音訊壓縮 － MP3 簡介

黃育銘
國立暨南國際大學資訊工程學系助理教授
E-Mail : ymhuang@csie.ncnu.edu.tw

1. 引言

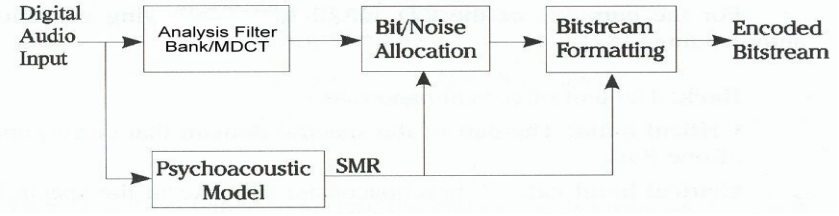
過去市面上 74 分鐘的光碟片（CD）大約僅可儲存 8-12 首音樂，如今它卻可儲存多達 80-120 首。20Hz-20kHz 頻率之音訊為人類聽覺系統所能感知的，因此於數位錄音時，如果以 44.1kHz（根據樣本取樣率（sampling rate）理論至少需為 20kHz 的兩倍）之音訊取樣率（也就是每秒錄製 44100 個音訊樣本）、每個樣本以 16 個位元來儲存，其資料量每秒多達 689K（(44100×16)/1024，單聲道）個位元。有鑑於此，MPEG（Moving Pictures Expert Group）組織提出了第一個音訊壓縮標準 – MPEG-1 音訊壓縮，它利用了人類聽覺系統的特性（所謂感官編碼（perception coding）方式）達到資料壓縮，且尚能維持近 CD 品質的聲音。在 MPEG-1 音訊壓縮標準裏，又可分成三種標準– MPEG-1 Layer I 音訊壓縮、MPEG-1 Layer II 音訊壓縮、及 MPEG-1 Layer III 音訊壓縮。表一為音訊在壓縮過程中，這三種壓縮標準分別每秒可以產生之位元流（bit-stream）量（例如 64Kbps：單聲道每秒產生 64K 位元流）。其中 Layer I 及 Layer II 在壓縮過程中僅提供固定大小之位元流輸出量，也就是說 bitrate_index 參數（隱藏在 MPEG-1 音訊位元流之標頭裏，如圖十所示），在壓縮過程中自始至終都是相同的；然而 Layer III 壓縮過程中，會隨著當時音訊的特性適時選用 bitrate_index 參數，以達到最好的壓縮效果，然而相對地其壓縮演算法的複雜度最高。MP3 即是 MPEG-1 Layer III 音訊壓縮標準的縮寫，一般常見的 MP3 音樂檔案，其平均每秒之資料量大約為 125K 個位元（雙聲道）。

	Bit rate (Kbps)		
Index	Layer I	Layer II	Layer III
0000	free format	free format	free format
0001	32	32	32
0010	64	48	40
0011	96	56	48
0100	128	64	56
0101	160	80	64
0110	192	96	80
0111	224	112	96
1000	256	128	112
1001	288	160	128
1010	320	192	160
1011	352	224	192
1100	384	256	224
1101	416	320	256
1110	448	384	320

Source: © 1993 ISO/IEC.

