# 第2章 文獻回顧

## 2.1 以特徵為基礎之信心度評估

在過去文獻中，有許多信心度評估的相關研究都是為了找出可以有效判斷語音辨識結果正確性的預估特徵。通常這些預估特徵是在辨識過程中從聲學模型分數、語言模型分數、語法(Syntax)等三種不同的資訊收集得來的。一些聲學、語言模型或語法相關常見的特徵如下:

(i) 正規化聲學對數相似度(Normalized Acoustic Log Likelihood):聲學對數分數除以辨識結果所佔的音框個數。

(ii) *N*-最佳(*N*-best)詞序列相關特徵:先由詞圖(Word Graph，在 2.2.1 節會詳細的介紹)產生 $N$ 條分數最高的詞序列。其相關的特徵有:某個候選詞在 $N$ 條最佳詞序列中出現的次數，如式(2.1) :

$$\frac{\sum_{w \in Wn} 1}{N} \tag{2-1}$$

其中 $w$ 代表一個詞，而 $W_n$ 代表在*N*-最佳詞序列中分數排名第 $n$ 高的詞序列;或是前*N*條中包含某個辨識結果的相似度權重比例(Weighted Ratio)，如式(2.2)所示:

$$\frac{\sum_{w \in Wn} f(n)P(W_n \mid X)}{\sum_{n=1}^{N} f(n)P(W_n \mid X)} \tag{2-2}$$

其中 $f(n)$ 為權重函式， $P(W_n \mid X)$ 代表聲學觀測序列 $X$ 產生詞序列 $W_n$ 的機率。

(iii) 聲學穩定度(Acoustic Stability):其主要概念為我們在尋找最佳詞序列時，可在語言模型分數的部份使用不同的權重β ，如式(2-3)所示:

$$p(X \mid W)P(W)^{\beta} \tag{2-3}$$

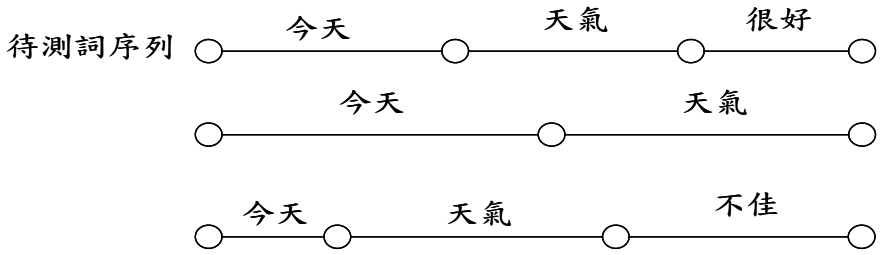


圖 2-1 三個不同語言模型權重所產生的詞序列

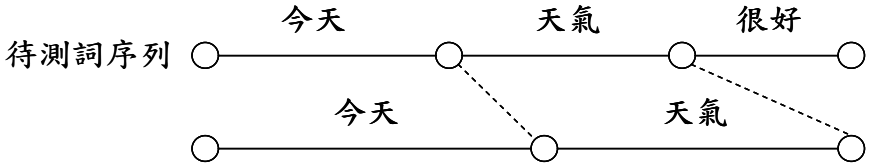


圖 2-2 第一條及第二條詞序列做完比對後之結果

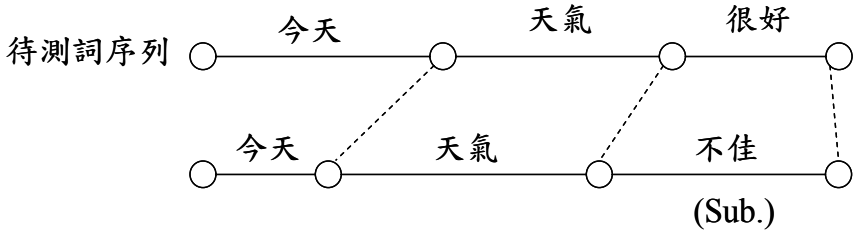


圖 2-3 第一條及第三條詞序列做完比對後之結果

因而得到不同的最佳詞序列。接下來針對某一個特定語言模型權重的詞序列(也就是實際上真正需做信心度評估的輸出結果，通常會事先由一組訓練資料中，找出能使訓練資料的辨識率為最高的權重值)，與其它不同語言模言權重所得到的詞序列做比對(利用Levenshtein Alignment)，計算某個辨識結果是否有出現在其它詞序列的一個位置。舉例來說，假設我們現在有三條不同語言模型權重所產生的詞序列，如圖 2-1所示。其中第一條是我們辨識器最後輸出的結果，也就是這裡所謂的待測詞序列。因此我們必須將其結果分別與第二條與第三條詞序列做比對，其比對後的結果如圖 2-2及圖 2-3所示。其中(Sub.)表示詞的替代(Substitution)。根據這三個圖，我們便可以知道在待測詞序列中，”今天”跟”天

氣”與其它兩條詞序列中都對齊至同一位置，其聲學穩定度的值便為2。而”很好”則沒有任何一條詞序列的詞對齊至同一位置，其聲學穩定度的值便是0。

(iv) 候選詞假設密度(Word Hypothesis Density):計算詞圖(Word Graph)中，在一個詞段(Word Arc)中，平均其它詞段出現的個數( $HD(\bullet)$ )，可以下列數學式表示

$$D(t') = \left|\{b : [w_b; s_b, e_b] \in WG \wedge (s_b \le t' \le e_b)\}\right| \tag{2-4}$$

$$HD(a : [w_a; s_a, e_a]) = \frac{1}{e_a - s_a + 1} \sum_{t=s_a}^{e_a} D(t) \tag{2-5}$$

其中 $w_a$ 代表詞圖 $WG$ 的一個詞， $s_a$ 及 $e_a$ 分別為 $w_a$ 的開始及結束時間， $D(t')$ 則是代表在詞圖中有多少不同詞段其開始時間及結束時間有包含 $t'$ 這個時間點。

(v) 持續時間(Duration)相關之特徵:一般而言，詞、音節或音素等辨識單位各自的持續時間差異性不太，因此，持續時間也可算是一個適合的預估特徵。

(vi) 語言模型:包括語言模型分數或語言模型回退(Back-off)行為[Chen and Goodman 1999]。假設我們現在有一個詞序列，包含 $w_1$ 、 $w_2$ 及 $w_3$ 三個詞。在我們要估計這個詞的三連語言模型時，加上回退行為，其計算可能有下列四種情況:

