

第二章 基礎理論與技術

2.1 語音辨識概論

目前較常見的語音辨識技術大致有：動態時間扭曲法 (DTW, Dynamic Time Warping)、隱藏式馬可夫模型 (HMM, Hidden Markov Model)。

2.1.1 動態時間扭曲法 (DTW, Dynamic Time Warping)

DTW 是語音辨識中最常用到也是最簡單的做法。其原理乃將兩個聲音分成測試語音與參考語音，將參考語音特徵置於縱軸、測試語音特徵至於橫軸，再對兩個語音訊號進行比對，而找出最短的路徑(即為答案)，屬於 Speaker Dependent 之方法。

優點：由於方法簡單，所以運算量相對於其他方法來的小很多，辨識速度快，適合 Speaker Dependent 語者辨識方面的應用。

缺點：由於每個人說話會有快慢不一的問題，所以比對時，經常要更動時軸；要針對每個人每一個語句都要建立一句參考語句，會造成比對語料庫過大，且不易建立多人共用之語料庫。

2.2.2 隱藏式馬可夫模型 (HMM, Hidden Markov Model)

HMM 是採用機率模型來描述發音的現象，將基本音節或音素建立語音模型，然後以一序列的語音模型來代表一句話，再依照機率模型來做狀態的轉移，藉以判斷辨認的結果。

優點：具有低記憶體、實作時有彈性和辨識時與語者無關(Speaker Independent)的特點，非常適合我們將語音辨識運用在嵌入式系統上。

缺點：要蒐集大量的語料來建立聲學模型，要花費一些時間訓練及建立模型，相對於 DTW 來的複雜許多，所以辨識時間較 DTW 所花的時間較長一些。

2.2 語音特徵參數擷取

在進行語音辨識之前，我們會先將輸入的語音訊號作特徵的擷取，然後再予以進行辨識。然而人類的語音訊號是屬於時變訊號(Time-Varying Signal)，即語音的特性為一非線性而隨著時間改變的訊號，所以無法以一般線性時變的方式來對訊號作分析。

因此我們通常是將語音訊號切割成為多個小塊的連續訊號集合，以一個音框為單位來對語音訊號擷取其特徵，一般來說，音框的長度約略取在 20~30 ms 間，因為語音訊號在這個範圍之間具有半穩態的特性，這就是短時距處理(Short-time Processing)的概念。

常見的語音特徵參數擷取方法有 LPC (Linear Predictive Coding)、MFCC (Mel-scale Frequency Cepstrum Coefficients)，在這邊我們採用 MFCC，雖然 LPC 有著運算量低的優點，但是缺點即在於為考量語音在頻譜上的特性，雜音越大取出來的語音特徵於辨識時就更不準確，而 MFCC 因為考慮到人類的聽覺系統對於低頻的聲音感知能力較強，所以 MFCC 在求取參數時會將低頻的部分多取，而高頻的部分少取。

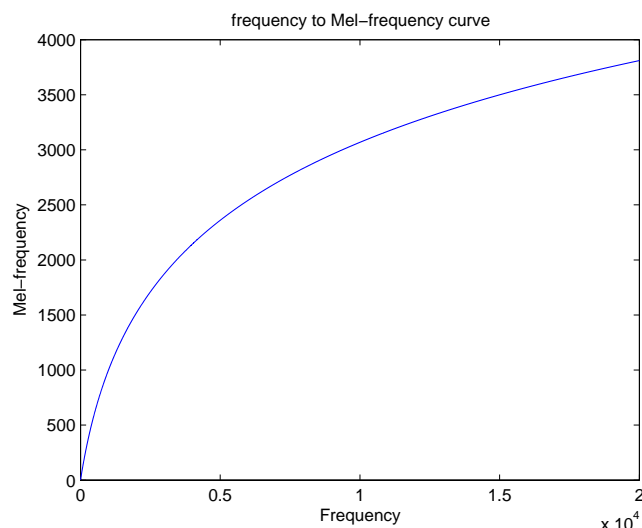


圖 2-1 語音頻率與人類感知頻率關係圖

整個 MFCC 特徵擷取流程如圖 2-2 所示，依序分為：音框化 (Frame Blocking)、計算能量 (Energy)、預強調 (Pre-emphasis)、漢明窗 (Hamming Window)、快速傅立葉轉換 (FFT, Fast Fourier Transformation)、三角帶通濾波器 (Triangular Band-pass Filters)、離散餘弦轉換 (DCT, Discrete Cosine Transformation)、差量倒頻譜參數 (Delta Cepstrum) 等步驟。