表一、MPEG-1 音訊編碼過程中，可能之位元流輸出率

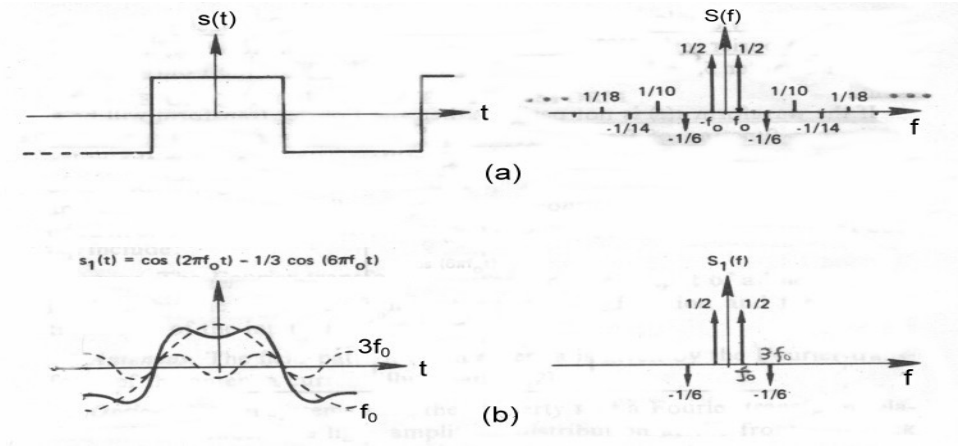
2. MPEG-1 音訊編碼過程



圖一、 MPEG-1 音訊編碼過程

基本上，MPEG-1 音訊演算法 — 利用聽覺心理學理論所衍生之一套有效音訊壓縮演算法。圖一為 MPEG-1 音訊編碼過程之模組架構，接下來我們將簡單地介紹各模組原理及功能。

2.1 頻率域



圖二、 時域（time domain）t 與頻率域（frequency domain）f

圖二(a)表示，訊號 s(t) 經由傅立葉（Fourier）轉換產生 S(f)，藉由 S(f)可分析出該訊號的頻譜（spectrum，頻率所分佈之範圍）及該訊號在這頻譜上之某一頻率相對之子訊號所含有的能量，其所分佈的頻率 f 等於 f_0、$3f_0$、$5f_0$、$7f_0$…，且該頻率上之子訊號的大小值 |f| 分別為 1/2、1/6、1/10…。s(t)與 S(f)的關係如下：

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt$$ 及

$$s(t) = \int_{-\infty}^{\infty} S(f) e^{j2\pi ft} dt$$ 。

圖二(a)中所示之 s(t) 訊號為由無窮多個餘弦波形子訊號（$\cos[2\pi f_0 t]$、$\cos[2\pi(3f_0)t]$、$\cos[2\pi(5f_0)t]$、…, 其頻率分別為 f_0、$3f_0$、$5f_0$、…）所合成，而圖二(b)所示僅為前兩個餘弦波形子訊號（其頻率分別為 f_0 及 $3f_0$）所合成之訊號 $s_1(t)= \cos(2\pi f_0 t)-1/3\cos(6\pi f_0 t)$，其已相當接近訊號 s(t)，這是因為 s(t)訊號之高頻子訊號其能量遠低於低頻子訊號的能量，因此如果捨棄該訊號之高頻子訊號，其失真度自

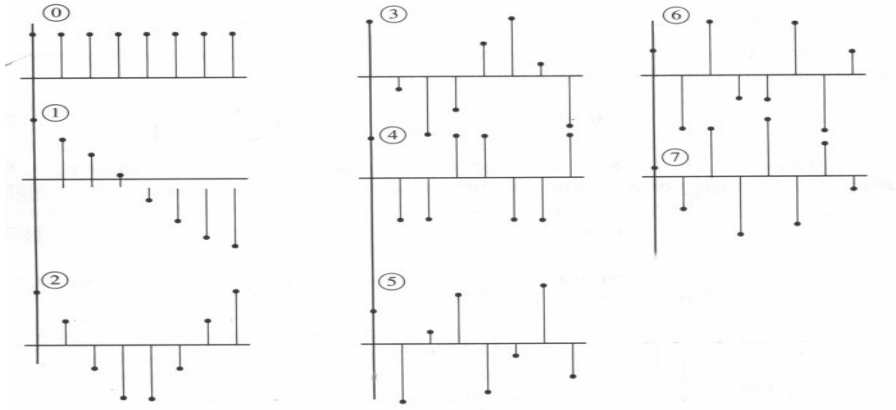
然不會太明顯。

在影像或聲音的壓縮演算法裏，離散餘弦轉換（Discrete Cosine Transform，DCT）常被用來作影像或聲音資料的頻譜分析。其觀念同如傅立葉轉換我們將以以下例子作說明：假設 s(n), 0≤ n≤ 7, 為時域上 8 點資料，T(m), 0≤ m≤ 7, 為頻率域上 8 點離散餘弦轉換輸出值，其關係為

