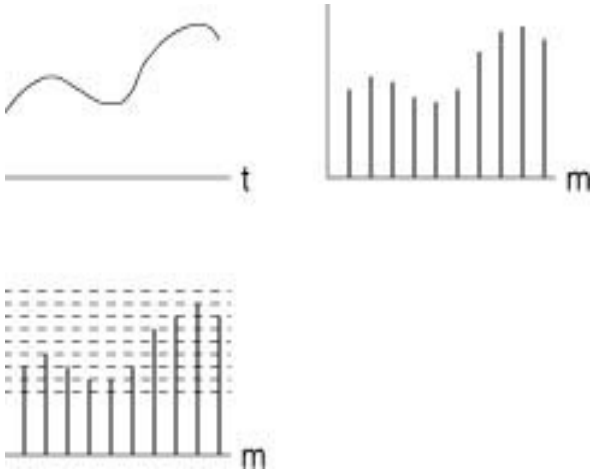
MP3 如何壓縮聲訊

杭學鳴　國立台北科技大學資訊工程系教授

MP3 數位隨身聽 — 現在的年輕人幾乎是人手一支。即便你不是特地買支 MP3 播放器隨身攜帶，你的手機通常也可播放 MP3 及其他多種格式的壓縮聲訊。

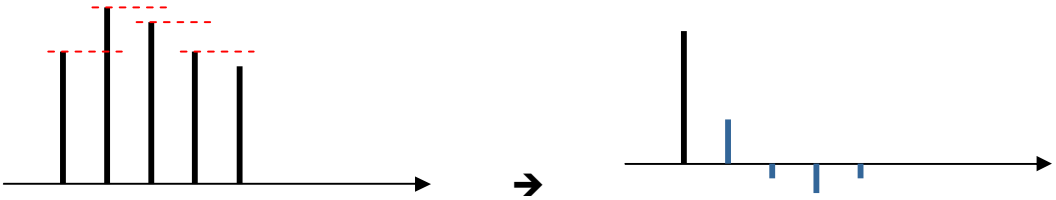
「輕便」，應該是造成 MP3 流行的主因，它雖然輕薄短小，可置於口袋，但它卻可連續播放數小時，甚至數百小時的音樂。MP3 播放器之所以能做的如此「小巧」，除了電子元件製造的進步外，還有一項技術的突破 —聲訊壓縮。(類似的技術亦可用於影像視訊，產生 DVD 影音光碟）。以下用較淺近的方式簡要說明「聲訊壓縮」的原理。

從物理的觀點，聲音訊號是空氣的振動。為了將聲音訊號儲存下來，我們用麥克風，將空氣振動波轉變成電子波動訊號。為了便於後續的數位訊號處理，我們先將連續波形類比式電子訊號取樣，取樣的示意圖為**圖一**。取樣訊號加以量化後，每一取樣點就可以整數來表示，因而可用二元(Binary)表示法。依據取樣定理，取樣頻率應為想要處理訊號頻寬的兩倍以上。由於人的耳朵能聽到的聲音頻率在 20KHz 以下，所以音訊一般**取樣頻率**是 40KHz 或 44.1KHz。44.1KHz 取樣即為目前 CD 音樂光碟的規範。至於為何要取 44.1 這個奇特的小數而非整數**?** 這是因為要配合電視訊號的規格，其細節不在此細說。



圖一 取樣: (左上)類比訊號在任何時間點(橫軸)都有數值，有無窮多數據點；(右上）取樣後訊號只有在取樣的時間點有數值，變成有限點數據；(左下）取樣訊號加以量化，每個取樣點用最接近的量化階梯值(橫虛線)代表，因為量化階梯值個數不多(數百到數萬)，因此可用整數去標示。

壓縮聲音訊號是利用兩種原理，第一類是統計（Statistics）的方法，第二類是人類聽覺特性（Humam perception）。第一類統計方法的一個簡單作法是預測(Prediction)。一般而言，前後兩個取樣的大小是相當接近的。如果我們用前一個訊號預測下一個訊號。或者說我們只傳送前後兩個訊號間的**差值**，如**圖二**，則一般而言，差值的大小(量)比原始訊號小。也就是說，相較於傳送原始訊號，傳送差值訊號所消耗的能量較少。