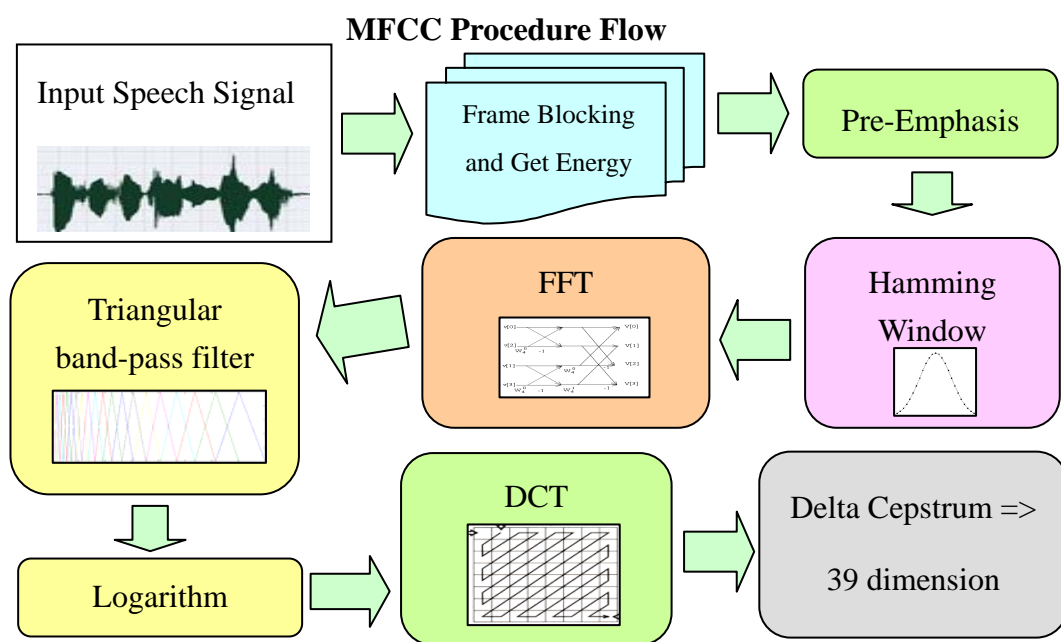


圖 2-2 MFCC 語音特徵擷取流程圖

2.2.1 音框化 (Frame Blocking)

先前提過，人類的語音訊號約在 20ms~30ms 會呈現半穩態，所以我們在觀測語音訊號的特徵會將 N 個取樣點集成一個觀測的單位---音框，並為了使音框與音框間的變化不要太大，通常我們會依一定的比例來重疊音框，例如 $1/3$ 、 $1/2$ 等等，在這邊我們取音框的大小 $N = 320$ 點，音框重疊的部分為 160 點，音框長度約 20ms。

2.2.2 計算能量 (Energy)

音框化過後，計算每一個音框的能量 ($energy = \sum_{n=1}^{frameSize} S(n)^2$) 值，作為之後 MFCC 第十三個參數。



2.2.3 預強調 (Pre-Emphasis)

將語音訊號通過一個高通濾波器： $H(z) = 1 - a \times z^{-1}$ ，其中 $a \in 0.9 \sim 1.0$ ，而假設我們的輸入訊號為 $S(n)$ ，預強調後的訊號為 $S_2(n) = S(n) - a \times S(n-1)$ 。而此一高通濾波器的目的即是為了消除並補償發聲過程中聲帶與嘴唇所產生壓抑高頻的效應，意即用來突顯高頻的共振峰，在這邊我們取 $a = 0.975$ 。

2.2.4 漢明窗 (Hamming Window)

為了使得音框與音框間的左右連續性增加，我們會將音框的訊號再乘上一個 Hamming Window，如果輸入的訊號一定保證是週期訊號，就沒必要再乘上

Hamming Window，只是人類的聲音訊號為時變訊號，所以如果不乘上 Hamming Window 會導致之後 FFT 分析訊號的時候會產生很多不必要的訊號，造成分析上的誤差及錯誤產生。假定音框的訊號為 $S(n), n = 0, \dots, N-1$ ，乘上 Hamming Window 過後我們會得到 $S'(n) = S(n) \times W(n)$ ， $W(n)$ 的公式如下：

$$W(n, a) = (1 - a) - a \times \cos(2\pi n / (N - 1)), 0 \leq n \leq N - 1 \quad (2.2.-1)$$

在這邊我們一般都是取 $a = 0.46$ 。圖 2-3 為 16khz，長度 5 sec 的語音訊號，其中一個音框（約 20ms）之原始訊號與乘上 Hamming Window 後所產生新的訊號的比較。

