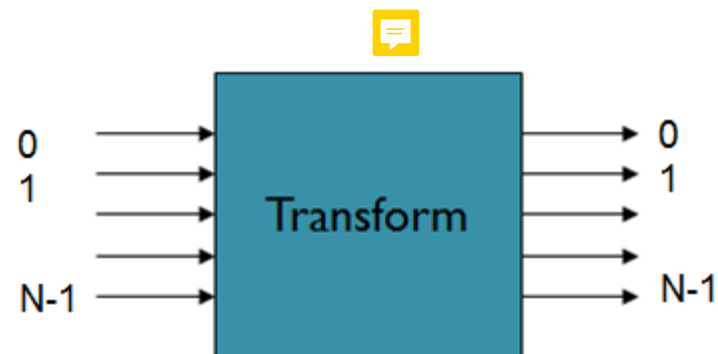


a type of data compression for "natural" data like audio signals  
typically lossless but is used to enable better quantization, which then results  
in a lower quality copy of the original input

# Transform Coding

# Transform Coding

The transform usually takes a block of  $N$  time domain input samples and produces a block of  $N$  frequency domain output samples. These output samples provide the information of frequency distribution of the input samples.

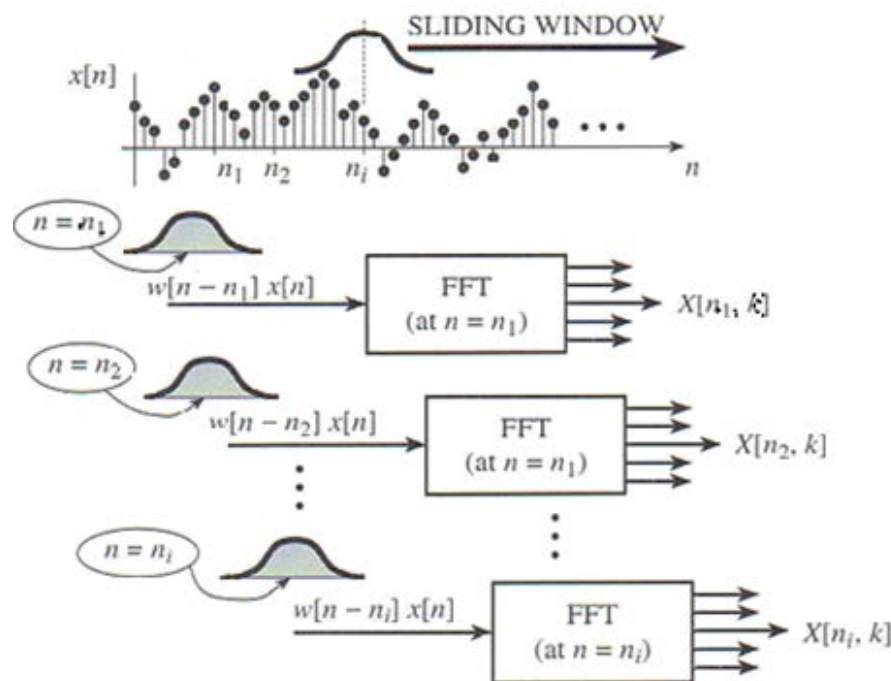


- In addition to filter bank, another approach to the time-to-frequency mapping is to use transform for spectrum analysis
- For example, transform  $N$  data samples into  $N$  freq bins based on a particular type of transform.
- Coders with a small number of freq sub-bands (e.g. MPEG I layer I and II) are called **sub-band coders**, and those with a large number of freq bins (e.g. Dolby Digital, MPEG AAC) are called **transform coders**
- MP3 (MPEG I layer III) uses hybrid of sub-band and transform coders. Other coders use either sub-band or transform approach only.

It first uses sub-band filters to divide the signal into a few sub-band signals that has a narrower bandwidth. Each of these narrow band signals is further processed by the modified discrete cosine transform (MDCT) to obtain many frequency bins.

# Issues of windows

- The transform coder, is operated on a finite number of data samples. eg., 512-point FFT collecting 512 samples for each transformation.
- A window function  $w(n)$  is to segment the input data stream,  $x(n)$ , see the figure for STFT
- The FFT outputs are from  $w(n)*x(n)$  rather than the FFT of  $x(n)$ .
- There will be significant effects on the quality of the audio at the output of the decoder if the window effects are not properly dealt with.



# Issues of windows

- The two main purposes of a window in the time domain are
  - to obtain a segment by multiplying the window function,  $w(n)$ , with the input stream  $x(n)$  in the time domain
  - To filter out the undesired frequency bands, which requires the knowledge of filter design
- The **time resolution** of the transform coder is defined as  $T_s * N$ , where  $T_s$  is the sample period and  $N$  is the window length
- The **frequency resolution** of a  $N$ -point transform coder is  $F_s/N$ , where  $F_s$  is the sample frequency.

can only keep one good

- In audio coding, window plays an important role in :
  - Maximizing frequency separation (or sufficient frequency resolution)
  - Minimizing the effects of audible blocking artifacts

relate to time domain resolution

# Issues of windows

- Two window parameters that are relevant to the above:
  - Window length
  - Window shape
- No single-shape window is optimal for all signals in terms of frequency and time domain resolutions

In audio coding, window length is a very important parameter to keep a balance of both good frequency and time resolutions with minimum computation costs.

## Issues to be dealt with:

- Longer window length means better frequency resolution, but poor time resolution.
- Shorter window length means better time resolution, but poor frequency resolution
- Dynamic signal contents need resolution variation.
- **Solution:** dynamically select the windows according to data contents

# Issues of windows

- Satisfying the **perfect reconstruction** condition on the window to remove the window effects on the data.
- Time and frequency resolutions

We shall deal with

- Transform coding
- Window function and window shape selection
- **Modified discrete cosine transform (MDCT)**

# General approach – Transform coding

- At the encoding (analysis) side
  - We divide input sequence into many small segments with a well-designed analysis window  $w_a$ .
  - Overlapping between adjacent windows are necessary to minimize the window effects.

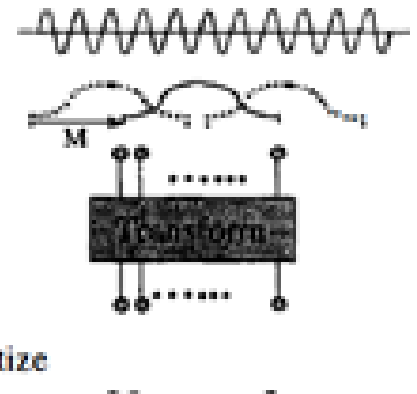
Each data segment is processed by a transform of  $N$  points.

- These transformed outputs are quantized and compressed according to the hearing threshold before transmission or stored in memory.

1) Slide  $M$  samples and Window (window length =  $N$ )

2) Perform an  $N$ -point Transform

3) Quantize, Store/Transmit, Dequantize

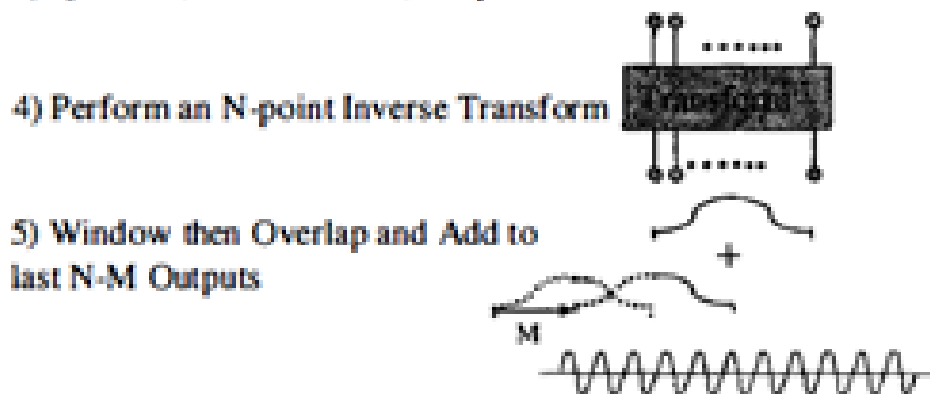


The outputs of the transform are further processed. According to the magnitudes and the frequency position, these transform outputs are quantized, transmitted or stored in the memory, while the others may be ignored since they are not useful for our hearing system.



## General approach – Transform coding

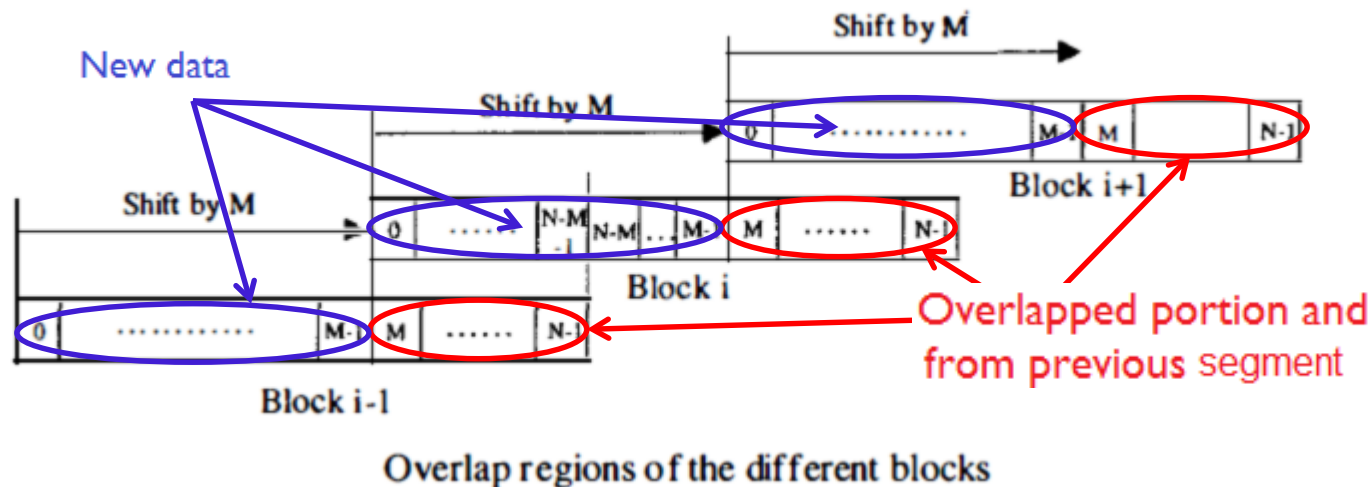
- At the decoding (synthesis) side
  - The received outputs are dequantized
  - Each dequantized segment is processed individually and these successively processed segments are combined to form the corresponding output data sequence



- By ignoring effects from step 3) in the figure, the basic requirement is that the output obtained at step 5) is to be the same as the input, except a constant scaling factor on the amplitude of sequence output and a constant delay.



