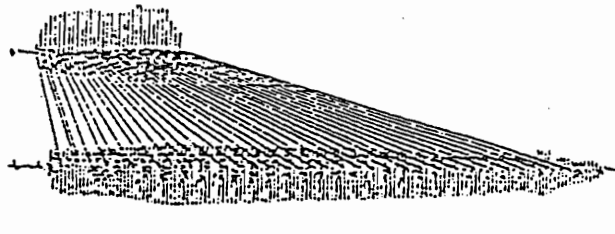


圖 7-b 連音中點音框與整個連音二字詞之距離曲線

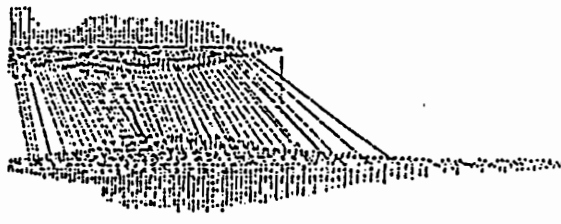


圖 7-c 圖 7-b 曲線之斜率變化走勢

圖 7 連音段之求取過程



前音節之 DTW 路徑對應圖示



後音節之 DTW 路徑對應圖示

圖 8 連音二字詞分別和前音及後音之以基週為單位之 DTW 對應關係圖

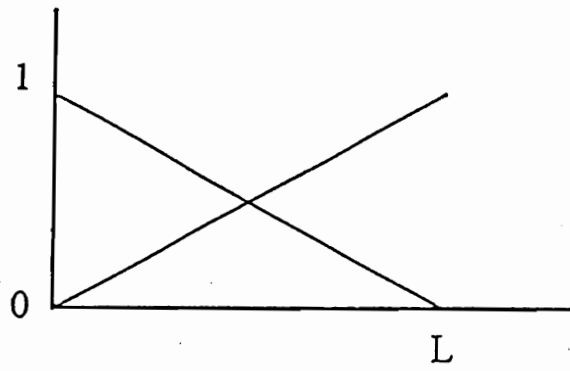


圖9 合成連音段之 A、B 係數走勢圖

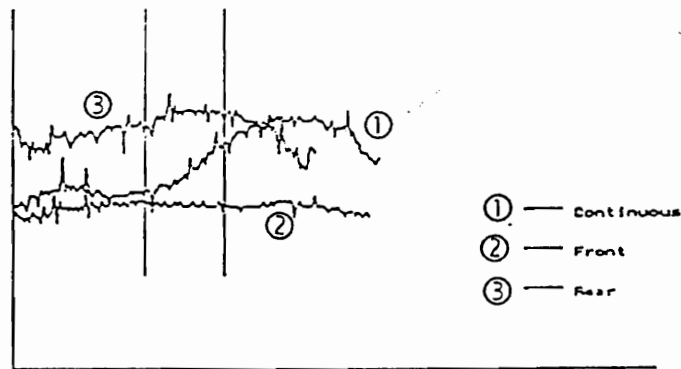


圖10 「官員」基週走勢比較圖，①是連音二字詞「官員」的基週走勢曲線，②是單音「官」的基週走勢曲線，③是單音「員」的基週走勢曲線