A. $w_1$ 、 $w_2$ 及 $w_3$ 此序列在訓練語料中的確存在，可由訓練語料直接估得此三連語言模型的機率 $P(w_3 \mid w_2, w_1)$ 。

B. 無法由訓練語料直接估得 $P(w_3 \mid w_2, w_1)$ 此三連機率，需由二連－二連語言模型回退行為所估得，如式(2-6)所示:

$$P(w_3 \mid w_2, w_1) \approx P(w_2 \mid w_1) + P(w_3 \mid w_2) \tag{2-6}$$

其中 $P(w_2 \mid w_1)$ 及 $P(w_3 \mid w_2)$ 可直接由訓練語料求得。

C. 無法由訓練語料直接估得 $P(w_3 \mid w_2, w_1)$ 此三連機率及 $P(w_2 \mid w_1)$ 二連機率，需由單連-二連語言模型回退行為所估得，如式(2-7)所示 :

$$P(w_3 \mid w_2, w_1) \approx P(w_1) + P(w_3 \mid w_2) \tag{2-7}$$

其中 $P(w_1)$ 及 $P(w_3 \mid w_2)$ 可直接於訓練語料估計。

D. 無法由訓練語料直接估得 $P(w_3 \mid w_2, w_1)$ 及 $P(w_3 \mid w_2)$ 此二項機率，需由二連－單連語言模型回退行為所估得，如式(2-8)所示:

$$P(w_3 \mid w_2, w_1) \approx P(w_3) + P(w_2 \mid w_1) \tag{2-8}$$

其中 $P(w_2 \mid w_1)$ 及 $P(w_3)$ 可直接由訓練語料求得。

而對上述四項情況，我們可以分別給定不同的信心度值。在[Uhrik and Ward 1997]中提到:能直接在訓練語料中求得的語言模型，其信心度會比較高，而如果需要回退行為才能求得的語言模型機率，其信心度較低。

(vii) 與語法剖析相關之特徵:一個詞是否能被文法(Grammar)正確地剖析[Sarikaya *et al*. 2005]。

有關預估特徵更詳細的資訊可以進一步參考[Cox and Rose 1996; Schaaf and Kemp 1997; Chase 1997; San-Segundo *et al*. 2001; Sanchis *et al*. 2003; Lane and Kawahara 2005; Benitez *et al*. 2000]。

在本節中我們將以自然貝氏分類器(Naïve Bayes Classifier)為例，介紹如何根據以上這些特徵值來判斷辨識結果的正確性。為了要使用自然貝氏分類器來從事信心度評估，我們必須事先定義兩種類別:$C_1$ 代表語音辨識結果為正確，$C_2$ 代表語音辨識結果為錯誤。另外，需為每一個辨識詞序列中的每一個辨識詞 $w$ 找出一組預估特徵參數向量 $\vec{f}$ (每一維代表一種預估特徵)，之後再利用貝氏定理可為每一個 $w$ 求取屬於 $C_1$ 之信心度，亦即

$$P(C_1 \mid \vec{f}, w) = \frac{P(C_1 \mid w) P(\vec{f} \mid C_1, w)}{\sum_{C'=C_1 \text{or} C_2} P(C' \mid w) P(\vec{f} \mid C', w)} \tag{2-9}$$

為了增加運算速度及減少參數估測量，可假設當給定詞及類別資訊 $C_i$ 時，每一維的特徵 $f_d$ 為互相獨立，所以 $P(\vec{f} \mid C_i, w)$ 可表示成:

$$P(\vec{f} \mid C_i, w) = \prod_{d=1}^{D} P(f_d \mid C_i, w) \tag{2-10}$$

而 $P(C_i \mid w)$ 與 $P(f_d \mid C_i, w)$ 中的每個式子皆可利用最大相似度估測(Maximum Likelihood Estimation, MLE)，由頻率統計(Frequency Count)求得:

$$P(C_i \mid w) = \frac{N(C_i, w)}{N(w)} \tag{2-11}$$

$$P(f_d \mid C_i, w) = \frac{N(f_d, C_i, w)}{N(C_i, w)} \tag{2-12}$$

其中 $N(\bullet)$ 代表某個事件(event)出現的次數，而上述兩式有機率為零的問題，可使用一般在語言模型常使用的絕對折扣平滑(Absolute Discounting Smoothing)技術來解決:

$$P(C_i \mid w) = \begin{cases} \dfrac{N(C_i, w) - b}{N(w)} & \text{if } N(C_i, w) > 0 \\ \dfrac{b}{N(w)} & \text{if } N(C_i, w) = 0 \end{cases} \tag{2-13}$$

$$P(f_d \mid C_i, w) = \begin{cases} \dfrac{N(f_d, C_i, w) - b}{N(C_i, w)} & \text{if } N(f_d, C_i, w) > 0 \\ M \dfrac{P(f_d \mid C_i)}{\sum_{f'_d : N(f'_d, C_i, w) = 0} P(f'_d \mid C_i)} & \text{if } N(f_d, C_i, w) = 0 \end{cases} \tag{2-14}$$

其中式(2-13)成立的原因為 $C_i$ 在這裡的討論假設僅只有分正確及不確定兩類，亦即 $C_i \in \{C_1, C_2\}$。而 $b$ 為介於0~1的數字，式(2-14)中的 $M$ 則代表 $N(f_d, C_i, w)$ 為非零值的事件個數乘上非零值的事件扣掉的值 $\dfrac{b}{N(C_i, w)}$。在最後決定信心度評估時，我們是計算 $P(C_1 \mid \vec{f}, w)$ 的值。當其值大於事先設定的門檻值時，便認為辨識詞為正確之語音辨識結果，否則就當成是錯誤的語音辨識結果。

一般而言，由於每個上述聲學或語言模型等方面的預估特徵都有極高的相關性[Kemp and Schaaf 1997; Schaff and Kemp 1997]。因此，即使將所有較有用的預估特徵都集合起來使用，對效能也不一定會有很大的提昇(與單一最有用的特徵比較)。因此，近來有學者嘗試將單一最有用的聲學預估特徵(如事後機率或*N*-最佳詞序列相關

特徵)，與純粹的語言特徵，如語意剖析(Semantic Parsing)或其它與語意相關資訊相互結合[Zhang and Rudnicky 2001; Guo *et al*. 2004]，的確對效能有些許的提昇。

## 2.2 事後機率之信心度評估

先前於1.1.4小節提到，傳統自動語音辨識演算法是在給定任何的聲學觀測序列(Acoustic Observation Sequence) $X$ 時，使用最大事後機率決策方式找出一條最有可能的詞序列 $\hat{W}$，使得它有最大的事後機率 $P(\hat{W} \mid X)$:

$$\begin{aligned} \hat{W} &= \arg\max_{W \in \overline{\mathbf{W}}_{\Sigma}} P(W \mid X) \\ &= \arg\max_{W \in \overline{\mathbf{W}}_{\Sigma}} \frac{p(X \mid W)P(W)}{p(X)} \\ &= \arg\max_{W \in \overline{\mathbf{W}}_{\Sigma}} p(X \mid W)P(W) \end{aligned} \tag{2-15}$$

其中 $\overline{\mathbf{W}}_{\Sigma}$ 代表語言中所有詞序列的組合，$P(W)$ 為 $W$ 的語言模型機率，$p(X)$ 為聲學觀測序列 $X$ 的事前機率，而 $p(X \mid W)$ 則是代表假設 $W$ 為辨識結果的情況下，產生 $X$ 的機率(也就是之前提到的聲學模型)。通常來說，此事後機率 $P(W \mid X)$ 對辨識結果的詞序列是一個很好的信心度評估。但是由於實際運用上的考量，在使用式(2-15)當作自動語音辨識的方法時，因為分母項 $p(X)$ 不影響詞序列排序，我們都會忽略 $p(X)$。這也說明了為什麼語音系統辨識結果的分數是不適合用來當作評估辨識結果可靠度的依據。但只要我們將語音系統辨識結果的分數再除以 $p(X)$ 的值，那麼此新的數值便是介於0~1的定量數值，可以用來判斷 $X$ 跟 $W$ 之間匹配的程度。就理論來說，我們可以依式(2-16)計算 $p(X)$ 的值:

$$p(X) = \sum_{W \in \overline{\mathbf{W}}_{\Sigma}} p(X, W) = \sum_{W \in \overline{\mathbf{W}}_{\Sigma}} P(W)p(X \mid W) \tag{2-16}$$