# General approach – Transform coding



Let us consider the processing details for every block of  $N$  data samples

- The window is shifted to get  $M$  new samples for the current segment.
- The overlapping length is  $N-M$  samples ( $\leq N/2$  for this application)

# General processing approach – Transform coding

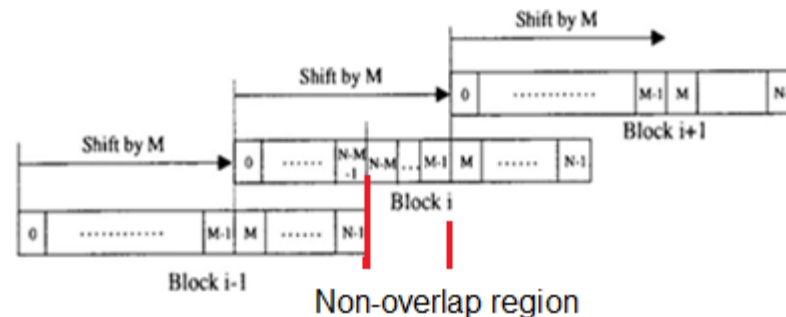
- In a similar way, at the decoder, we need to perform  $N$ -point inverse transform.
- The transform outputs are processed by a  $N$ -point synthesis window  $w_s(n)$
- Then  $M$  samples are used as the decoder output for each segment. *same amount of overlapping is used in both analysis and synthesis processes*
- This arrangement also keeps the output data rate to be the same as the input data rate.
- **Perfect reconstruction** is to achieve

$$x'(n) = Cx(n-D),$$

where  $x'(n)$  is the output of the decoder and  $x(n)$  is the input of the encoder (we have seen the same condition for subband coders) and  $D$  is a positive integer.

# General processing approach – Transform coding

- In any segment region without overlapping, for  $N=N-M, \dots, M-1$ , we require the signal windowed with both the analysis and synthesis windows be equal to the original signal



- The condition in terms of window function is

$$w_a^i[n] * w_s^i[n] = 1 \quad \text{for } n = N - M, \dots, M - 1$$

region without any overlapping

math equation to achieve perfect reconstruction. This equation is derived by considering both the transform and inverse transform operations.

We consider the current segment indexed by  $i$  and the previous segment indexed by  $i-1$ , In this region, the product of the window used for analysis process and that used for the synthesis process should be one.

It means that there is no net effects of the two windows on the input data.

# General processing approach – Transform coding

- Two overlap regions:
  - One is  $n=0, 1 \dots N-M-1$  with the previous segment  $i-1$
  - The other is  $n=M, \dots, N-1$  with the next segment  $i+1$
- The condition for perfect reconstruction is

$$w_a^i[n] * w_s^i[n] + w_a^{i-1}[M+n] * w_s^{i-1}[M+n] = 1$$

the sum of the windowed signals from both segments should be the original signal

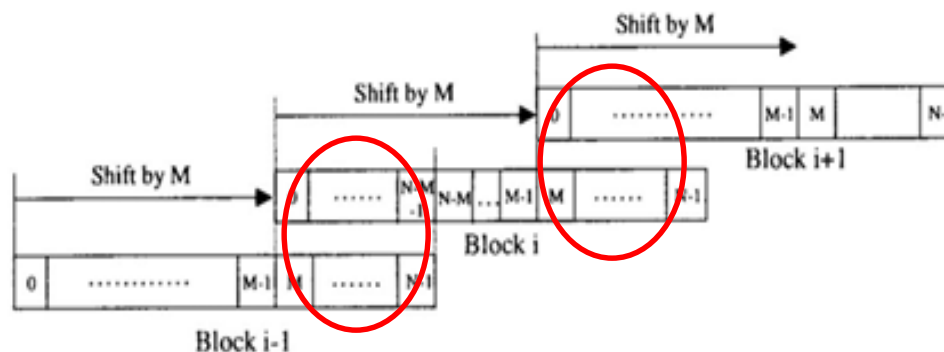


Figure 7. Overlap regions of the different blocks

In the equation or in the figure, it is seen that this condition relates the right side of the window of one segment with the left side of the window of the following segment.

# General processing approach – Transform coding

- If we choose to work with identical analysis and synthesis windows, then we find that the perfect reconstruction conditions simplify to the equation below.

$$\begin{aligned} w^i[n]^2 + w^{i-1}[M+n]^2 &= 1 \quad \text{for } n = 0, \dots, N-M-1 \\ w^i[n]^2 &= 1 \quad n = N-M, \dots, M-1 \end{aligned}$$

- It has been found that the sine window defined below meets the condition shown in the above equation.

$$w[n] = \begin{cases} \sin\left[\frac{\pi}{2} \frac{n+1/2}{N-M}\right] & n = 0, 1, \dots, N-M-1 \\ 1 & n = N-M \dots M-1 \\ \sin\left[\frac{\pi}{2} \frac{N-n-1/2}{N-M}\right] & n = M \dots N-1 \end{cases}$$

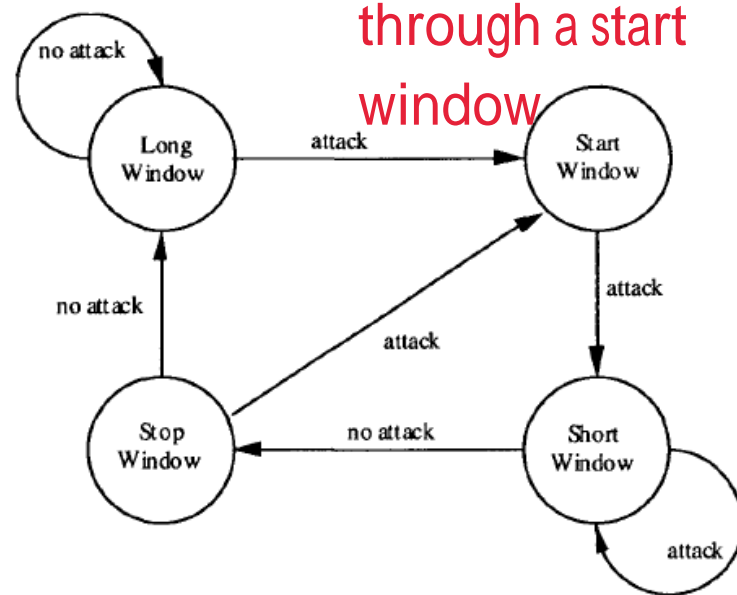
- Note that the window impulse response has three segments, the first one and the last one are overlapped with the previous and succeeding windows, respectively.

# Window Switching

- Long window provides a good frequency resolution, but poor time resolution, and vice versa.

- The distribution of signal energy is monitored and a flat spectrum indicates a sudden change of signal energy (known as **attack**) 突然的变化

- The Figure shows the state transition diagram between long and short window switching.



Block switching state diagram for Layer III from [ISO/IEC 11172-3]

- without losing the perfect reconstruction property
- The figure uses four types of windows – long window, short window, start window and stop window.

Assuming long window is being used, it is switched to short window through a start window when attack is detected, .

When non-attach is detected, the short window is switched into long window through a stop window.

# Window Switching

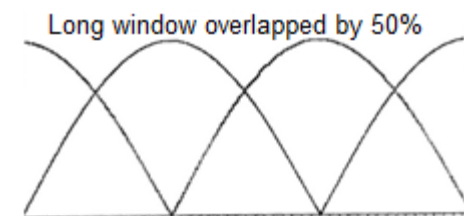
- The long or short windows has different length and their shapes are given in the Figure.

- Their mathematic expression is

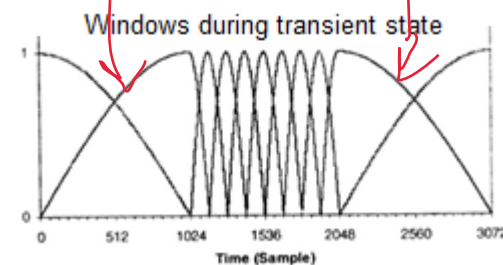
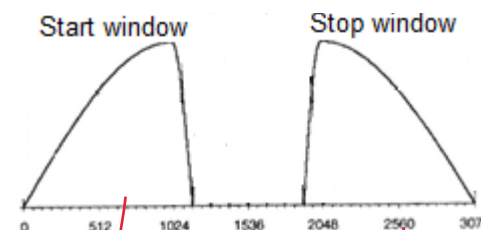
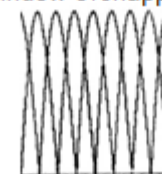
$$w[n] = \begin{cases} \sin\left[\frac{\pi(n+1/2)}{N}\right] & n = 0, 1, \dots, N/2 - 1 \\ \sin\left[\frac{\pi(N-n-1/2)}{N}\right] & n = N/2, \dots, N - 1 \end{cases}$$

- The start window is used for a transition from long to short window, and the stop window is used for a transition from short to long window.