$$T(m) = \sum_{n=0}^{7} s(n)\cos[\pi m(2n+1)/16] \text{ 及}$$

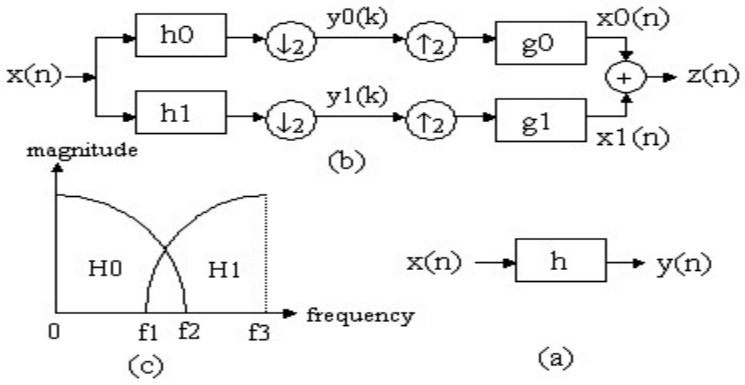
$$s(n) = \sum_{m=0}^{7} T(m)\cos[\pi m(2n+1)/16] \text{ 。}$$

圖三分別為 m=0,1, …, 7 之對應 8 個 (第 m 個向量為 $\cos[\pi m(2n+1)/16]$, n=0, 1, …, 7) 向量。也就是說 s[n]訊號可視作由圖三八種不同頻率之子訊號所合成，該子訊號之大小分別為 T(0), T(1), T(2), …, T(7)。圖三中第一種子訊號頻率最低，第八種子訊號頻率最高，且每一種子訊號之波形有如餘弦波形。



圖三、 八種不同頻率的子訊號

2.2 濾波庫 （filter bank）



圖四、 (a) Single Filter

(b) Two-Channel Filter Bank (c) Frequency responses of h0 and h1

所謂濾波器（Filter），即對所輸入的訊號僅保留其某一範圍頻率之所有子訊號

其餘頻率之子訊號則會被過濾掉。圖四(a)所示為一單一濾波器，輸入訊號為 x(n)，h(n)為該濾波器（h）之系統參數值，輸出訊號為 y(n)，則其關係如下：

$$y(n) = \sum_{m=-\infty}^{\infty} h(n-m)x(m) \equiv x(n) \otimes h(n)\text{，}n = -\infty \sim \infty\text{。} \quad (1)$$

圖四(b)所示為一雙頻道（channel）之濾波庫，其由一個低頻（low-pass）濾波器（h0）及一個高頻（high-pass）濾波器（h1）所組成，h0(n)及 h1(n)之頻譜分別為 H0（0~f2）及 H1（f1~f3），如圖四(c)所示。在圖四(b)中，↓2（在分析濾波庫中稱作 decimation by 2）表示對所輸入的訊號序列兩點僅取一點，因此

$$y0(k) = x(n) \otimes h0(n)|_{取\ n=2k} = \sum_{m=-\infty}^{\infty} h0(2n-m)x(m) \quad (2a)$$

$$y1(k) = x(n) \otimes h1(n)|_{取\ n=2k} = \sum_{m=-\infty}^{\infty} h1(2n-m)x(m)\text{。} \quad (2b)$$

↑2 表示對所輸入的訊號序列每兩點之間插入一個 0（在合成濾波庫中稱作 interpolation by 2），因此

$$x0(n) = \sum_{k=-\infty}^{\infty} g0(n-2k)y0(k) \quad (3a)$$

$$x1(n) = \sum_{k=-\infty}^{\infty} g1(n-2k)y1(k)\text{。} \quad (3b)$$

假設 h0、h1、g0、g1 等濾波器之系統參數值分別設計如下：h0(-1)=h0(0)=1/2；h(-1)=-1/2, h(0)=1/2；g0(0)=g0(1)=1；g1(0)=1, g1(1)=-1 其它未指定的參數均為 0。