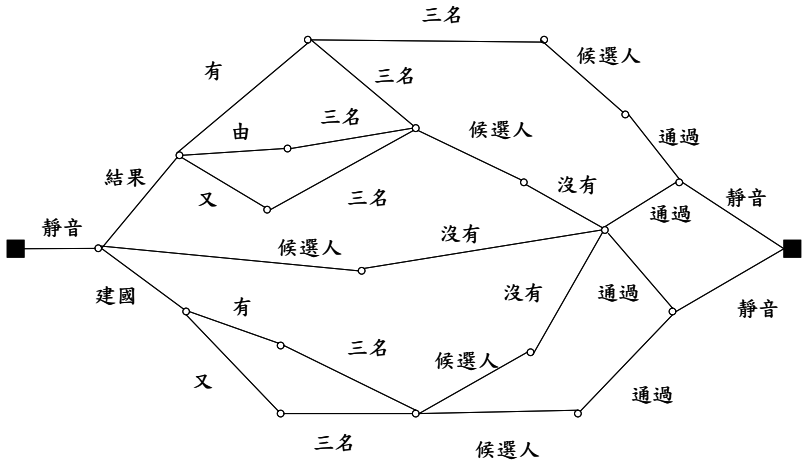


圖 2-4 詞圖 $\Psi^X$，為所有可能詞序列 $\overline{\mathbf{W}}_\Sigma$ 的近似表示

其中 $W$ 代表 $X$ 一個可能的辨識詞序列。顯而易見的，如果我們沒有對此假設做一些限制，要求得 $p(X)$ 是一件很困難的事，畢竟我們無法加總語言中的所有可能詞序列。所以我們通常會做一些限制或是用近似的方法計算 $p(X)$。如1.2.2小節所提到的，通常會有兩類方法，亦即圖形化基礎法與填充化基礎法，而本論文主要是討論圖形化基礎法。

在圖形化基礎法中，通常會先對每一個 $X$ 產生對應的詞圖 $\Psi^X$，如圖 2-4所示，由於只有 $p(X|W)P(W)$ 機率較大的詞序列才會有可能留在 $\Psi^X$ 中。因此，詞圖可說是所有可能詞序列的近似表示，以用來求得式(2-16)的近似值。在有了詞圖及每條詞序列的事後機率後，可進而求得每個辨識詞的事後機率。在下面的小節中，本論文將會更詳細的介紹有關於事後機率的計算及其相關變形。

### 2.2.1 詞圖簡介

詞圖為一有向性，非環狀(Acyclic)的圖形表示法，圖 2-4中的每一個節點代表一個時間點，每個詞段(Word Arc) $a$ 由三個變數表示，$a:[w_a;s_a,e_a]$，其中 $w_a$ 代表其對應的詞編號為何，$s_a$ 代表詞段開始時間，$e_a$ 則代表結束時間。每個詞段通常會給定某種分數，較常見的為此詞段產生語音段落的聲學分數 $p(X_{s_a}^{e_a}|w_a)$。而每個詞圖都會有兩個特殊的節點，分別代表詞圖的開始及結束(如圖 2-4的兩個實心點)。只要是從開

始節點到結束節點的任何路徑都可視為一條完整路徑(Complete Path)，而任一條完整路徑都可以代表聲學觀測序列 $X$ 的一條可能辨識詞序列。

2.2.2 計算一個詞段的事後機率

根據2.2.1小節的說明，我們可以將詞圖 $\Psi^X$ 上某個詞段 $a:[w_a;s_a,e_a]$ 的事後機率，$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)$(在詞圖上計算一個詞段的事後機率，與原本給定一個聲學觀測序列 $X$ 計算事後機率並不相同, 因為 $\Psi^X$ 只為 $\overline{\mathbf{W}}_\Sigma$ 之近似表示)看成是所有通過這個詞段的所有完整路徑的分數總和除以詞圖上所有完整路徑分數總和，如式(2.17)所示:

$$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)=\frac{\sum_{\{\overline{W}:[w^n;s^n,e^n]_{n=1}^{N}\}\in\Psi^X,\,a\subset\overline{W}}\left\{\prod_{n=1}^{N}p(X_{s^n}^{e^n}\,|\,w^n)\cdot P(w^n\,|\,h^n)\right\}}{\sum_{\{\overline{W}':[w^m;s^m,e^m]_{m=1}^{M}\}\in\Psi^X}\left\{\prod_{m=1}^{M}p(X_{s^m}^{e^m}\,|\,w^m)\cdot P(w^m\,|\,h^m)\right\}} \quad (2\text{-}17)$$

其中，$\overline{W}$ 代表在詞圖的一條完整路徑，共有 $N$ 個詞段，$a\subset\overline{W}$ 代表包含詞段 $a$ 的完整路徑 $\overline{W}$，$h^x$ 為 $w^x$ 的詞歷史(Word History)。$p(X_{s^n}^{e^n}\,|\,w^n)$ 代表開始時間 $s^n$ 至結束時間 $e^n$ 此段聲學觀測序列的聲學相似度，而 $P(w^n\,|\,h^n)$ 代表其語言模型分數。$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)$ 可以用前向後向(Forward-backward)演算法有效率地求解。演算法詳細過程如圖 2-5所示，其主要概念為在計算通過某詞段 $a:[w_a,s_a,e_a]$ 的所有完整路徑分數總和時，首先加總由其結束時間為 $s_a-1$ 的所有詞段其轉移至此詞段機率乘上先前每個詞段所累積的分數，最後乘上 $a:[w_a,s_a,e_a]$ 的聲學分數，代表由從 $t$ 為0至此詞段的所有路徑分數總和(也就是所謂的"前向"部份，如圖 2-5中前兩個最外層的for迴圈);接著再加總由其開始時間為 $e_a+1$ 的詞段其轉移至此詞段機率乘上每個開始時間為 $e_a+1$ 的詞段先前所累積的分數及其對應的聲學分數，做為由 $t=T-1$ 至 $e_a$ 的所有路徑分數總和(也就是所謂的"後向"路份，如圖 2-5中第三及第四個最外層的for迴圈)。將此前向及後向的分數相乘之後，便是通過詞段 $a:[w_a,s_a,e_a]$ 的所有完