圖二（左）原始訊號，(右）除第一點為原始訊號外，其餘為前後兩個原始訊號之差值訊號。將次一點差值訊號加到其前一點原始訊號即還原回次一點的原始訊號。

運用統計方法壓縮也有較複雜的方式。假如我們用 16 位元（bit）代表一個取樣值，這些數值(從 0 到 2^{16})的機率是不大均勻的。特別是我們取差值後，差值多半在 0 附近，但偶而也有較大幅度的。如果每一個差值，我們用同樣多位元(例如 16 bits)去表示(傳送)，我們沒節省到位元的使用。但是我們如果用較少的位元表示較常出現的數值，用較多位元表示不常見的數值，那麼平均而言，使用的位元數仍較少，這是所謂**可變長度編碼**（Variable-length coding, VLC)。**表 1** 為一個簡化的例子。假設有四個事件(A,B,C,D)，個別機率為 0.64, 0.16, 0.16, 0.04。如果每個事件，我們都用兩個位元（2 bits）表示，平均傳送每個事件的位元數為 2。但若用 VLC 方式，如第四欄，則平均位元率為 1.56。這種作法又稱為**熵編碼**(Entropy coding)。

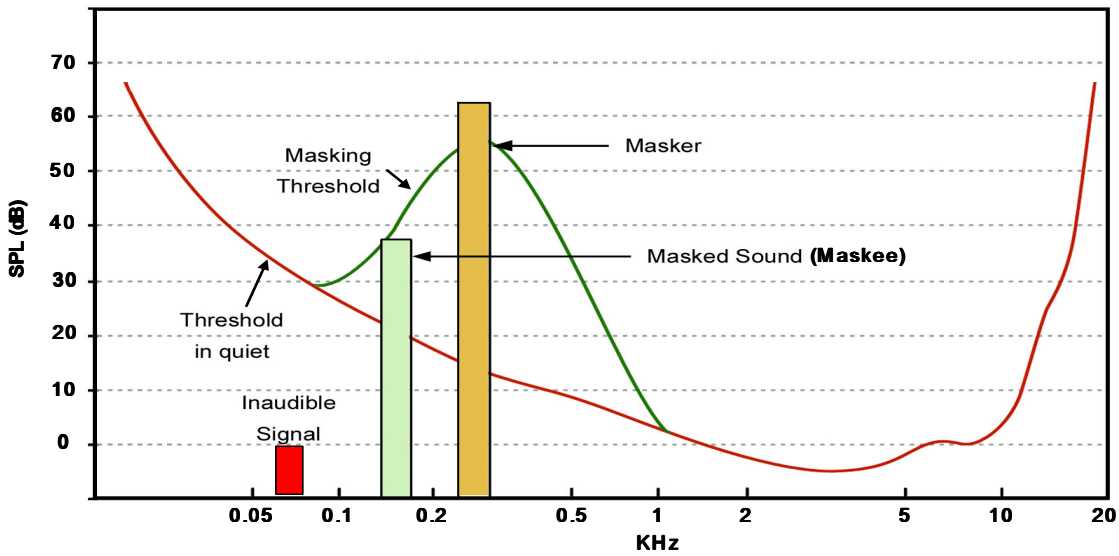
表 1 可變長度編碼示例

事件	機率	固定長度編碼	VLC
A	0.64	00	1
B	0.16	01	01
C	0.16	10	001
D	0.04	11	000

VLC 平均位元率： $\bar{b} = (1\times.64 + 2\times.16 + 3\times.16 + 3\times.04) = 1.56$

壓縮聲訊的第二項利器是人類聽覺。人的耳朵因其生理結構對聲音的收聽有數個特性: 第一個是我們耳內聲音接受器的靈敏度。為方便說明，在此先介紹聲音頻域的表示法。簡單的說，頻域表示法就是將音波分解成為不同頻率弦波的組合; 具體的說，是將聲音分解成許多不同頻率的高低音的組合。從物理觀點，高音是高頻率的音波，低音是低頻率的音波。樂器所發出的聲音，不是單一頻率的弦波。在彈奏的主要音符(對應一主要頻率)外，亦含有諧振弦波(主要頻率整數倍)，以及較小振幅的臨近頻率弦波。