- The Figure at the bottom shows the window changes into short window, and then change into long window again.



Short window overlapped by 50%



perfect reconstruction condition has been always met even during the process of window switching.

# Modified Discrete Cosine Transform (MDCT)

- When performing overlap computation in audio coding (as described), the data rate increases due to the overlapping between the adjacent segments.
- For example, the audio coder processes  $N$  samples for every  $M$  new samples. Since  $M < N$ , the data rate increases by  $(N/M)$ .
- Therefore, it is desired to have a transform that allows overlapping without increasing the data rate.
- The modified discrete cosine transform is a class of transforms that satisfy the above condition.
- MDCT has its inverse operation that recovers signal from overlapping blocks.
- The pair of MDCT and the inverse MDCT together achieves the **perfect reconstruction condition**.

by using the windowing strategy that has been discussed previously



# Modified Discrete Cosine Transform (MDCT)

- The MDCT transform of a block of  $N$  samples,  $x[n]$  (from the  $N/2$  new data samples at frame  $i$  and  $N/2$  from the previous frame  $i-1$ ) to form  $N/2$  frequency domain outputs

$$X_i[k] = \sum_{n=0}^{N-1} \left\{ \underbrace{w_a^i[n] x_i[n]}_{\text{Windowed data}} \right\} \cos\left[\frac{2\pi}{N}(n + n_0)(k + 0.5)\right], \quad k = 0, 1, \dots, \frac{N}{2} - 1$$

window function,  $w[n]$  is multiplied with the data function,  $x[n]$ . The subscript,  $i$ , is the segmentation index, and the subscript,  $a$ , means the analysis process.

- The IMDCT takes the  $N/2$  frequency domain samples  $X[k]$  and returns a  $N$ -point time-domain signal, ready to be overlapped-and-added:

$$x'_i[n] = w_s^i[n] \frac{4}{N} \sum_{k=0}^{N/2-1} X_i[k] \cos\left[\frac{2\pi}{N}(n + n_0)(k + 0.5)\right] \quad n = 0, \dots, N - 1$$

- The original data is recovered by adding IMDCTs of subsequent blocks in their overlapping halves.

# Applications in Audio Coders

three layers of MPEG 1 coding standards.

**TABLE 10.3 Comparison of filter-bank properties.**

Feature	Layer 1	Layer 2	Layer 3
Filterbank type	PQMF	PQMF	Hybrid PQMF/ MDCT
<u>Frequency resolution at 48 kHz</u>	750 Hz	750 Hz	41.66 Hz
<u>Time resolution at 48 kHz</u>	0.66 ms	0.66 ms	4 ms
<u>Impulse response (LW)</u>	512	512	1664
<u>Impulse response (SW)</u>	—	—	896
<u>Frame length at 48 kHz</u>	8 ms	24 ms	24 ms

window size

long window

window duration  
in time

good freq. resolution

poor time resolution

the layer 3 achieves more  
balanced performance  
for various kinds of  
signal contents.

- Detailed information on **spectrum analysis** of standards
- Compare the frequency and time resolution of different coders
- Note that level 3 coding uses a hybrid filterbank performed by the PQMT (Pseudo-Quadrature Mirror Filter) and then the MDCT for better frequency resolution.

both layer 1 and layer 2 use only PQMF to perform sub band filtering.

# MPEG-I – An Example of Signal analysis

- **Step I:** Spectral analysis (discussed previously) and SPL normalization sound pressure level
  - The SPL normalization guarantees that a 4 kHz signal of +/-1 bit amplitude will be associated with a SPL of 0 dB (close to the hearing threshold) smallest dB expressed
  - Let us assume a full scale sinusoid is associated with the SPL of 90 dB, based on the equation

$$x(n) = s(n)/\{N(2^{b-1})\}, \quad (1)$$

where  $s(n)$  is the input signal,  $N$  is the length of window  $w(n)$  and  $b$  is the maximum number of bits per sample. length of window

- In this example, we assume  $x(n)$  is a 12-ms segment of 512 samples.
- With a Hann window, the power spectrum density  $p(k)$  is obtained by

$$P(k) = PN + 10 \log_{10} \left| \sum_{n=0}^{N-1} w(n)x(n)e^{-j2\pi nk/N} \right|^2, \quad 0 \leq k < N/2 \quad (2)$$

here is DFT computation of  $w(n)x(n)$ .

where  $PN=90.302$  dB is a normalization constant

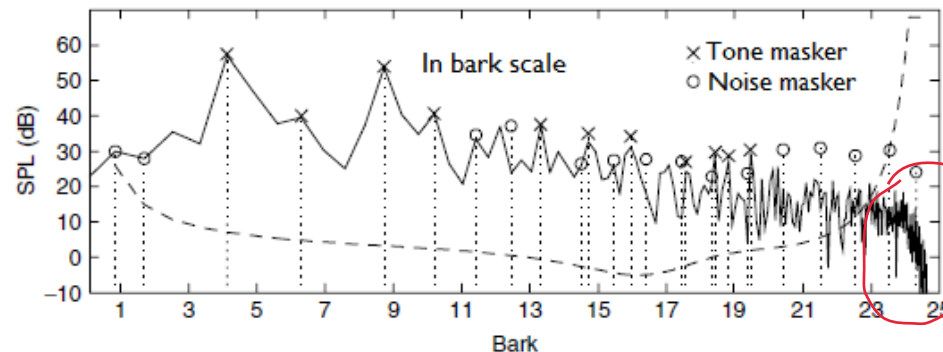
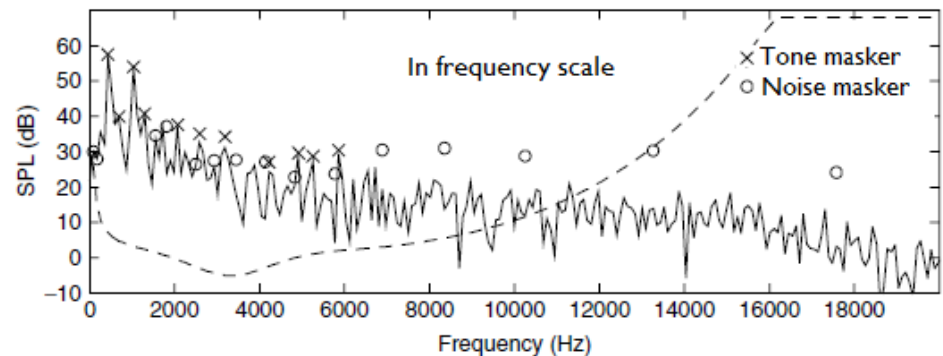
- The second term is the squared DFT (or power spectrum density (PSD)) of the windowed input segment in terms of dB.

# MPEG-I – An Example

- The window function used is

$$w(n) = \frac{1}{2} \left[ 1 - \cos \left( \frac{2\pi n}{N} \right) \right]$$

- The normalization procedure in (1) and the parameter PN in (2) are used to estimate SPL from the input signal
- A full scale sinusoid is processed by a 512-point FFT producing a spectrum line have 84 dB SPL.
- With a 16-bit sample resolution, the SPL of a very-low-amplitude signal will be at or below the hearing threshold.



# MPEG-I – An Example

## Step 2: *Masker identification*

- The frequency component meets the following conditions are classified as the tonal maskers.

$$S_T = \left\{ P(k) \left| \begin{array}{l} P(k) > P(k \pm 1) \\ P(k) > P(k \pm \Delta_k) + 7\text{dB} \end{array} \right. \right\}$$

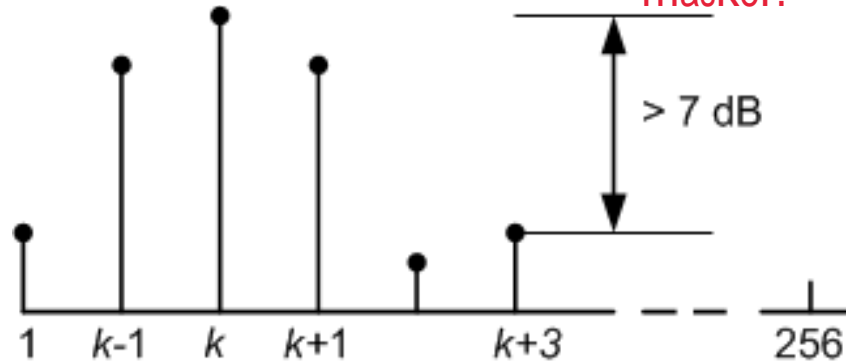
where  $P(k)$  is the PSD and  $k$  is the bark index, and

$$\Delta_k \in \left\{ \begin{array}{lll} 2 & 2 < k < 63 & (0.17 - 5.5 \text{ kHz}) \\ [2, 3] & 63 \leq k < 127 & (5.5 - 11 \text{ kHz}) \\ [2, 6] & 127 \leq k \leq 256 & (11 - 20 \text{ kHz}) \end{array} \right\}$$

- $P(k) > P(k \pm 1)$  means  $P(k)$  is larger than its adjacent neighbors.
- $P(k) > P(k \pm \Delta_k) + 7\text{dB}$  means for non adjacent neighbors, the  $P(k)$  should be larger than 7 dB.
- For higher frequency masker, it should be larger than more neighboring components. For example,  $\Delta_k$  is in the range from 2 to 6 when  $k > 127$ .

# MPEG-I – An Example

The magnitude requirement of a masker.