根據(2)之式子，可推得

y0(k) = [x(2k)+x(2k+1)]/2

y1(k) = [x(2k)-x(2k+1)]/2

根據(3)之式子，可推得

x0(2k) = y0(k); x0(2k+1) = y0(k);

x1(2k) = y1(k); x1(2k+1) = -y1(k);

因此，

z(2k) = x0(2k)+x1(2k) = y0(k)+y1(k)；

z(2k+1) = x0(2k+1)+x1(2k+1) = y0(k)-y1(k)。

我們再以以下例子作說明：

假設 x(n)={10, 14, 10, 12, 14, 8, 14, 12, 10, 8, 10, 12}, n=0, …, 11，其經過一個低頻濾波器（h0）及一個高頻濾波器（h1），分別產生較低頻 a(n) 子訊號及較高頻 b(n) 子訊號（a(n)及 b(n)仍為時域上的值）。

a(n) = [x(n)+x(n-1)] / 2 = {12, 12, 11, 13, 11, 11, 13, 11, 9, 9, 11, 12}，

b(n) = [x(n)-x(n-1)] / 2 = {-2, 2, -1, -1, 3, -3, 1, 1, 1, -1, -1, 0}，

x(n) = a(n)+b(n), 其中假設 x(12)=12。

然而如果對 a(n)或 b(n)只取偶數點（n=0, 2, 4, …）部份，分別為

y0(k) = {12, 11, 11, 13, 9, 11},

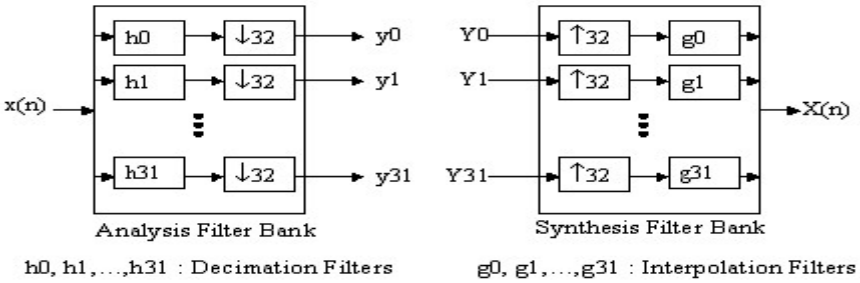
y1(k) = {-2, -1, 3, 1, 1, -1}, k=0, 1, …, 5。