整路徑分數總和。只要將所有結束時間點為$t = T - 1$的詞段對應之前向分數加總，即可獲得詞圖所有完整路徑總和。

假設 $a$ 與 $r$ 均為詞圖中之詞段，$P(a|r)$ 表示由詞段 $r:[w_r;s_r,e_r]$ 至詞段 $a:[w_a;s_a,e_a]$ 的轉移機率(Transition Probability)，$\gamma_a$ 則為詞段 $a$ 的事後機率，$\kappa$ 為聲學分數的權重。詞圖起始時間為 0，結束時間為 $T$-1

for 開始時間為 0 的詞段 $a$

$\alpha_a = p(X_{s_a}^{e_a} | w_a)^{\kappa}$

end //前向初始化

for $t$=1 to $T$-1

for 開始時間為 $t$ 的詞段 $a$

$\alpha_a = 0$

for 結束時間為 $t$-1 的詞段 $r$

$\alpha_a = \alpha_a + \alpha_r P(a|r)$

end

$\alpha_a = \alpha_a \cdot p(X_{s_a}^{e_a} | w_a)^{\kappa}$

end

end //前向遞迴

for 結束時間為 $T$-1 的詞段 $a$

$\beta_a = 1$

end //後向初始化

for $t$=$T$-2 to 0

for 結束時間為 $t$ 的詞段 $a$

$\beta_a = 0$

for 開始時間為 $t$+1 的詞段 $r$

$\beta_a = \beta_a + \beta_r P(r|a) p(X_{s_\gamma}^{e_\gamma} | r)^{\kappa}$

end

end

end //後向遞迴

for 每一詞段 $a$

$$\gamma_a = \frac{\alpha_a \beta_a}{\sum_{\text{所有在時間}T-1\text{結束的詞段}a} \alpha_a}$$

end

圖 2-5 利用前向後向演算法求得詞圖中某一詞段的事後機率

2.2.3 計算一個辨識詞的信心度

基本上，我們可以直接使用某個詞段 $a:[w_a;s_a,e_a]$ 的事後機率，$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)$，當作是辨識詞 $w_a$ 的信心度評估，便能得到一定的效果。但我們知道，在詞圖中，對詞段 $a:[w_a;s_a,e_a]$ 而言，會有許多詞段 $\gamma$ 與詞段 $a$ 對應到相同的詞編號，只是開始時間及結束時間有些許不同，因此求某個詞 $w_a$ 的信心度時，除了考慮自己本身的事後機率 $P(a:[w_a;s_a,e_a]\,|\,\Psi^X)$ 之外，也可以加入那些開始及結束時間有些微的差距，但是詞編號是一樣的詞段。在前人的研究中，[Wessel *et al.* 2001] 提出了三個計算方法:

$$C_{\text{sec}}(a:[w_a;s_a,e_a]) = \sum_{\substack{r:[w_r;s_r,e_r],w_r=w_a \\ \{s_r,\ldots,e_r\}\cap\{s_a,\ldots,e_a\}\neq\phi}} P(r:[w_r;s_r,e_r]\,|\,\Psi^X) \tag{2-18}$$

$$C_{med}(a:[w_a;s_a,e_a]) = \sum_{\substack{r:[w_r;s_r,e_r],w_r=w_a \\ s_r\le[(s_a+e_a)/2]\le e_r}} P(r:[w_r;s_r,e_r]\,|\,\Psi^X) \tag{2-19}$$

$$C_{\max}(a:[w_a;s_a,e_a]) = \max_{t\in\{s_a,\ldots,e_a\}} \sum_{\substack{r:[w_r;s_r,e_r],w_r=w_a \\ s_r\le t\le e_r}} P(r:[w_r;s_r,e_r]\,|\,\Psi^X) \tag{2-20}$$

其中 $P(r:[w_r;s_r,e_r]\,|\,\Psi^X)$ 的算法可參考式(2-17)。式(2-18)的意思為將時間上與現在這個詞段 $a:[w_a;s_a,e_a]$ 相交，而且對應詞編號相同的所有詞段事後機率相加，當作辨識詞 $w_a$ 的信心度。而式(2-19)則是累加詞編號相同，但是必須與詞段 $a:[w_a;s_a,e_a]$ 的時間中點相交之詞段事後機率。最後，式(2-20)則是不單單只看與 $a:[w_a;s_a,e_a]$ 時間中點相交的相同詞編號詞段之累加事後機率，而是也考慮到相交於 $s_a$ 到 $e_a$ 之中任意時間點相同詞編號詞段之累加的事後機率，然後再取有最大累積值的時間點之分數作為辨識詞 $w_a$ 的信心度。

另外，在[Lo and Soong 2005]中則是提到對於在計算 $w_a$ 的事後機率時，應該注意的幾點事項:

(i) 縮減的搜尋空間(Reduced Search Space):如2.2.2小節所提到的，由於我們無法針對語言中所有可能的詞序列做加總的動作，另外也為了避免搜尋空間太過龐大,所以通常都是在一些較簡化的搜尋空間(如詞圖或*N*-最佳詞序列)來計算辨識結果的事後機率。

(ii) 放寬時間的限制:因為一個詞段是由詞編號、開始時間及結束時間三項要素所構成，而辨識詞其開始及結束時間會在搜尋最佳詞序列時，因其搜尋空間大小而有許多不同的可能。因此，針對有不同的開始及結束時間，但其時間點有重疊(Overlap)且詞編號為相同的詞段，應該視為同樣的詞段。

(iii) 給定聲學及語言模型分數不同的權重:考慮這項因素的原因，主要是因為下列二項特徵:

A. 聲學模型分數區間範圍為 0 到正無窮大，但語言模型的分數卻只介於 0~1 之間

B. 每個音框都會計算其聲學模型分數，而語言模型分數則通常是在經過一段時間後才會計算(於 3.1.4 小節說明此緣由)

綜合了以上各項要點後，在[LO and Soong 2005]中，一個詞段的事後機率計算如下:

$$P(a:[w_a;s_a,e_a]\,|\,\Psi^X)=\frac{\sum_{\substack{[w^n;s^n,e^n]_{n=1}^N\in\Psi^X\\ \exists i,1\le i\le N,w^i=w_a,(s_a,e_a)\cap(s^i,e^i)\neq\phi}}\left\{\prod_{i=1}^{N}\ p^{\alpha}(X_{s^i}^{e^i}\,|\,w^i)\cdot P^{\beta}(w^i\,|\,h^i)\right\}}{\sum_{[w^m;s^m,e^m]_{m=1}^M\in\Psi^X}\left\{\prod_{m=1}^{M}\ p^{\alpha}(X_{s^m}^{e^m}\,|\,w^m)\cdot P^{\beta}(w^m\,|\,h^m)\right\}}\quad(2\text{-}21)$$

其中 $\alpha$ 及 $\beta$ 分別代表聲學及語言模型的權重。

而在[Sanchis *et al*. 2004]中，作者不再只對單一的詞圖作計算，而是在多個詞圖估算信心度評估，其演算法如圖 2-6所示:

令 $M_w$ 為詞典中所有的詞，$\Psi_t^X$ 為在某個時間點 $t$ 時所建構時間持續從 0 至 $t$ 的詞圖，WPost 則是取得當時詞圖上每個詞段的事後機率的函數，$\gamma$ 則為詞段的事後機率，$C(w,t)$ 用來儲存每個詞 $w \in M_w$ 的每個時間點 $t \in \{0,\ldots,T-1\}$ 之事後機率總和。

for 對每個詞 $w \in M_w$

  for $t$=0 to $T$-1

    $C(w,t) = 0$ //初始化

  end

end

for $t$=0 to $T$-1

  $A = \text{WPost}(\Psi_t^X)$

  for 每個詞段 $a:([w_a;s_a,e_a],\gamma) \in \Psi_t^X$

    for $t' = s_a$ to $e_a$

      $C(w_a,t') = C(w_a,t') + \gamma$

    end

  end

end

令 $h_{best}$ 為詞圖 $\Psi_{T-1}$ 中分數最高的詞序列

for 每一詞段 $a:[w_a;s_a,e_a] \in h_{best}$

$$W_{avg}(w_a;s_a,e_a) = \frac{1}{e_a - s_a + 1}\sum_{t=s_a}^{e_a} \frac{C(w_a,t)}{T-t+1}$$

$$W_{\max}(w_a;s_a,e_a) = \max_{s_a \le t \le e_a} \frac{C(w_a,t)}{T-t+1}$$

圖 2-6 在多重詞圖上求辨識詞 $w$ 信心度

在圖 2-6中，$W_{avg}$ 與 $W_{\max}$ 便可以作為辨識詞 $w$ 兩種的信心度評估。另一方面，詞段上的分數除了可以是聲學分數外，也可以是語言模型分數，或是聲學及語言模型分數的組合。