## Example:

- The Figure shows the SPL for  $k=1 \dots, 256$ .
- The  $k$ th component has a magnitude larger than its neighbors
- The magnitude of the  $k$ th component has a magnitude larger than the  $(k \pm 2)$ th and  $(k \pm 3)$ th component by more than 7 dB
- Therefore, the  $k$ th component is a tone masker.

# MPEG-I – An Example

## Step 3: *Magnitude of maskers*

- The **tonal markers**,  $P_{TM}(k)$ , are computed from the spectral peaks listed in  $S_T$  as follows

$$P_{TM}(k) = 10 \log_{10} \left( \sum_{j=-1}^1 10^{0.1P(k+j)} \right) \text{dB}$$

- This equation shows that the **energy of the  $k$ th masker is the energy sum from three frequency components ( $k-1$ ,  $k$  and  $k+1$ )**
- A single **noise masker** for each critical band is computed from the remaining spectrum line not within the  $\pm \Delta k$  neighborhood of a tonal masker using the sum

$$P_{NM}(\bar{k}) = 10 \log_{10} \left( \sum_j 10^{0.1P(k+j)} \right) \text{dB} \quad \forall P(j) \notin \{P_{TM}(k, k \pm 1, k \pm \Delta_k)\}$$

where  $\bar{k}$  is defined to be the geometric mean spectral line of the critical band

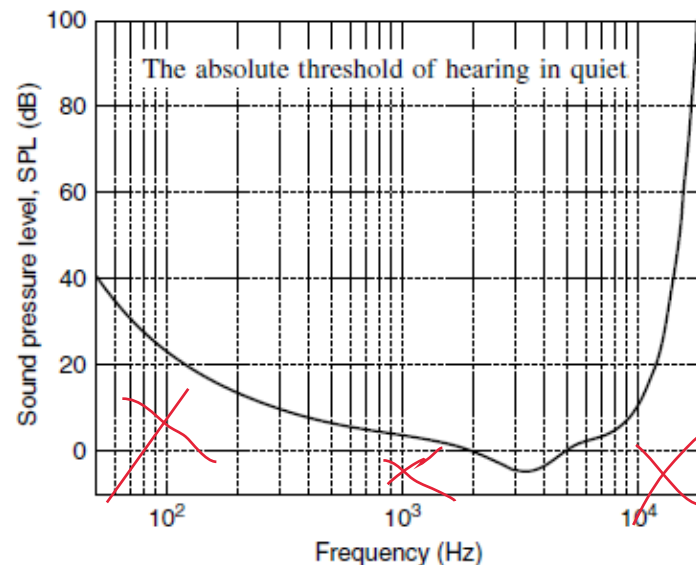
$$\bar{k} = \left( \prod_{j=l}^u j \right)^{1/(l-u+1)}$$

where  $u$  and  $l$  are the upper and lower spectrum line boundaries of the critical band, respectively.

# MPEG-I – An Example

## Step 3: Decimation and reorganization of maskers

- This step to reduce the number of maskers (minimizing computation load)
- (a) Keeping only maskers that satisfy  $P_{TM,NM}(k) \geq T_q(k)$ 
  - where  $T_q(k)$  is the SPL of the absolute hearing threshold at spectrum line  $k$ .
- It is approximated by
 
$$T_q(k) = 3,64\left(\frac{k}{1000}\right)^{-0.8} - 6.5e^{-0.6(k/1000-3.3)^2} + 10^{-3}\left(\frac{k}{1000}\right)^4 \text{ (dB)}$$
- It could be naively interpreted as a maximum allowable energy level for coding distortion





# MPEG-I – An Example

## Step 3: Decimation and reorganization of maskers

- (b) A sliding window (of 0.5 Bark-wide) is used to replace any pair of maskers occurring within a distance of 0.5 Bark by the stronger of the two. one stronger masker is used to represent other maskers within 0.5 Bark distance
- (c) The masker frequency bins are reorganized according to the subsampling scheme

$$P_{TM,NM}(i) = P_{TM,NM}(k), \quad P_{TM,NM}(k) = 0$$

where

$$i = \begin{cases} k, & 1 \leq k \leq 48 \\ k + (k \bmod 2) & 49 \leq k \leq 96 \\ k + 3 - ((k - 1) \bmod 4) & 97 \leq k \leq 232 \end{cases}$$

here is low freq. range,  $i=k$  means do not do anything, keep the info.

- The net effect of this step is a 2:1 decimation of masker bins in critical band 18-22 and 4:1 decimation of masker bins in critical band 22-25. means take  $i$  out of 4, other 3 take to 0
- This procedure reduces the total number of tone and noise masker bins to 106. reduce the processing complexity.

# MPEG-I – An Example

- Step 4: Calculation of Individual masking thresholds
  - Individual tone and noise masking thresholds are computed.
  - Each threshold represents a masking contribution at frequency bin  $i$  due to the tone or noise masker at bin  $j$
  - For each  $j$  and all  $i$ , the **tonal** masker thresholds,  $T_{TM}(i,j)$ , is given by two terms with minus sign, i.e.,  $-2.75Z_b(j)-6.025$ , which are the off-set values for the threshold.

$$T_{TM}(i, j) = P_{TM}(j) - 2.75Z_b(j) + SF(i, j) - 6.025(\text{dB})$$

where  $P_{TM}(j)$  is the SPL of the tonal masker at frequency bin  $j$ ,  $Z_b(j)$  is the bark frequency of bin  $j$ , and the spread of masking from masker  $j$  to maskee bin  $i$  is given by

$$SF(i, j) = \begin{cases} 17\Delta_{Z_b} - 0.4P_{TM}(j) + 11 & -3 \leq \Delta_{Z_b} < -1 \\ (0.4P_{TM}(j) + 6)\Delta_{Z_b} & -1 \leq \Delta_{Z_b} < 0 \\ -17\Delta_{Z_b} & 0 \leq \Delta_{Z_b} < 1 \\ (0.15P_{TM}(j) - 17)\Delta_{Z_b} - 0.15P_{TM}(j) & 1 \leq \Delta_{Z_b} < 8 \end{cases}$$

where the Bark maskee-masker separation is  $\Delta_{Z_b} = z_b(i) - z_b(j)$

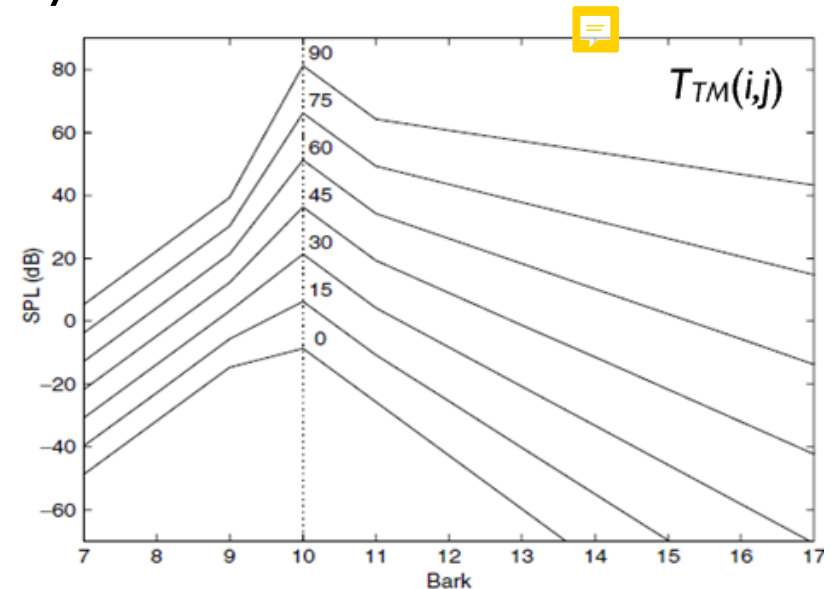
# MPEG-I – An Example

## Observations on $T_{TM}(i,j)$

- The Figure shows  $T_{TM}(i,j)$  at  $Z_b=10$  Barks with different levels of the masker
- The slope decreases with the increasing masker level to reflect the psychophysical nature of our hearing system, i.e., the frequency sensitivity selectivity decreases as stimulus levels increase.
- The spread of the masking is constrained to 10 Barks neighborhood to reduce computational complexity