回到耳朵靈敏度的問題，人類的耳朵大約只能聽到 20KHz 頻率的高音，這在各種動物當中可算是相當傑出的了。其次，人類耳朵的靈敏度隨頻率而不同，在 3KHz 左右最靈敏。**圖三**的橫軸為頻率（或可說是音高），縱軸為聲壓（Sound Pressure Level, SPL）或可理解為聲音的大小。圖中有一條較靠近橫軸底線的曲線，代表一般人聽力的靈敏度，稱為**靜音門檻**(Threshold in Quiet)。聲壓必須超過這條曲線的幅度，人耳才可以感受到。由於縱軸是採用對數值，每上升一格即為增加 10 倍，因此在 10KHz 以上*的*頻率，聲壓門檻值其實甚高。人類耳朵對高頻不太敏感，如果原來聲音的高頻部份不夠強，人耳是不太聽得到。人耳對極低頻亦是如此。既然人耳不易分辨，在進行音訊壓縮時，對極低頻與較高頻區域，可以做的粗糙些，也就是少送一些位元。

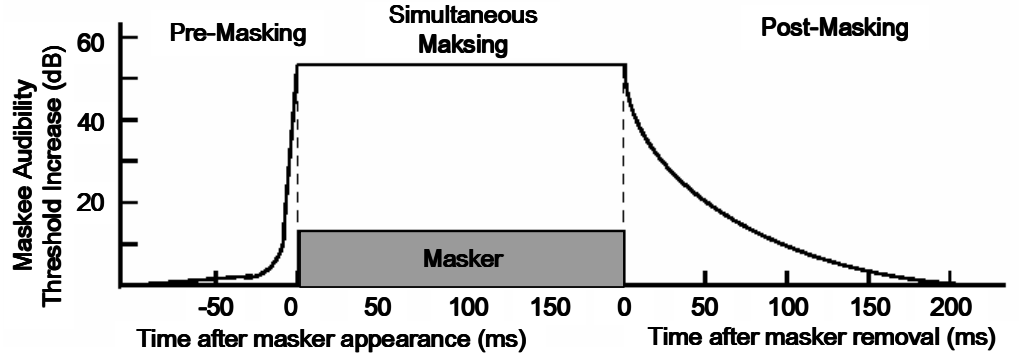


圖三 同時遮蔽效應: 圖中最底下中凹的曲線為靜音門檻，3KHz 左右為最低點。低於此靜音門檻的聲音人耳察覺不到，例如左下的 Inaudible Signal。如有一極強的聲音如中央突出的 Masker，它可遮蔽掉附近較弱的聲音，如其左之 Masked Sound (Maskee)。這個「遮蔽」是主觀的，亦即物理上實際存在這個 Maskee，但

人耳聽不出其存在。Masker 的遮蔽範圍(Masking Threshold)如圖中在 Masker 兩側類似金字塔之三角曲線。

人耳聽覺的第二個特性是**遮蔽效應**（Masking effect），而**圖三**主要表達的正是這個部份。遮蔽效應的概念是: 如果同時有兩個單頻弦波，其頻率相當接近，能量(振幅)大者可能在主觀上遮蔽能量小者。所謂主觀，意指物理上有兩個弦波，但人耳只能聽到聲量較強者。**遮蔽者**(Masker)的強度必須超過**被遮蔽者**(Maskee)某個幅度（如**圖三**金字塔三角形所顯示），這幅度的大小與兩者的頻率差距有關，也與其本身頻率大小有關。這個遮蔽效應亦可擴展到非弦波，只是其狀況較為複雜。