在[Razik *et al.* 2005]中也提出了幾種詞段信心度評估的作法。首先，最簡單的概念為在計算每個詞段的信心度時，將此詞段的聲學分數乘上單連語言模型的分數，而分母項的部份只考慮有同樣的開始、結束時間但詞編號不相同的其它詞段，如式(2-22)所示:

$$C(a:[w_a;s_a,e_a]) = \frac{p(X_{s_a}^{e_a} \mid w_a)\cdot P(w_a)}{\sum_{\substack{r:[w_r,s_r,e_r] \\ s_r=s_a,e_r=e_a}} p(X_{s_r}^{e_r} \mid w_r)\cdot P(w_r)} \tag{2-22}$$

不過，由於這樣的限制太嚴格，會導致$C(a:[w_a;s_a,e_a])$容易趨近於1。因此，在[Razik *et al.* 2005]中試著放寬關於開始、結束時間或詞段時間長度的限制。但由於詞圖上有許多詞段都符合現在所規定的限制，因此在決定分子或分母項的聲學分數$p(X_s^e \mid w)$時，便決定取相同的詞編號且符合目前的限制，具有最大聲學分數的詞段:

$$C(a:[w_a;s_a,e_a]) = \frac{\max_{\substack{a':[w_{a'};s_{a'},e_{a'}] \\ w_{a'}=w_a,a'\in E}} p(X_{s_{a'}}^{e_{a'}} \mid w_{a'})^{\alpha}\cdot P(w_a)^{\beta}}{\sum_{r:[w_r;s_r,e_r]\in E} \max_{\substack{r':[w_{r'};s_{r'},e_{r'}] \\ w_{r'}=w_r,r'\in E}} p(X_{s_{r'}}^{e_{r'}} \mid w_{r'})^{\alpha}\cdot P(w_r)^{\beta}} \tag{2-23}$$

其中$\alpha$及$\beta$分別代表聲學及語言模型的權重，而$E$代表那些符合放寬時間及詞段長度後但為相同詞編號的其它詞段，如開始時間及結束時間不一定要相同，只要有交集就可以或其它任意放寬的限制。值得注意的是式(2-22)及式(2-23)都只考慮到單連語言模型，這樣的資訊可能太過於粗糙，所以作者便進一步考慮藉由使用雙連語言模型來包含前後詞的資訊。首先，只多看前一個詞的話，便可將式(2-23)改為:

$$C(a:[w_a;s_a,e_a]) = \frac{\max_{\substack{a':[w_{a'};s_{a'},e_{a'}] \\ w_{a'}=w_a,a'\in E}} p(X_{s_{a'}}^{e_{a'}} \mid w_{a'})^{\alpha}\cdot \sum_{\substack{a_p:[w_{a_p};s_{a_p},e_{a_p}] \\ e_{a_p}=s_{a'}-1}} P(w_a \mid w_{a_p})^{\beta}}{\sum_{r:[w_r;s_r,e_r]\in E} \max_{\substack{r':[w_{r'};s_{r'},e_{r'}] \\ w_{r'}=w_r,r'\in E}} p(X_{s_{r'}}^{e_{r'}} \mid w_{r'})^{\alpha}\cdot \sum_{\substack{r_p:[w_{r_p};s_{r_p},e_{r_p}] \\ e_{r_p}=s_{r'}-1}} P(w_r \mid w_{r_p})^{\beta}} \tag{2-24}$$

若再多考慮後面一個詞，最後的式子便可以寫成:

$$C(a:[w_a;s_a,e_a]) = \frac{\max\limits_{\substack{a':[w_{a'};s_{a'},e_{a'}]\\ w_{a'}=w_a, a'\in E}} (p(X_{s_{a'}}^{e_{a'}} \mid w_{a'}))^{\alpha} \cdot \Gamma_{a':[w_{a'};s_{a'},e_{a'}]}}{\sum\limits_{r:[w_r;s_r,e_r]\in E} \max\limits_{\substack{r':[w_{r'};s_{r'},e_{r'}]\\ w_{r'}=w_r, r'\in E}} (p(X_{s_{r'}}^{e_{r'}} \mid w_{r'}))^{\alpha} \cdot \Gamma_{r':[w_{r'};s_{r'},e_{r'}]}} \tag{2-25}$$

其中

$$\Gamma_{a:[w_a;s_a,e_a]} = \sum_{\substack{a_p:[w_{a_p};s_{a_p},e_{a_p}]\\ e_{a_p}=s_a-1}} \sum_{\substack{a_n:[w_{a_n};s_{a_n},e_{a_n}]\\ s_{a_n}=e_a+1}} \left\{P(w_a \mid w_{a_p})P(w_{a_n} \mid w_a)\right\}^{\beta} \tag{2-26}$$

其中 $w_{a_p}$ 及 $w_{a_n}$ 分別代表 $w_a$ 前一個及後一個連接詞。

## 2.3 引用高階資訊(High Level Information)之信心度評估

除了上述所提到關於聲學、語言及語法方面的信心度評估外，[Cox and Dasmahapatra 2002]認為人們通常還可以藉由語意資訊(Semantic Information)來辨別辨識結果的正確性。因此，他們便提出利用潛藏語意分析(Latent Semantic Analysis, LSA)來判別辨識結果的正確與否，另外[Guo *et al*. 2004]也提出利用詞與詞之間交互訊息(Inter-word Mutual Information)來與事後機率做適當地結合，進而提昇信心度評估的正確率。以下兩小節將會分別介紹此兩種方法。

### 2.3.1 潛藏語意分析(Latent Semantic Analysis, LSA)

潛藏語意分析近年被廣泛運於資訊檢索(Information Retrieval)[Furnas *et al*. 1988]及語音辨識的語言模型[Belllegarda 1998]等領域。主要是利用線性代數的奇異值分解(Singular Value Decomposition, SVD)來將原本高維度且不相關的詞向量(Word Vector)與文件向量(Document Vector)投影到較低維度的潛藏語意空間(Latent Semantic Space)。如果兩個詞向量在此潛藏語意空間利用餘弦估測(Cosine Measure)的值愈大(在空間上愈接近)，則此兩個詞也有比較接近的語意。

在進行奇異值分解之前，我們必須先建立一個詞-文件矩陣(Word-document Matrix) $A$，此矩陣可經由事先收集大量文件資料求得。假設我們詞典有 $m$ 個詞，而文件有 $n$ 篇，則此矩陣 $A$ 的維度大小便是 $m \times n$，而每個元素的算法通常可表示為