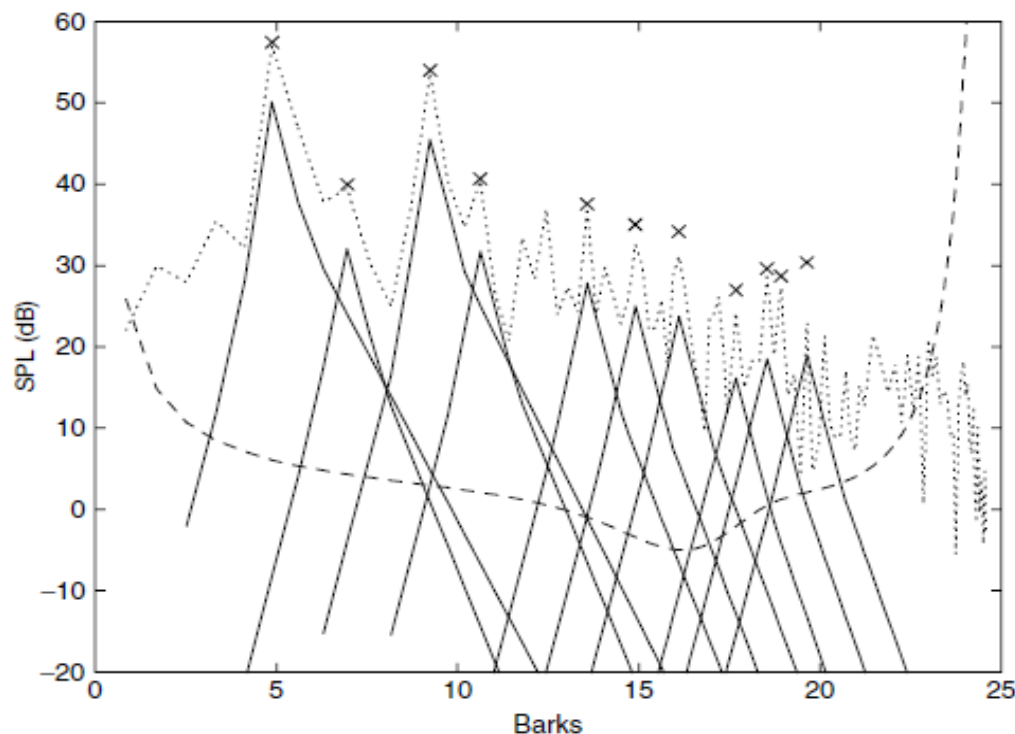
$$SF(i, j) =$$

$$\left\{ \begin{array}{ll} 17\Delta_{Z_b} - 0.4P_{TM}(j) + 11 & -3 \leq \Delta_{Z_b} < -1 \\ (0.4P_{TM}(j) + 6)\Delta_{Z_b} & -1 \leq \Delta_{Z_b} < 0 \\ -17\Delta_{Z_b} & 0 \leq \Delta_{Z_b} < 1 \\ (0.15P_{TM}(j) - 17)\Delta_{Z_b} - 0.15P_{TM}(j) & 1 \leq \Delta_{Z_b} < 8 \end{array} \right.$$



# MPEG-I – An Example

- The figure shows the individual masking thresholds associated with the tonal maskers shown previously
- The short-dashed line shows the sum of the effects (solid line) from the tonal maskers and the absolute hearing threshold (long-dashed line)



# MPEG-I – An Example

- Similarly, the noise masker thresholds,  $T_{NM}(i,j)$ , is given by

$$T_{NM}(i, j) = P_{NM}(j) - 0.175Z_b(j) + SF(i, j) - 2.025(\text{dB})$$

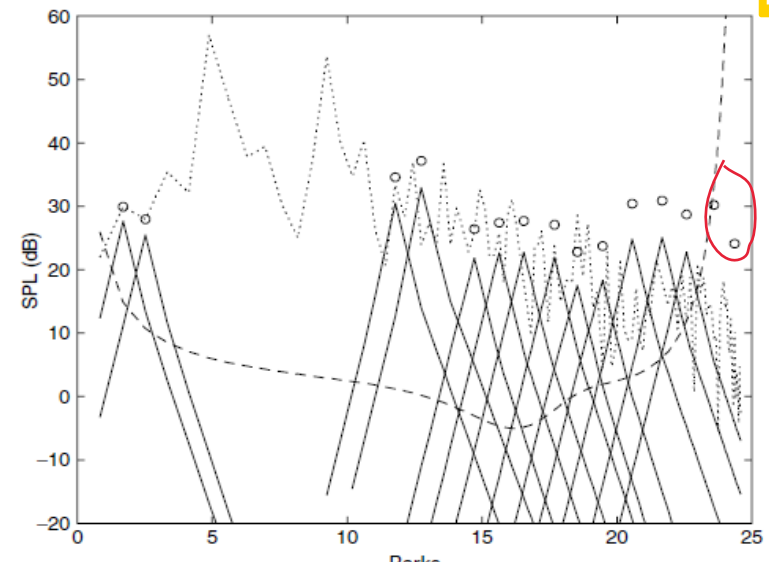
where  $P_{NM}(j)$  is the SPL of the noise masker at frequency bin  $j$ ,  $z_b(j)$  is the bark frequency of bin  $j$ , and  $SF(i,j)$  is obtained by replacing  $P_{TM}(j)$  with  $P_{NM}(j)$  in the definition of  $SF(i,j)$  given previously.

two negative terms  $-0.175Z_b(j) - 0.2025$

- The offset  $0.175Z_b(j) + 2.025$  for noise masking is smaller than  $0.275Z_b(j) + 6.025$  for tonal masking. Therefore, noise masker has more masking effects than the tonal masker.

- The figure shows the individual masking thresholds associated with the noise maskers identified previously

for the same magnitudes of noise and tone maskers, the noise masker produces more masking effects.



# MPEG-1 – An Example

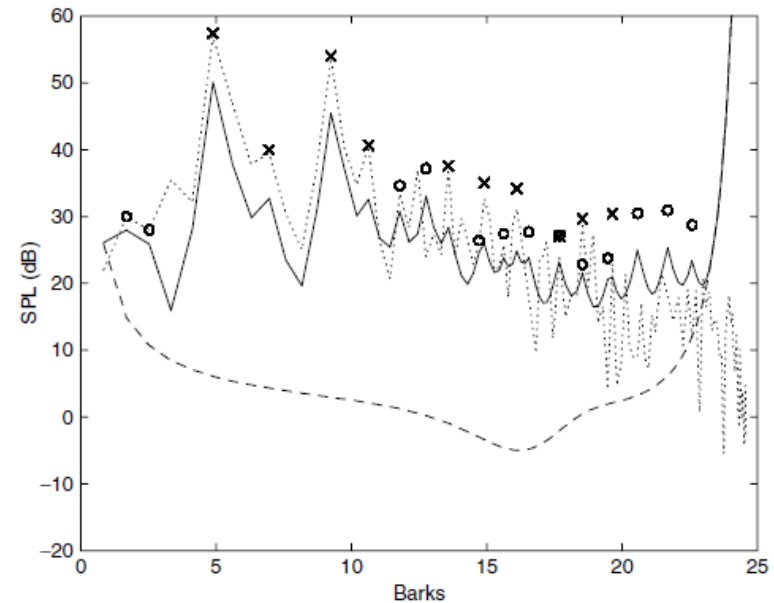
## STEP 5: Calculation of Global Masking Thresholds

- The **global masking threshold** is obtained by combining individual masking thresholds with

$$T_g(i) = 10 \log_{10} (10^{0.1T_q(i)} + \sum_{l=1}^L 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^M 10^{0.1T_{NM}(i,m)}) \text{ (dB)}$$

where  $T_q(i)$  is the hearing threshold for bin  $i$ ,  $T_{TM}(i,l)$  and  $T_{NM}(i,m)$  are the individual masking thresholds from Step 4, and  $L$  and  $M$  are the no. of tonal and noise maskers, respectively, identified in Step 3.

- Note that the two high frequency noise maskers that occur below the absolute threshold have been eliminated in the above Figure.
- Temporal masking should be also included (not discussed here).

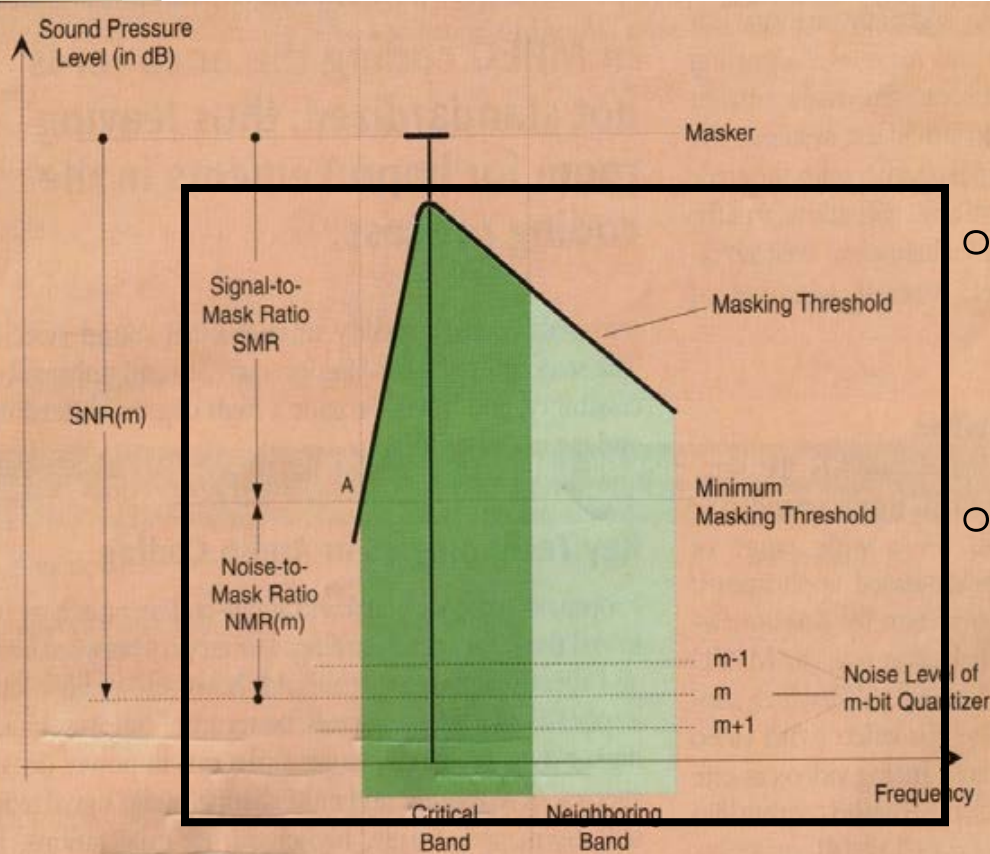


# Bit Allocation and Frame Format

The bit allocation process determines the number of code bits allocated to each subband based on information from the psychoacoustic model