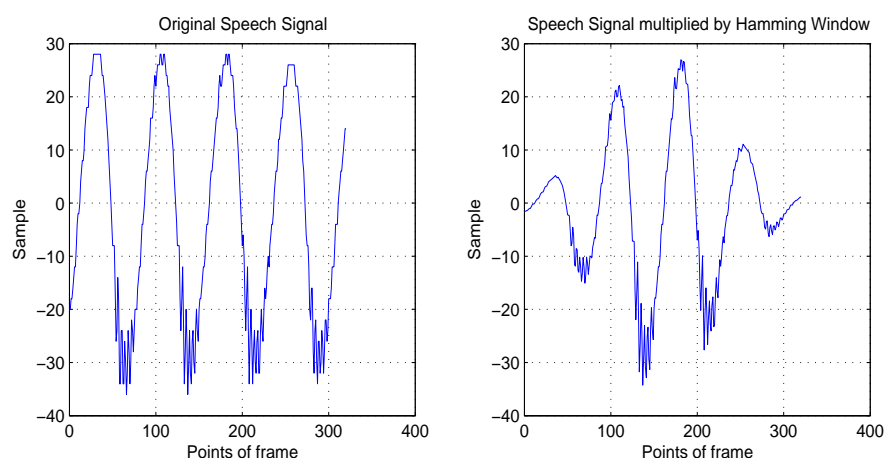


圖 2-3 原始訊號與乘上 Hamming Window 過後的訊號比較

2.2.5 快速傅立葉轉換 (FFT, Fast Fourier Transform)

在語音訊號處理中，最常被使用到的技術就是將時(間)域 (Time Domain) 的訊號轉換至頻率域 (Frequency Domain) 上來觀察能量分佈的情形，從能量的頻譜分佈，我們可以找到許多參數，讓我們來區隔不同的發音現象。

2.2.6 三角帶通濾波器 (Triangular band-pass filter)

FFT 轉換過後，我們可以得到能量的頻譜，然後將能量的頻譜乘上一組 M 個的三角帶通濾波器以降低資料的維度，最後再將每一個頻帶計算其對數量值得到對數能量，在這邊我們依據[8]取 26 個三角帶通濾波器。

2.2.7 離散餘弦轉換 (DCT, Discrete Cosine Transformation)

最後再將求出來的 26 個能量對數 E_k 代入離散餘弦轉換，求出 Mel-scale 倒頻譜參數，將之前轉到頻率域的訊號再轉回時域，DCT 的公式如下：

$$C_m = \sum_{k=1}^M E_k \cdot \cos\left[m\left(k - \frac{1}{2}\right)\frac{\pi}{M}\right], m = 1, \dots, L \quad (2.2-2)$$

在這邊 M 為 26 個 Triangular Band-pass Filters， L 為倒頻譜參數之個數 12， E_k 則為 FFT 運算過後的能量值。

2.2.8 差量倒頻譜參數 (Delta Cepstrum Coefficients)

在實際的運用上，我們還會再加上差量倒頻譜參數，用來顯示其對時間的變化。Delta Cepstrum 的意義即在於倒頻譜參數對於時間的斜率，意即代表倒頻譜參數在時間上的動態變化，公式如下：

$$\Delta C_m(t) = \frac{\sum_{\tau=-M}^M \tau \cdot C_m(t+\tau)}{\sum_{\tau=-M}^M \tau^2} = \frac{\sum_{\tau=1}^M \tau \cdot (C_m(t+\tau) - C_m(t-\tau))}{2 \cdot \sum_{\tau=1}^M \tau^2}, m = 1, 2, \dots, L$$

... (2.2-3)

在這邊 M 取 2， τ 代表第 N 個音框，因此到最後我們還會再產生出 24 個差量倒頻譜參數，總共會得到 36 個參數，再加上能量的參數共 39 個，也就是我們一般語音辨識中最常用的 39 維參數。

2.3 聲學模型訓練

訓練聲學模型的部分，我們採用以 HMM 為理論基礎的 HTK (HMM Toolkits) [9] 來訓練我們的聲學模型。隱藏式馬可夫模型可分為連續與離散型 (CHMM and DHMM)，為了得到較好的辨識率我們採用 CHMM [10]。

CHMM 使用連續的機率密度函數 Gaussian Mixture Model 來計算狀態機率，HMM 本身即是一個有限狀態機包含了次轉移機率 (Next-transition Probability) 與自身轉移機率 (Self-transition Probability) 統稱為轉移機率 (Transition Probability) 如圖 2-4。