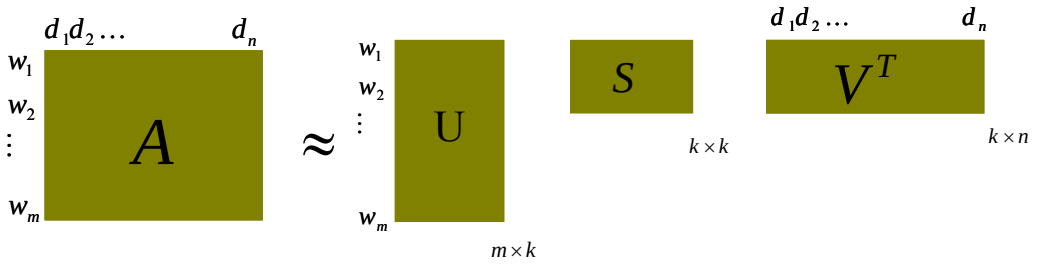


圖 2-7 奇異值分解

$$a_{ij} = (1 - E_i) \times \log(1 + \frac{c_{ij}}{n_j}) \tag{2-27}$$

$c_{ij}$ 是詞 $w_i$ 出現在第 $j$ 篇文件 $d_j$ 的次數;$n_j$ 則是代表 $d_j$ 的大小。而 $E_i$ 則可視為 $w_i$ 的正規化熵值(Normalized Entropy) :

$$E_i = -\frac{1}{\log_2(N)} \sum_{j=1}^{N} f_{ij} \log_2 f_{ij} \tag{2-28}$$

其中

$$f_{ij} = \frac{c_{ij}}{t_i} \tag{2-29}$$

$t_i$ 代表詞 $w_i$ 出現在 $n$ 篇文件的總次數。而 $0 \le E_i \le 1$(在有 $f_{ij} = 1$ 以及所有 $f_{ij} = \frac{1}{N}$ 時分別有最小和最大值),因此當 $E_i$ 的值愈大,則代表 $w_i$ 在各個文件中出現的次數趨近相同,其重要性便比較低;反之當 $E_i$ 的值愈小,$w_i$ 出現次數集中於某幾篇文件,其重要性因而提高。另外 $\frac{c_{ij}}{n_i}$ 可視為 $w_i$ 在 $d_j$ 的重要性,其值愈大,則代表其重要性愈高;反之,其重要性愈低。雖然有了此詞-文件矩陣 $A$,但由於每篇文件不可能包含所有的詞,通常此矩陣 $A$ 裡面的元素會有許多的值為0,且詞與文件的維度不同,其代表的意義也不同。因此,可以利用進行奇異值分解來做降維的動作。而奇異值分解的公式如下:

$$A \approx USV^T \tag{2-30}$$

其中$U$代表$m \times k$維的左奇異矩陣;$S$為$k \times k$維的對角矩陣;$V$是$n \times k$維的右奇異矩陣;$V^T$則是代表$V$的轉置矩陣(Transposition Matrix);$k$為小於等於矩陣$A$的秩$R$的一個整數值。奇異值分解的概念可以用圖 2-7來表示。經過奇異值分解後，詞和文件就都被投影到維度較低的潛藏語意空間，而原本在矩陣$A$的列向量便可用$U$的列向量$\vec{u}_i$來表示，而$A$的行向量可以改用$V^T$的行向量$\vec{v}_j^T$來代表。其中$\vec{u}_i$與$\vec{v}_j^T$的每一維度有一對一的對應關係，代表某一種潛藏的語意空間[Bellegarda 2000; 2005]。因此，如果我們想計算兩個詞$w_i$及$w_j$在語意上是否有相關聯，便可以藉著$\vec{u}_i$及$\vec{u}_j$的餘弦估測值來決定。

在實作上，當我們將一個聲學觀測序列辨識成一詞序列$W = w_1, w_2, \ldots, w_N$時，一個詞$w_i$的信心度評估可用下式計算:

$$MSS_i = \frac{1}{N} \sum_{j=1}^{N} \text{Cos}(U(w_i), U(w_j)) \tag{2-31}$$

$MSS_i$代表$w_i$的平均語意相似度(Mean Semantic Similarity)，而$U(w_i)$代表做完奇異值分解後的$\vec{u}_i$向量。Cos(·,·)則是兩個向量的餘弦估測函式。但由於功能詞(Function Word)對其它的詞在語意上幾乎都很接近，加上功能詞常常會一直出現在辨識詞序列中，使得$MSS_i$的值容易變大，而影響(2-31)式的正確性。為了避免這樣的情況，在求$w_i$的平均語意相似度時，通常都不考慮功能詞，而採用式(2-32):

$$MSS_i = \frac{1}{N - N_{w_j \in W^f}} \sum_{j=1, w_j \notin W^f}^{N} \text{Cos}(U(w_i), U(w_j)) \tag{2-32}$$

其中$W^f$代表所有的功能詞集合，$N_{w_j \in W^f}$則代表詞序列中功能詞的個數，除了$MSS_i$之外，更進一步關於利用潛藏語意分析計算某詞$w_i$的信心度，請參考[Cox and Dasmahapatra 2002]。

2.3.2 交互資訊(Mutual Information, MI)

交互資訊可以視為是兩個變數(Variables)相依(Dependence)程度。而當給定兩個詞 $w_i$ 及 $w_j$ 時，其交互訊息的計算如式(2-33)所示:

$$MI(w_i, w_j) = \log\left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}\right) \tag{2-33}$$

其中

$$P(w_i) = \sum_{w_j} P(w_i, w_j) \tag{2-34}$$

而

$$P(w_i, w_j) = \frac{N(w_i, w_j)}{\sum_{w_l, w_k} N(w_l, w_k)} \tag{2-35}$$

其中 $N(w_i, w_j)$ 代表 $w_i$ 及 $w_j$ 同時在訓練資料出現的次數，而式(2-35)的分母項則是代表語料庫中所有詞對(Word Pair)個數。因此，一個辨識詞序列 $W = w_1, w_2, \ldots, w_N$ 中某個詞 $w_i$ 的信心度可以表示成 $w_i$ 與辨識詞序列其它詞的平均交互資訊(Average Mutual Information, AMI) :

$$AMI_i = \frac{1}{N}\sum_{j=1}^{N} MI(w_i, w_j) \tag{2-36}$$

雖然單獨使用上述的兩種較高階資訊的信心度評估沒有辦法比聲學方面的信心度評估獲得較好的效用[Jiang 2005]。但如果將此兩種高階資訊與事後機率相關的信心度評估做適當的結合，如線性插補法(Linear Interpolation)，可以獲得單獨只用事後機率相關的信心度評估更佳的效果[Guo *et al*. 2004]。

目前信心度評估除了用於驗證語音辨識結果的可信度之外，在進行詞彙樹複製搜尋的往前觀測(Look-ahead)，或是降低詞錯誤率(Word Error Rate, WER)也有相關的研究，在稍後的2.4及2.5兩小節將會一一說明。