# Perceptual Bit Allocation



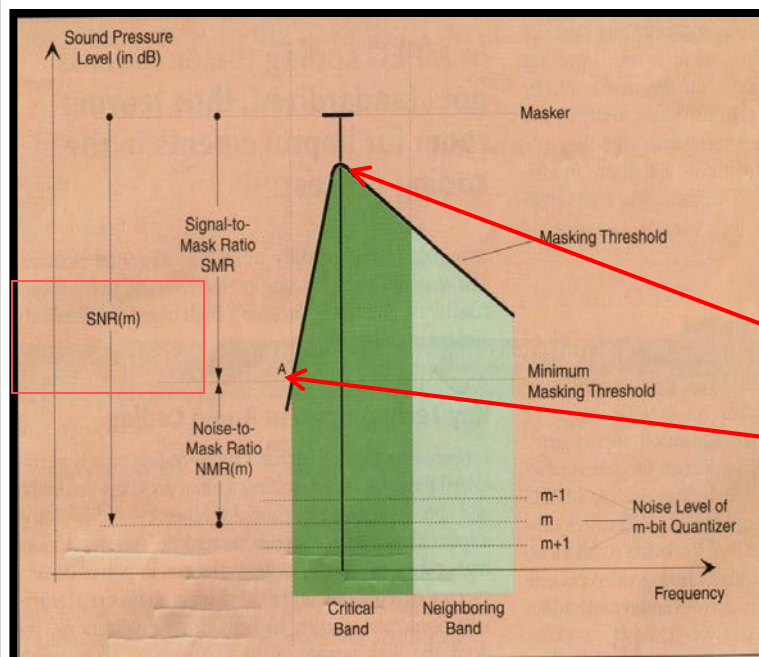
- This diagram defines many important parameters specifying the masking curve within the critical band (CB).
- The masker located at the center of the CB creates a masking threshold to mask signals falling below this threshold.

- The dark green shade represents the area whereby the signals are to be masked (within the same CB).
- It also shows the quantization noise level corresponding  $m$  bits sample representation.
- The SNR of the masker is defined as the no. of bits used to quantize the signal; 6 dB with each bit, e.g., SNR = 90 dB for 15 bits.

the figure relates the number of bits used for quantization and the signal quality.



# Perceptual Bit Allocation

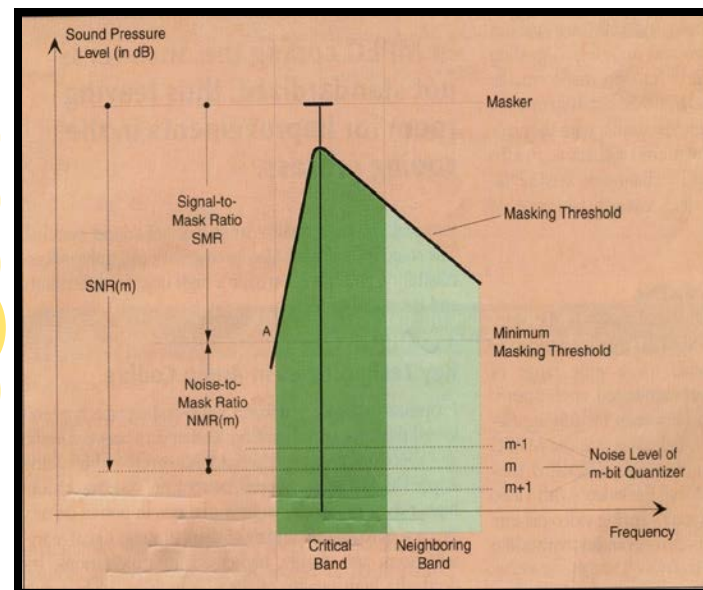


- The slope of the masking threshold is steeper toward lower freq, than higher freq.
- $SMR$  = dist. between level of masker and masking threshold.
- Min  $SMR$  = freq of masker, (e.g. 6dB from masker.)
- Max  $SMR$  = point A towards the lower freq
- Distant between masker and masking threshold is smaller in noise-masking tone than in tone-masking noise.

- If  $SNR > SMR$ , quantization noise will not be audible, Noise-to-mask ratio,  $NMR(m) = SNR(m) - SMR$  dB
- It means that the number of bits,  $m$ , should be large enough so that coding noise is not audible
- The  $NMR(m)$  is an important parameter to determine the allocated number of bits to a CB.

# Perceptual Bit Allocation

- Distance between masker and masking threshold is generally smaller in noise-masking-tone experiments than tone-masking-noise experiments. Therefore, noise is a better masker than a tone.

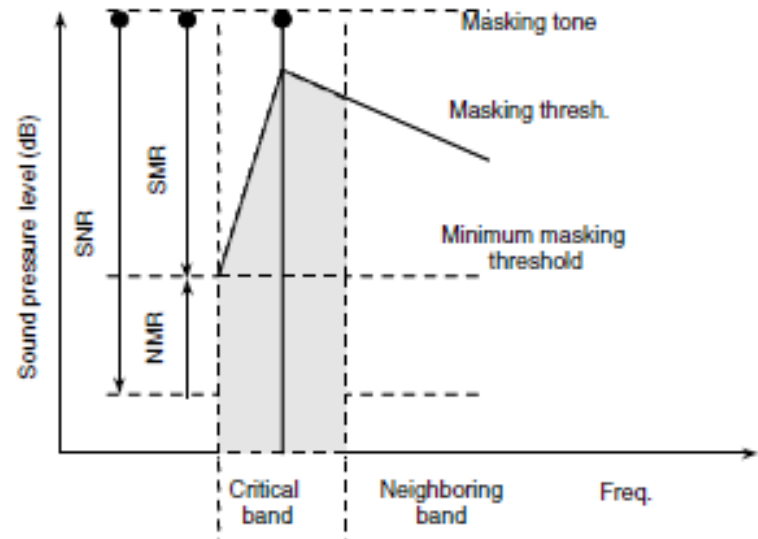


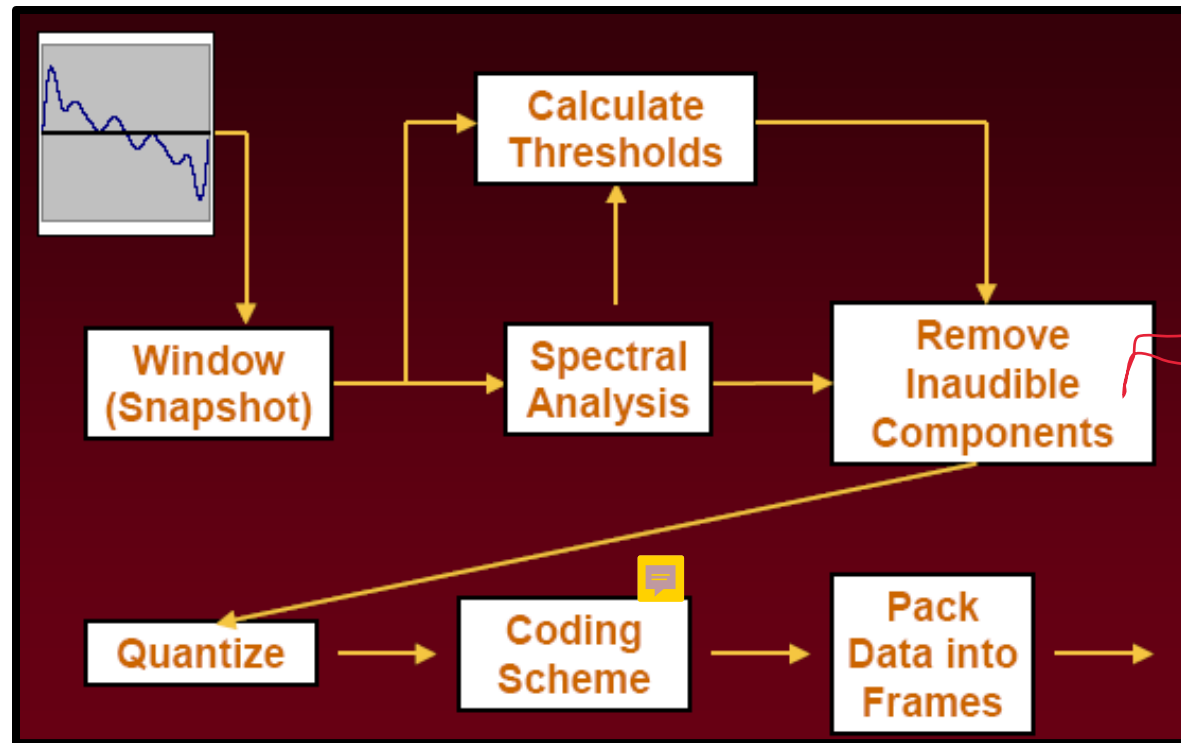
- Music and speech signals consist of many maskers, being tonal and noise-like. A global masking threshold should be obtained from all these maskers.
- When the SPL level of the masking signal increases, the upper frequency slope of the masking curve becomes more shallower and masks a wider range of upper frequencies.

# Perceptual Bit Allocation

- In perceptual bit allocation method, the no. of bits allocated to different bands is determined based on the global masking thresholds obtained from the psychoacoustic model.
- The signal-to-mask ratio (SMR) determines the number of bits to be assigned in each band for perceptually transparent coding of the input audio.
- The noise-to-mask ratios (NMRs) are computed by subtracting the SMR from the SNR in each subband, i.e.,  

$$\text{NMR} = \text{SNR} - \text{SMR} \text{ (dB)}$$
- The main objective in a perceptual bit allocation scheme is to keep the quantization noise below a masking threshold, i.e.,  $\text{NMR} > 0$ .
- If the minimum masking threshold is 40 dB and the SPL of tone masker is 68 dB. How many bits should be used at least?