上述的遮蔽效應現象，是發生在「同時」(Simultaneous) 的兩個聲音。另有一種遮蔽效應現象，是出現在前後不同時間的兩個聲音。**圖四**的橫軸是時間，縱軸是聲壓。**圖四**顯示，先有一個較強的聲音（**遮蔽者**，Masker）維持 200ms (10^{-3} 秒，毫秒）後停止，但隨後接著另一較小的聲音。如果後起聲音不夠大，則人類主觀是聽不到的，因為聽覺細胞經過聲音激發後，需一小段時間休息才能再有反應，此稱為**後遮蔽效應**（Post-masking)。此外，還有一種「後一段聲音遮蔽前一段聲音」的**前遮蔽效應**（Pre-masking)，至於為何會產生「後發而遮蔽先發」現象的原因較難解釋。

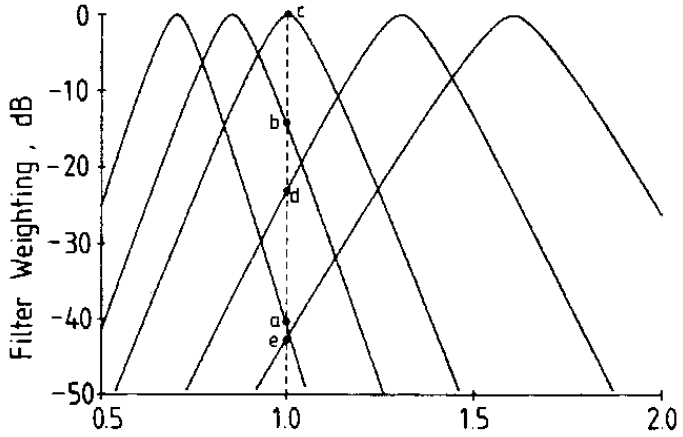


圖四 前後時遮蔽效應: 圖中橫軸是時間，Masker 在中段的 0 秒鐘發生，經過 200ms 消失。但它可遮蔽其前(遮蔽期較短)和其後(遮蔽期較長)的聲音訊號。

上面兩個遮蔽效應，特別是同時隣頻的遮蔽效應在壓縮過程中扮演非常重要的角色。既然人耳聽不到，我們就不用紀錄下來或傳送出去。所以，我們觀察經 MP3 壓縮過的波形，與原聲音波形常有很大差別，只是人耳聽不出來。有一點需要略作說明的是，因為遮蔽等人耳特性是「**主觀的**」，每個人的耳朵靈敏度略有不同，因此，這些靈敏度值因人與環境而異，不是絕對的。人耳經過訓練後，在寂靜

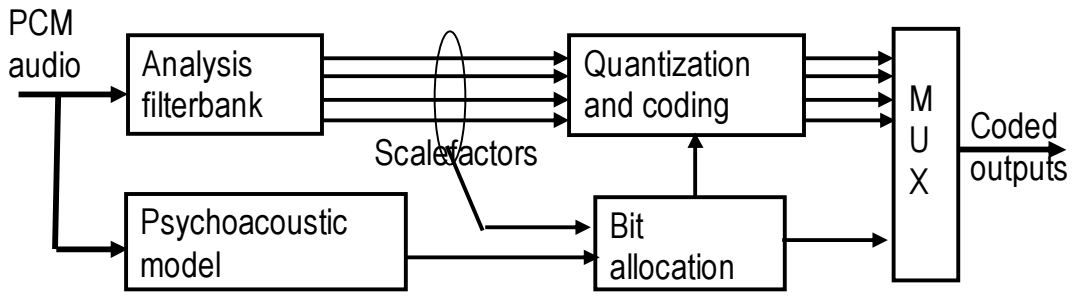
背景下，仍可能察覺到被遮蔽聲音對整體音質的細微影響。但是對大部分使用者及其環境(如嘈雜的公車上)而言，背景雜音很高，現有 MP3 及其他壓縮法提供的品質多可接受。

與上述遮蔽效應有關的人耳特性是所謂**主頻帶**(Critical band)概念，也就是說，人耳聽覺範圍的頻率可切割為 (略相重疊)二十幾個主頻帶。在同一頻帶內多個聲音會互相遮蔽影響，但頻帶與頻帶之間彼此影響較弱而可忽略。這些頻帶在低頻較窄，高頻較寬，差距較大，如**圖五**所示。注意其橫軸座標是對數，每格差 10 倍。