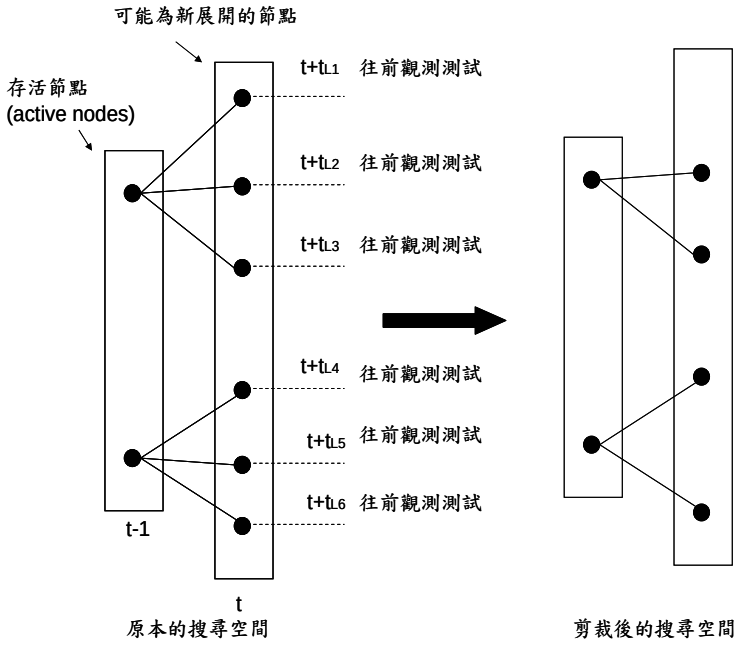


圖 2-8 往前觀測基本概念

## 2.4 信心度評估於詞彙樹複製搜尋之研究

在大詞彙連續語音辨識的研究領域中，要如何降低詞彙樹複製搜尋的搜尋空間(Search Space)及運算時間(Computation Time)是極為重要的課題。而往前觀測(Look-ahead)便是希望能降低運算時間卻又不影響語音辨識系統的正確率，其基本概念為在進行詞彙樹複製搜尋時，往前多看幾個音框或是事先看完整個語句，只有通過往前觀測測試的節點才會存活下來，如圖 2-8所示。在[Afify *et al*. 2005]中，作者提出了以驗證為基礎(Verification-based)的往前觀測法。作者將往前觀測當作是一個假設檢定(Hypothesis Testing)問題，針對某個音素 $\alpha$ 提出虛無假設(Null Hypothesis) $H_0$ 及對立假設(Alternative Hypothesis) $H_1$:

$$H_0 \ : \ \alpha \ \text{ starts at time } \ t$$
$$H_1 \ : \ \alpha \ \text{ does not start at time } \ t$$

接下來便是採用兩元假設檢定(Binray Hypothesis Testing)，以決定要接受兩個假設中那一個。其檢定式如式(2-37)所示:

$$LRT = \frac{p(X \mid H_0)}{p(X \mid H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \tag{2-37}$$

其中$\tau$代表我們預先設定的門檻值(Threshold)，如果$LRT$的值大於$\tau$，則採用虛無假設，也就是將音素$\alpha$留下來。反之，則裁剪音素$\alpha$。實作上，$p(X \mid H_0)$及$p(X \mid H_1)$的機率估測可以分別寫成$p(x_t^{t+d_\alpha} \mid \alpha)$及$p(x_t^{t+d_\alpha} \mid \bar{\alpha})$，其中$d_\alpha$代表要往前看多少時間(在這邊設定虛無假設及對立假設往前看的時間是一樣的)，而$\alpha$及$\bar{\alpha}$分別代表音素對應的隱藏式馬可夫模型及反對應模型(Anti-model)，更詳細的資訊請參照[Afify *et al*. 2005]，另外在[Fabian *et al*. 2005]及[Abdou and Scordilis 2003]也都有探討信心度評估運用於往前觀測的相關探討。

## 2.5 信心度評估於降低詞錯誤率之研究

語音辨識最終的夢想便是希望能讓電腦可以像人一樣，在辨識人們所說的每一句話時，詞或字正確率能達到百分之百。因此，不論是研究聲學、語言模型或是強健性語音辨識等方面的研究學者，其實都是為了達到這個目標而努力。近年來，開始有研究學者也試著將信心度評估應用於詞圖搜尋或是*N*-最佳詞序列重新排序，以增進語音辨識的正確率。接下來的幾個小節將分別介紹搜尋最佳詞序列的基礎觀念，以及探討應用信心度評估於提昇語音辨識系統的正確率之相關研究。

### 2.5.1 最小化貝氏風險(Minimum Bayes Risk)

當我們將一段聲學觀測序列$X$辨識成某個詞序列$W$時，有時難免會產生辨識錯誤的情況(如2.1小節所提到會有詞的刪除、插入及替代等錯誤)。而一個好的語音辨識系統對$X$的辨識錯誤率當然是愈小愈好。更進一步而言，我們可以將辨識系統視為一個將$X$對應到詞序列的一個映射函式(Mapping Function)，或稱為分類器(Classifier):

$$F(X): X \rightarrow W_h^X \tag{2-38}$$

其中$W_h^X$代表一個假設空間(Hypothesis Space)，為詞典中所有詞的可能組合$\overline{\mathbf{W}}_\Sigma$的子集合(Subset)。另外，我們先定義一個成本函式(Cost Function)$\ell(W,W')$($W$屬於$\overline{\mathbf{W}}_\Sigma$，而$W'$屬於$W_h^X$)，代表當將一個詞序列$W$的聲學觀測序列辨識成$W'$時的成本(為一個實數值)。此成本函式的定義通常與任務相關(Task-dependent)，如對語音辨識的領域來說，此函式通常定義成Levenshtein距離(此為語音辨識系統中判別辨識率好壞的評估標準，將會於3.2.3小節介紹)，或是與Levenshtein距離有關的函式。在估計此映射函式的期望錯誤時，通常都是利用貝氏風險來估算:

$$\sum_{W\in\overline{\mathbf{W}}_\Sigma} \ell(W,W')P(W\mid X) \tag{2-39}$$

有了風險評估函式之後，此映射函式$F(X)$對$X$的最佳辨識結果便是相當於在$W_h^X$此假設空間中選擇一條貝氏風險為最低的詞序列:

$$F^*(X)=\arg\min_{W'\in W_h^X}\sum_{W\in\overline{\mathbf{W}}_\Sigma} \ell(W,W')P(W\mid X) \tag{2-40}$$

2.5.2 最大事後機率與最小化貝氏風險尋找最佳詞序列之關聯

如果將$\ell(W,W')$定義成一個簡單的0/1對稱(Zero-one and Symmetric)成本函式:

$$\ell_{0/1}(W,W')=\begin{cases}1 & \text{if } W'\neq W\\ 0 & \text{otherwise}\end{cases} \tag{2-41}$$

則式(2-40)可以改寫成

$$\begin{aligned}F(X)&=\arg\min_{W'\in W_h^X}\sum_{W\in\overline{\mathbf{W}}_\Sigma}\ell_{0/1}(W,W')P(W\mid X)\\&=\arg\min_{W'\in W_h^X}\sum_{W\in\overline{\mathbf{W}}_\Sigma,W\neq W'}\ell_{0/1}(W,W')P(W\mid X)\\&=\arg\min_{W'\in W_h^X}1-P(W'\mid X)\\&=\arg\max_{W'\in W_h^X}P(W'\mid X)\end{aligned} \tag{2-42}$$