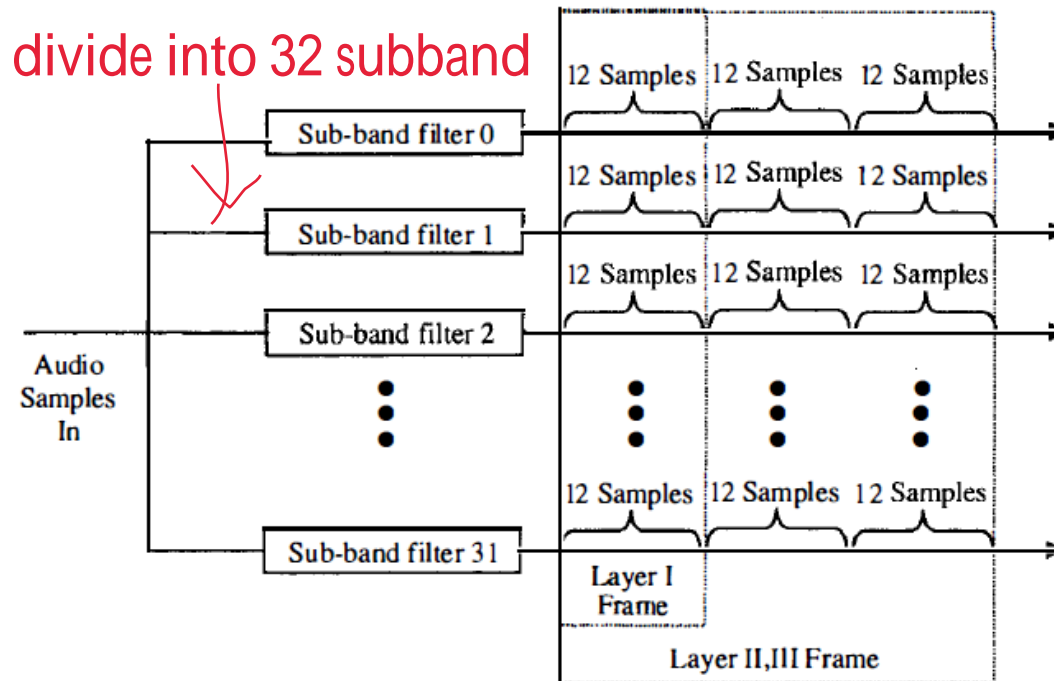
- Considering practical encoder based on the knowledge learned previously
- With the global masking threshold, we shall remove the inaudible components
- Adding additional information that is necessary for the decoding process
- We shall look at some practical MPEG coding standards

# ISO/MPEG-1 Audio Coding : Comparison

MPEG-I Layer I	MPEG-I Layer II	MPEG-I Layer III
Analysis filterbank (PQMF) Freq. resolution = 750Hz	Analysis filterbank (PQMF) Freq. resolution = 750Hz	Hybrid filterbank (Subband (PQMF) filter+ 18-pt MDCT) <b>better freq. resolution</b> Freq. resolution = 42 Hz
Each block has 12 subband samples	Each block has 36 subband samples	Each block has 36 subband samples
Least complex, delay and poorest quality	Moderate complex, delay and good quality	More complex, delay and better quality
L+R mode in intensity coding	L+R mode in intensity coding	Intensity coding + additional option of M/S mode
FFT =512 point in psychoacoustic model	FFT =1024 point in psychoacoustic model	FFT =1024 point in psychoacoustic model
Did not exploit redundancy between scalefactors	Exploit redundancy between scalefactors	Exploit redundancy between scalefactors
No Huffman coding	No Huffman coding	Huffman Coding
Bit reservoir buffer technique	Bit reservoir buffer technique	Bit reservoir buffer technique

# MPEG-I, II, III

- Let us overview the MPEG-I coding details.



Data frame structure for MPEG-1 Audio Layers I, II, and III

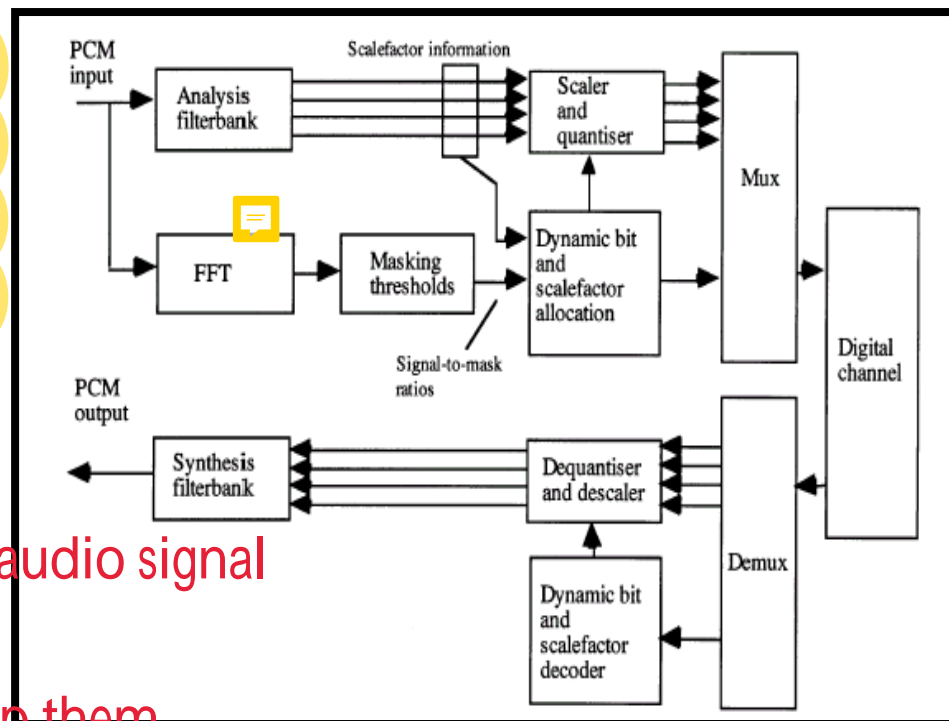
- The figure shows the arrangement for MPEG-I spectrum analysis:
    - Input data are divided into 32 frequency sub-bands in a frame
    - For layer I, each sub-band contains 12 samples, i.e.,  $12 \times 32 = 384$  samples
    - For layers II and III, each sub-band has 36 samples, i.e.,  $36 \times 32 = 1152$  samples
- sample number increase

# MPEG-1 Layer I, II audio coding

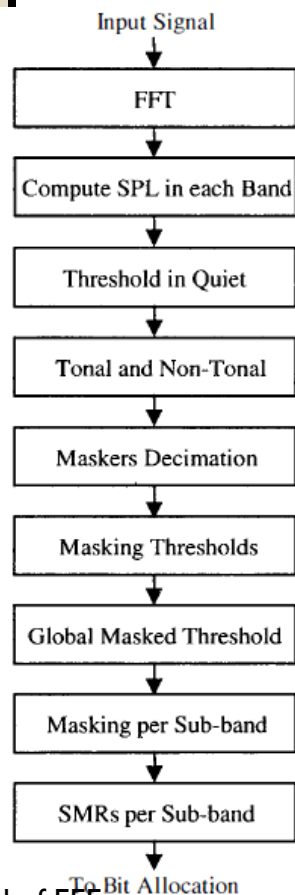
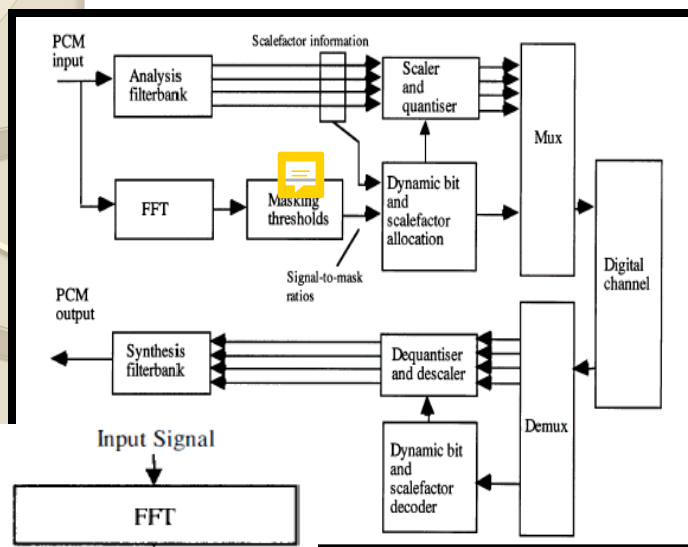
- **Analysis Filterbank** : splits the audio signal into 32 sub-bands. Filters are linearly spaced, each having a BW=750Hz, i.e.,  $750 \times 32 = 24 \text{ kHz}$ .
- Samples in each band are critically decimated, and split into blocks of 12 decimated samples.