將 y0 及 y1 序列分別作相加及相減得到

y0(k)+ y1(k) = {10, 10, 14, 14, 10, 10}，

y0(k)-y1(k)= {14, 12, 8, 12, 8, 12}。

所以 z(n) = {10, 14, 10, 12, 14, 8, 14, 12, 10, 8, 10, 12}。

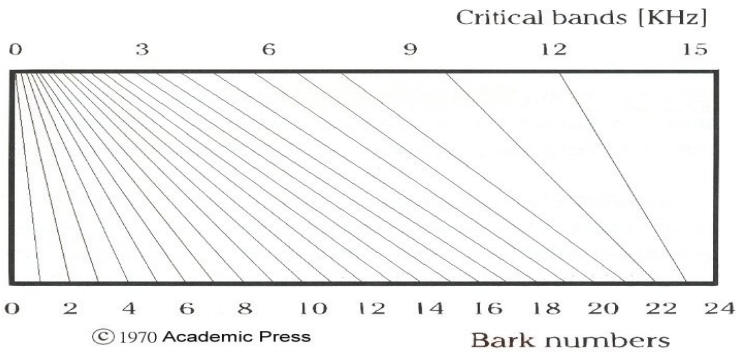


圖五、 分析濾波庫及合成濾波庫

類似上述雙頻道之濾波庫原理，在 MPEG-1 音訊編碼器中（如圖五所示為 32 頻道之濾波庫），分析濾波庫（analysis filter bank）將訊號 x(n) 原先之頻譜均分成 32 等份，每一等份稱之一個次頻帶（subband），其頻寬（bandwidth，該頻帶的總寬度）均相等（等於 Fs/64, Fs 表示樣本取樣率且等於訊號 x(n)之頻寬的兩倍），也就是說將原先訊號 x(n)，透過 32 個中頻濾波器（band-pass filters），轉換成 32 組子訊號 y0(k), y1(k), …, y31(k)。其中輸入訊號 x(n)之點數與 32 組輸出子訊號 y0(k), y1(k), …, y31(k)的總點數相等，一般我們稱之”critically sampled”。而在 MPEG-1 音訊解碼器中，則藉由合成濾波庫（synthesis filter bank）， 可將 32 組中頻子訊號 Y0(k), Y1(k), …,Y31(k)還原成一組全頻（原先）訊號 X(n)。

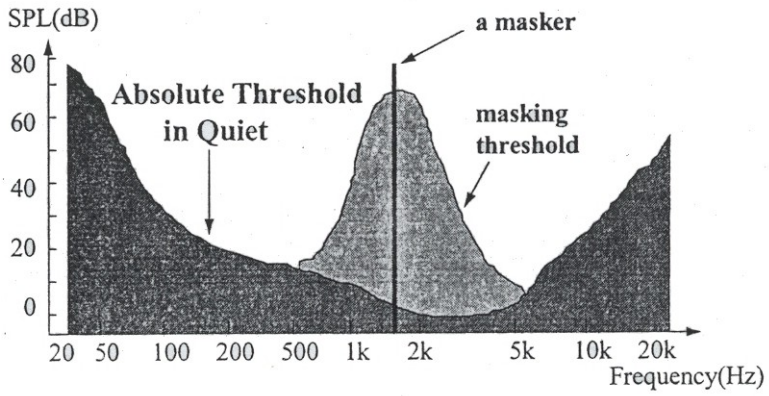
人類聽覺系統對於不同頻率的聲音，其敏感度與頻率並不是保有一線性關係圖六為將一個 15.5 KHz 頻帶分割成 24 個次頻帶（Bark1, Bark2,…, Bark24），每一個次頻帶（稱之一 critical band）其頻寬不相等。對於同一個 critical band 裏的聲音，敏感度幾近相同。

因此，在 MPEG-1 Layer III 標準裏，為了達到更高頻率域上的解析度（resolution in frequency domain），針對圖五中之 32 組次頻帶上之中頻子訊號，每一組再經由 MDCT（Modified Discrete Cosine Transform）轉換成 18 組頻率域上的訊號，亦即共有 576（32×18）組頻率域上之子訊號。接下來這 576 組子訊號，根據每一個 critical band 的頻率範圍群組（grouping）成 24 組，每一組稱之一 scale factor band（非常逼近於某一 critical band），藉由 2.3 節即將介紹之聽覺心理學模型，算出每一組之 SMR 值。