也就變成我們目前一般使用的最大事後機率尋找最佳詞序列的方法。

2.5.3 事後機率詞圖搜尋

在[Wessel *et al*. 2000]中，作者提出了使用詞圖計算出來的事後機率來增加辨識結果的正確率。假設我們有了詞圖 $\Psi^X$ 之後，在實作最大事後機率辨識法則時，會以 $\Psi^X$ 取代假設空間 $W_h^X$ 。另外，在產生詞圖之後，我們便有了每個詞段的詞編號及對應的開始和結束時間。因此，(2-42)式便可以改寫成式(2-43):

$$\begin{aligned} F(X) &= \arg \max_{[w^n;s^n,e^n]_{n=1}^N \in \Psi^X} P([w^n;s^n,e^n]_{n=1}^N \mid \Psi^X) \\ &= \arg \max_{[w^n;s^n,e^n]_{n=1}^N \in \Psi^X} \prod_{n=1}^M P([w^n;s^n,e^n] \mid [w^n;s^n,e^n]_{n=1}^{n-1}, \Psi^X) \qquad (2\text{-}43) \\ &\approx \arg \max_{[w^n;s^n,e^n]_{n=1}^N \in \Psi^X} \prod_{n=1}^M P([w^n;s^n,e^n] \mid \Psi^X) \end{aligned}$$

其中 $P([w^n;s^n,e^n] \mid \Psi^X)$ 也就是2.2.2小節中所介紹的詞段之事後機率。

2.5.4 最小化音框錯誤率詞圖搜尋

當我們將貝氏風險中的成本函式訂為0/1函式後，便可以將最大事後機率的辨識方法視為找出一條正確率為最高的詞序列(因為詞序列的事後機率也可想像為詞序列正確的機率)。但是由於一般是以詞錯誤率(Word Error Rate)為語音辨識系統好壞的評估標準，因此會有成本函式與估評標準不匹配的情況出現，在[Mangu *et al*. 2000]中便有提出最小化詞序列錯誤率並不等於同時最小化詞錯誤率的例子。要解決這個不匹配的問題，最簡單的方法便是將成本函式改為與評估標準一致。雖然這樣本質上解決了不匹配的問題，但是由於計算詞錯誤率時，必須計算兩個詞序列的Levenshtein距離(也就是統計兩個詞序列中的取代、刪除及插入的次數)。如果要在詞圖中眾多可能的詞序列中，針對任兩個詞序列都要計算Levenshtein距離的話，其組合有太多種，計算複雜度便成了這個方法的主要缺點。為了解決此計算複雜度所造成的缺點，[Stolcke *et al*. 1997]中將此兩兩比對(Pairwise Alignment)限制在*N*-最佳詞序列範圍內。幾年後，[Mangu *et al*. 2000]將原本比對的對象再度從*N*-最佳詞序列改成詞圖。但是考量前面所提到的高計算複雜度缺點，[Mangu *et al*. 2000]試著將詞圖原本包含

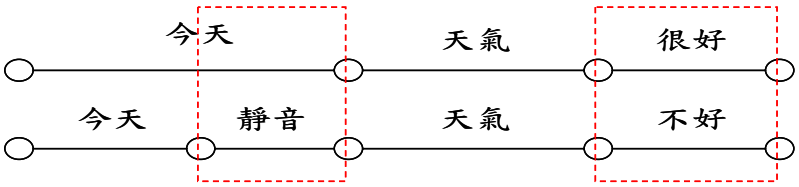


圖 2-9 音框錯誤率圖解

不同長度的詞序列改成長度一模一樣，一般稱為一致性網路(Consensus Network or Sausage)。因此只要執行複雜度較低的多重字串比對(Multiple String Alignment)。

字串比對中的刪除及插入錯誤是造成兩兩比對的複雜度會如此高的原因。當字串比對只剩下取代錯誤時，原本動態規畫(Dynamic Programming)比對法的Levenshtein距離成本函式之計算複雜度將大幅下降。[Wessel *et al.* 2001]便是以此想法為出發點，提出了新的成本函式-音框錯誤率(Time Frame Error)。

由於詞圖擁有每個詞的開始及結束時間資訊，當我們比對詞圖上任意的兩條詞序列時，如圖 2-9所示，虛線框框的地方便是代表有音框錯誤。從圖 2-9可以看出，原本字串比對時會存在的詞刪除、插入或替代錯誤，現在都已可以用音框錯誤來取代。基於這個觀念，便可以定義一個新的不對稱成本函式，稱為音框錯誤率[Wessel *et al.* 2001] :

$$\ell([w^n;s^n,e^n]_{n=1}^{N},[w^m;s^m,e^m]_{m=1}^{M})=\sum_{n=1}^{N}\frac{\sum_{t=s^n}^{e^n}1-\delta(w^n,w_t^m)}{1+\alpha(e^n-s^n)} \tag{2-44}$$

其中$[w^n;s^n,e^n]_{n=1}^{N}$及$[w^m;s^m,e^m]_{m=1}^{M}$分別代表詞圖上的兩條完整路徑，其詞的個數分別為$N$及$M$，而

$$\delta(w^n,w_t^m)=\begin{cases}1, & \text{if } w^n=w_t^m\\ 0, & \text{if } w^n\neq w_t^m\end{cases} \tag{2-45}$$

$w_t^m$代表在$t$這個音框時$w^m$的詞編號。$\alpha$則是決定是否要做正規化(Normalization)的參數。當$\alpha$為1時，代表採取正規化的動作，換言之此辨識方法有長詞優先的傾向。現在，將音框錯誤率成本函式代入最小化貝氏風險的準則，則式(2-40)可以改寫為

$$
\begin{aligned}
F^{*}(X) &= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{[w^m;s^m,e^m]_{m=1}^M} \ell([w^n;s^n,e^n]_{n=1}^N,[v^m;s^m,e^m]_{m=1}^M) P([v^m;s^m,e^m]_{m=1}^M \mid \Psi^X) \right\} \\
&= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{[w^m;s^m,e^m]_{m=1}^M} \sum_{n=1}^N \frac{1-\delta(w^n,w_t^m)}{1+\alpha(e^n-s^n)} P([v^m;s^m,e^m]_{m=1}^M \mid \Psi^X) \right\} \\
&= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{n=1}^N \frac{\sum_{t=s_n}^{t=e_n} \left[ 1 - \sum_{[w^m;s^m,e^m]_{m=1}^M} \delta(w^n,w_t^m) P([v^m;s^m,e^m]_{m=1}^M \mid \Psi^X \right] )}{1+\alpha(e^n-s^n)} \right\} \\
&= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{n=1}^N \frac{\sum_{t=s_n}^{t=e_n} \left[ 1 - \sum_{[w^m;s^m,s^m]_{m=1}^M} \sum_{w^m:s^m\le t\le e^m} \delta(w^n,w_t^m) P([v^m;s^m,e^m]_{m=1}^M \mid \Psi^X) \right]}{1+\alpha(e^n-s^n)} \right\} \\
&= \arg\min_{[w^n;s^n,e^n]_{n=1}^N} \left\{ \sum_{n=1}^N \frac{\sum_{t=s_n}^{t=e_n} \left[ 1 - \sum_{[w^m;s^m,s^m]:s^m\le t\le e^m} \delta(w^n,w_t^m) P([v^m;s^m,e^m] \mid \Psi^X) \right]}{1+\alpha(e^n-s^n)} \right\}
\end{aligned}
\tag{2-46}
$$

其中我們將 $\sum_{[w^m;s^m,e^m]:s^m\le t\le e^m} \delta(w^n,w_t^m) P([w^m;s^m,e^m] \mid \Psi^X)$ 表示成 $P(w^n \mid t,\Psi^X)$，可以視為在音框 $t$ 時，通過 $w^n$ 的機率為多少，也就相當於 $w^n$ 在音框 $t$ 時正確的機率。