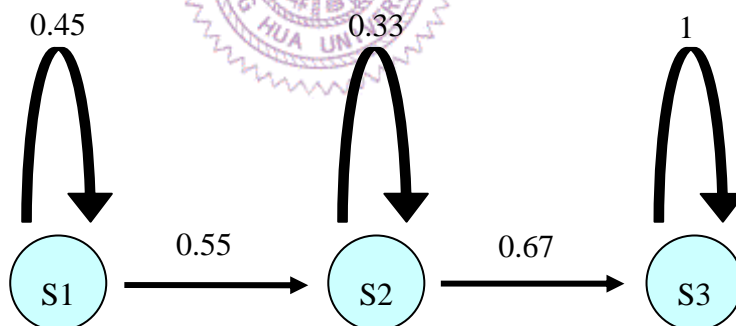


圖 2-4 狀態轉移示意圖

其機率密度函數為：

$$b_j(x) = \sum_{k=1}^M c_{jk} N(x, \mu_{jk}, \Sigma_{jk}) = \sum_{k=1}^M c_{jk} b_{jk}(x) \dots (2.3-1)$$

其中 $N(x, \mu_{jk}, \Sigma_{jk})$ 或是 $b_{jk}(x)$ 代表狀態 j ，包含一個共變數矩陣 Σ_{jk}

(Covariance Matrix)及平均向量 μ_{jk} (Mean Vector)的高斯密度函數， M 代表 Mixture 的數目， C_{jk} 代表第 k 個 Mixture 的權重。

2.4 語音訊號與聲學模型之比對

我們使用 Viterbi Algorithm [10]來比對受測語句與聲學模型，因為 Viterbi 演算法已經是很成熟的動態規劃演算法，可以大幅度降低搜尋 HMM 最佳狀態序列的計算量。最佳路徑機率定義如下：

$$V_t(i) = P(X_1^t, S_1^{t-1}, s_t = i | \phi) \dots (2.4-1)$$

$V_t(i)$ 代表在時間 t 時最有可能的狀態序列機率。而整個 Viterbi 步驟如圖 2-5

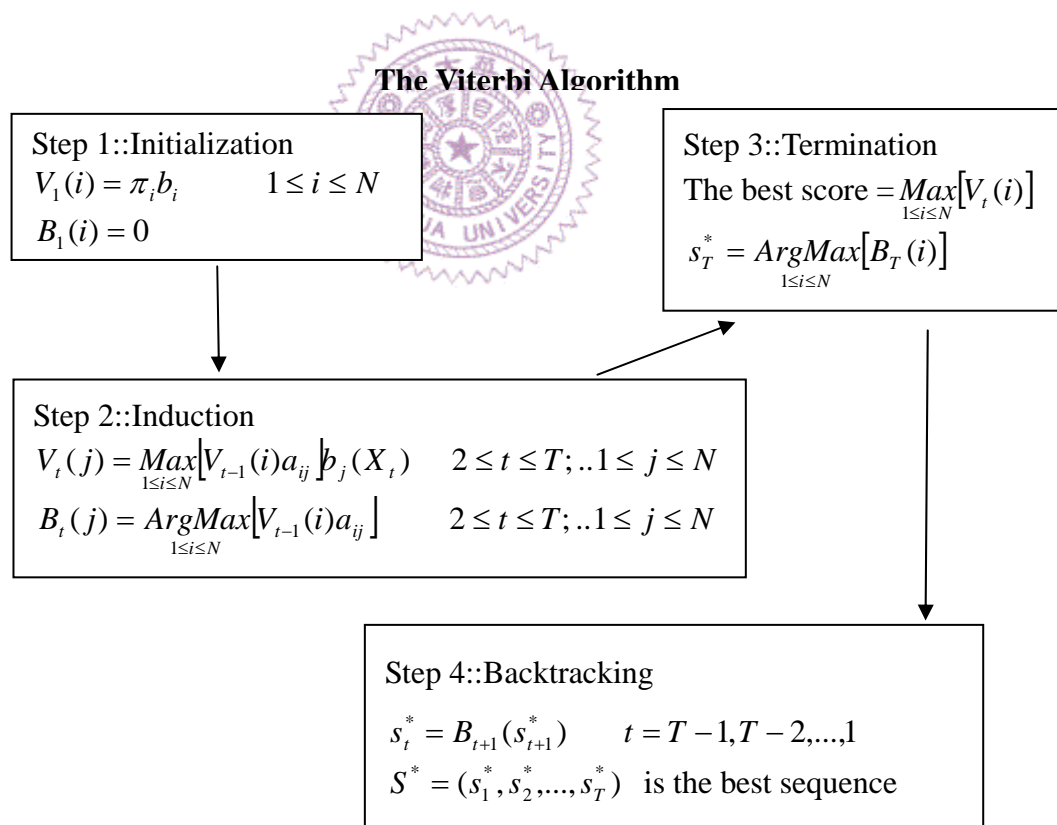


圖 2-5 Viterbi 計算流程圖