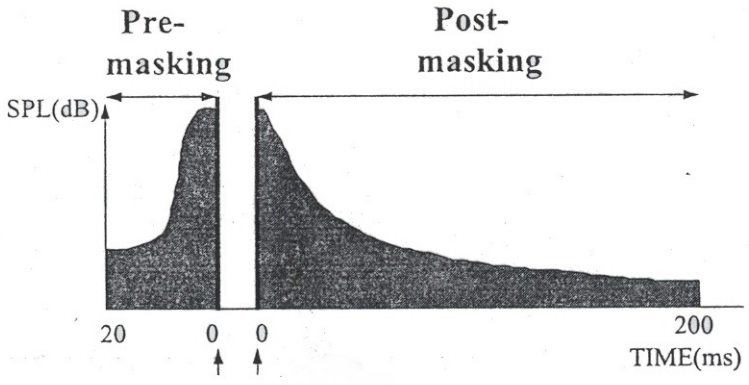


圖六、 Nonlinear critical bands

2.3 聽覺心理學模型（Psychoacoustic Model）



圖七、 頻譜域上之遮蔽效應（Spectral Masking Effect）



圖八、 時域上之遮蔽效應（Temporal Masking Effect）

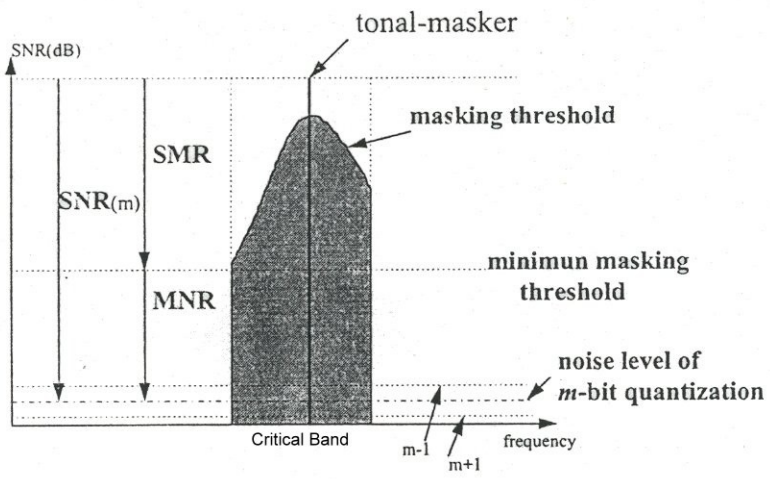
如圖四所示，在不同的頻率人耳所能聽到的最小音量（Sound Press Level（SPL），等於 20logM dB，M 為該頻率聲音之音量）有所不同，即所謂的”靜音的絕對門檻（Absolute Threshold in Quite）”，也就是說在靜音的絕對門檻的曲線下之音量為人耳所不能察覺的。因此，人類聽覺系統對於頻率小於 40Hz 或大於

10KHz 之音訊較不敏銳，而最敏銳的音訊其頻率大約介於 2~4KHz 。除了考慮上述人耳的自然現象外，還利用了音訊訊號頻譜域上之遮蔽效應及時域上之遮蔽效應以達到對音訊資料作壓縮。

圖七所示為一頻譜域上之遮蔽效應現象，當某一個頻率的聲音（稱之 masker）很強時，對於鄰近頻率的聲音，若其聲音大小小於某個範圍（稱之 masking threshold），則這些頻率的聲音也是人耳所不能察覺的。masking threshold 曲線會隨者 masker 聲音之頻率及大小而有所不同。

圖八所示為一時域上之遮蔽效應現象，在時間軸上突然聲音強度遠高於鄰近聲音的強度（大約 40dB 以上），則對於先前已發生的聲音（echo）或爾後即將產生的聲音，同樣地若其聲音大小小於某個範圍，則這些聲音也是人耳所不能察覺的。前者稱之 Pre-masking ，大約長達 5~12ms ；後者稱之 Post-masking ，大約長達 50~200ms。

2.4 位元/雜訊 分配法則（Bit/Noise Allocation）



圖九、 Mask-to-Noise Ratio (MNR)