圖五 人耳主頻帶分割示意圖: 橫軸為頻率(KHz)。圖中顯示主頻帶在不同頻率時寬度不同，並互相重疊。又本圖僅為示意圖，未將所有二十幾個主頻帶準確完整畫出。

將以上幾項工具合在一起，形成所謂**聲訊感知型編碼**(Audio Perceptual Coding) 。**圖六**是個簡化的一般系統圖。聲音訊號經過取樣後，先經過**濾波器群組**(Analysis Filter Bank)。其目的是將訊號分解成多個頻帶，以配合人耳主頻帶特性。其次，每個頻帶(或鄰頻帶)作遮蔽效應等的分析，去掉不傳送的部份，即圖中**心理聲學模式** (Psychoacoustic Model)。剩餘部分，每個頻帶再分別經量化(每個取樣取其近似值)，最後做 VLC，即圖中**量化編碼部分** (Quantization and Coding)。儲存或傳送時，多個頻帶輸出要排列在一個檔案或位元流中，此程序稱為**多工**(Multiplexing，MUX)。圖中有**一位元分配方塊**(Bit Allocation)，因為內容較專業，此處略過。 接收端 (解碼器或播放器) 將上述步驟反轉過來，還原回原來聲訊。MP3 和下一段中敘述的其他聲訊標準都有類似這樣的大架構，但各規格的細節有些不同。經過多年來的經驗累積，在大架構中每一方塊裡都有許多技術細節，關於那些細節在此就不詳述。



圖六 聲訊感知型編碼架構圖: 左邊取樣後聲音訊號經過幾項程序處理後(參內文解釋)，成為最右邊二元編碼數據輸出。

最後，我們略述有關 MP3 等音樂壓縮的歷史。現在市場上流行的壓縮音樂格式，有國際標準組織制定的，也有私人商業公司自行發展的。前者最知名的是 MP3 與 AAC (Advanced Video Coding)，後者有微軟(Microsoft) 的 Windows Media Audio (WMA) 與杜比 (Dolby) 的 AC3 (Adaptive Transform Coder 3)。MP3 與 AAC 都是由同一標準組織 ISO/IEC MPEG (Moving Picture Experts Group) 群組所制定出來的。ISO (International Standard Organization) 與 IEC (International Electrotechnical Commission) 是兩個獨立的國際標準組織，他們合組成一個委員會，稱之為 JTC1 (Joint Technical Committee One (on Information Technology))，MPEG 為其下一個工作專家組。自 1988 年成立後，MPEG 標準委員會於 1992 年制定用於 VCD 的 MPEG-1，1994 年制定用於 DVD 的 MPEG-2，相關產品在商業市場上極為成功。國內參與 MPEG 標準的單位不多。十餘年前工研院曾斷續參加過會議，但未持續。1999 年開始，因為之前在美國工作時的相關經驗，交通大學蔣迪豪教授和我又再參加 MPEG 標準會議。2002 年蔡淳仁教授由美返台加入，並得交大李素瑛等教授合作支援，遂較積極參與至今。我們做的比較偏視訊，曾完成 MPEG-4 Part 7 Optimized Reference Software，和 MPEG-21 Part-12 Multimedia Test Bed for Resource Delivery。

最後解釋一下 MP3 的名稱來由。MPEG-1 標準的第三部分 (Part 3) 是聲訊編碼，其中由易而難，分成三層 (Layer)。MP3 是最複雜但壓縮效率最好的第三層 (Layer 3)，廠商自取名為 MP3，此非官方正式名稱。AAC 是 MPEG-2 的第七部分 (Part 7)，其壓縮效率更好，是 Apple iPot 的主用格式。其後，MPEG 標準委員會持續改進，更新的標準有 HE-AAC (High-efficient AAC 或 AAC+) 和 MPEG Surround。(HE-AAC 屬於 MPEG-4 Part 3AAC 的延伸，MPEG Surround 則屬於 MPEG-D (ISO/IEC 23003)。)有興趣技術細節朋友可參考，中文書有交大吳炳飛教授，AUDIO CODING MP3 篇技術手冊(全華圖書，2007)，英文書有 M. Bosi and R.E.

Goldberg, *Introduction to Digital Audio Coding and Standards*, Kluwer, 2003。