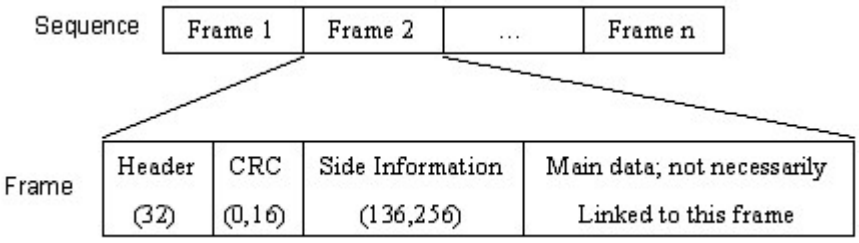
如圖九所示，假設在某一 critical band 上有 5 筆資料要編碼，其大小值介於 0~8 且分別為 1.5, 7.0, 4.5, 5.0,及 3.0 。如果僅用兩個位元來編碼，00 表示大小為 1, 01 表示大小為 3, 10 表示大小為 5, 及 11 表示大小為 7。則這 5 筆資料將會被編碼成 00, 11, 10, 10, 及 01（即分別表示大小值為 1, 7, 5, 5, 及 3 ，與原先資料大小值有些微差距）， 該編碼過程稱作量化 (quantization)，其差距值稱作量化雜訊 (quantization noise)。原先 5 筆資料能量(energy)定義成 10logM dB，M 為每個資料大小平方和再取平均；而量化雜訊（ 0.5, 0, 0.5, 0, 及 0 ）之能量為 10logN dB ，N=(0.25+0+0.25+0+0)/5 。當然隨者可用來編碼之位元數增加時，相對的量化雜訊之能量會降低，也就是用 m+1 個位元來編碼會比用 m 個位元來編碼所造成之量化雜訊稍低。

Signal-to-Noise Ratio (SNR)等於 10log(M/N) (亦即 10logM-10logN)。在每一個 critical band 裏之 masking threshold 曲線，找出最小的 masking threshold 值 (等於 SMR(Signal-to-Mask Ratio))，所以 MNR（Mask-to-Noise Ratio）= SNR-SMR。

在位元或雜訊分配模組裏，即是在處理如何將可用來編碼之有限位元數最佳地分配給每一個 critical band，使得每一個 critical band 之 MNR 值會升高，也就是說儘量降低每一個 critical band 裏之量化雜訊值（因對在這 critical band 上的資料作量化，進而達到資料壓縮）。

在 Layer III 標準裏，量化採用非線性量化器 (non-linear quantizer), 且經過量化編碼過的資料會再以 Huffman 碼作進一步編碼。

2.5 編碼後之音訊位元流(Audio Encoded Bitstream)



圖十、 MPEG-1 Layer III 編碼後之音訊位元流格式

整個 MPEG-1 編碼後之音訊位元流由一連串的 frame 位元流所組成，如圖十所示。 每個 frame 中擁有 384（從 32 組子訊號中，每組各取 12 點，共有 12×32=384）點子訊號的資料，而 Layer II 及 Layer III 每個 frame 中擁有 1152（從 32 組子訊號中，每組各取 36(12×3)點，共有 36×32=1152）點。根據 SMR 值，這些資料會被量化編碼，之後再經 Huffamn 編碼成一串位元流存在該 frame 中之 Main data 欄位。當這一串位元流資料量過大時，則會將部份位元流存入另一個 frame 中之 Main data 欄位，這種觀念在 Layer III 裏稱作”bits reservoir”。

標頭(Header)欄位含有 32 個位元，CRC 欄位含有 0 或 16 個位元，side information 欄位含有 136 或 256 個位元，這些額外訊息讓解碼器方便作解碼或作進一步編輯。

3. 結論

當一首歌曲中有比較急促的聲音或聲音不是很圓滑的時候，MP3 歌曲裏的雜訊就會很顯著。

參考文獻

[1] ISO/IEC JTCI/SC29, “Information Technology – coding of moving pictures and

associated audio for digital storage media at up to about 1.5 Mpbs – IS 11172 (Part 3, Audio),” 1992.
[2] K. R. Rao and J. J. Hwang, “Techniques & Standards For Image • Video & Audio Coding,” PTR/PH 1996
[3] 古媄君, “MPEG-1 音訊編解碼器之研究與其即時軟體之實作,” 台大資工系碩士論文, 1996.
[4] S. Shlien, “Guide to MPEG-1 Audio Standard,” IEEE Trans. On Broadcasting, Vol. 40, No. 4, pp. 206-218, Dec. 1994.
[5] D. Pan, “A Tutorial on MPEG/Audio Compression,” IEEE Multimedia Magazine, pp. 60-74, Summer 1995.
[6] P. Noll, “MPEG Digital Audio Coding,” IEEE Signal Processing Magazine, pp. 59-81, Sep. 1997.