- **Scale factors** are calculated to normalize the amplitude of the samples in each band.

reduce number of bits  
some important detail of audio signal  
come from small amp.  
need to scale it bigger to keep them



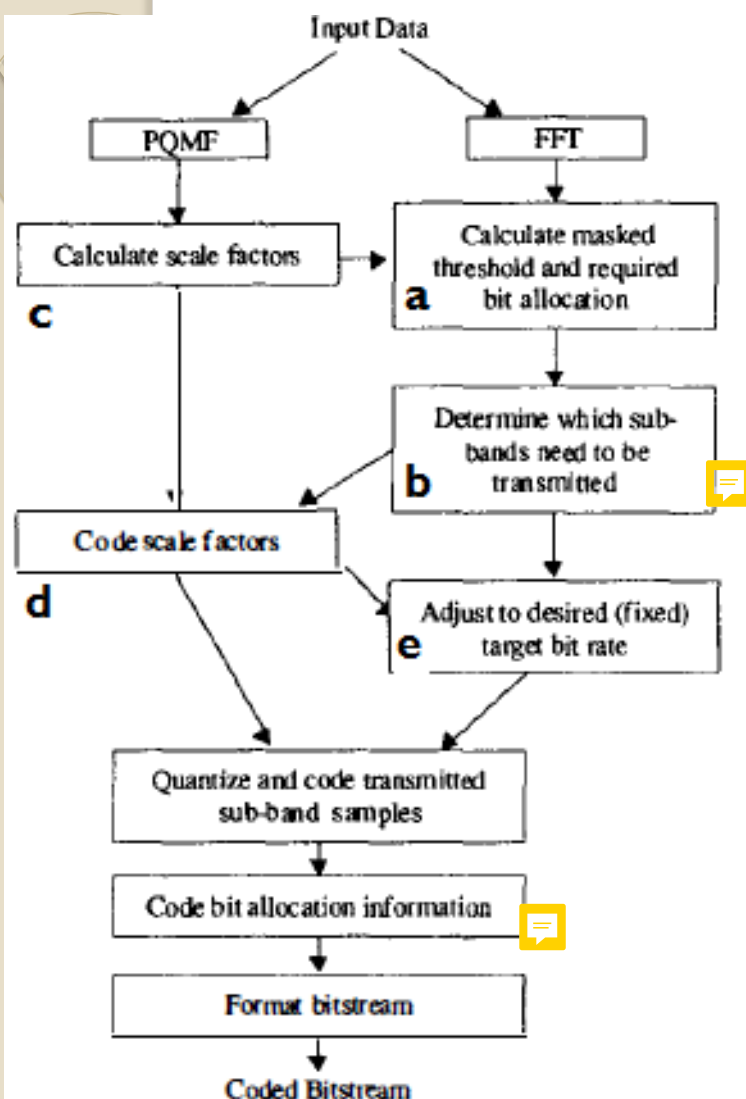




- A **Psychoacoustic model I** calculates the masked threshold from the spectrum of the current block. SMR for each band is computed.
- Signal is channeled to the **512-point FFT(layer I)** or **1024-point FFT(layer II)** to calculate the spectrum of the current audio block.
- Finally, the quantized samples, scale factors, and control info are multiplexed together for transmission or storage.



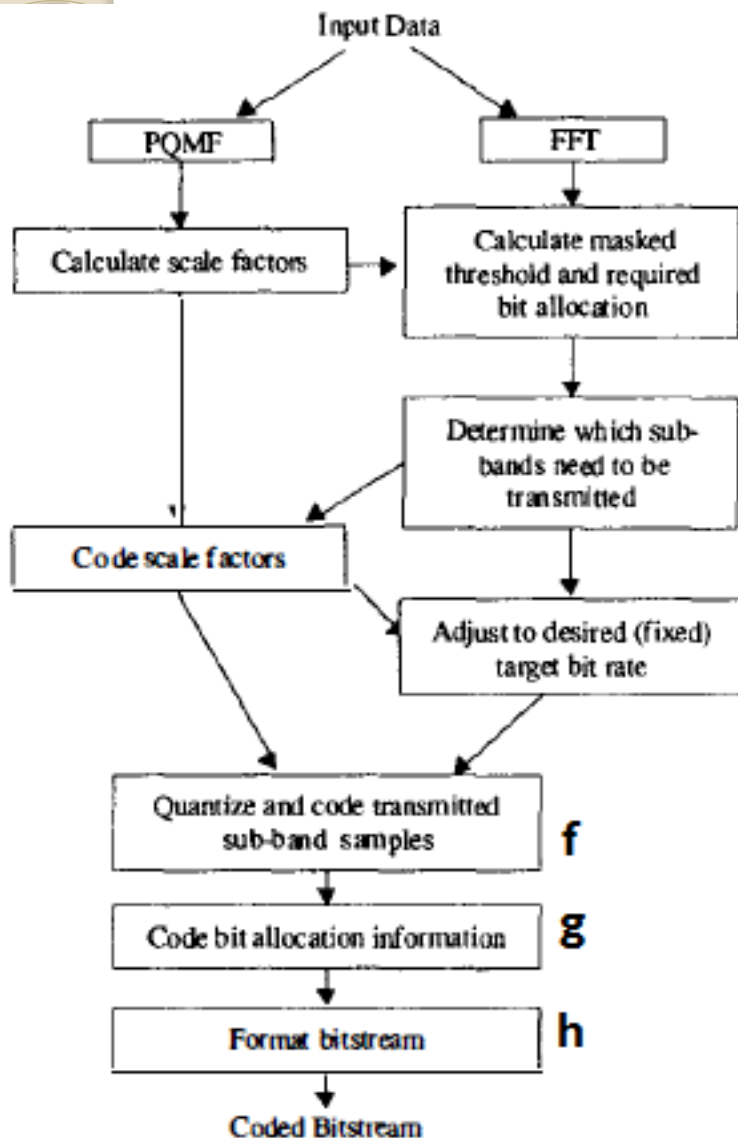
# MPEG-I Audio Syntax-Bit Allocation



Consider the MPEG-I layers I and II.

- Bit allocation is performed according to the global masking threshold
- Decide which subband is to be used.
- A 6-bit **scale factor** is calculated for each of 32 subbands
  - For layer I, each subband has 12 bins
  - The scale factor is used for maximizing the signal to noise ratio for each subband.
- Side information is also needed to properly represent the scale factor
- Need adjustment for fixed bit rate

# MPEG-I Audio Syntax-Bit Allocation



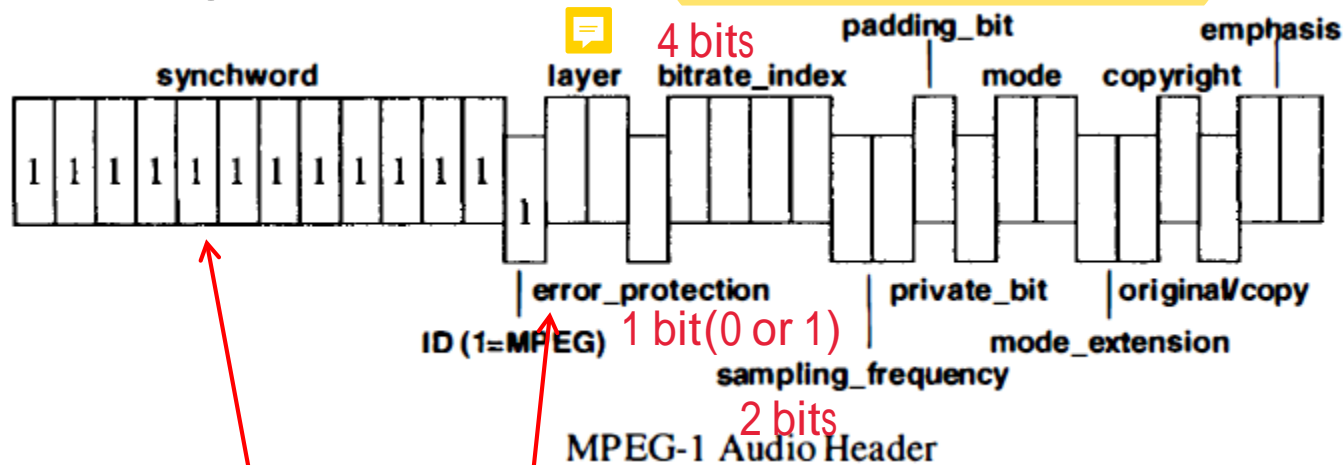
f. Quantization process is performed according to the information of scale factor and the information on the selected bit allocation

g. Additional bit allocation information is added

h. Putting all information into the frame with the defined frame format.

# MPEG-I Audio Syntax

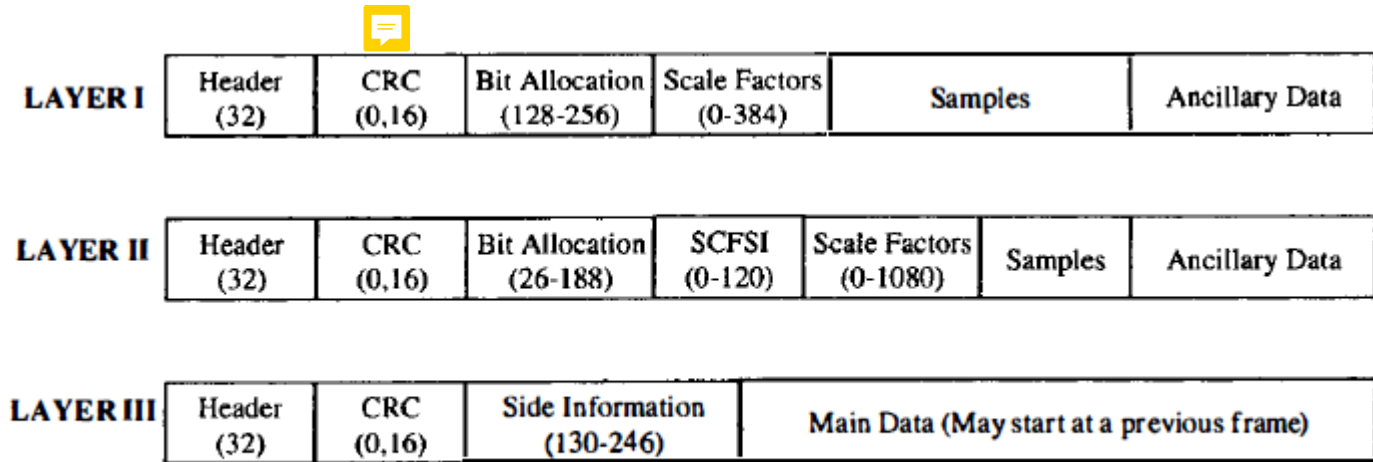
- The figure shows a 32-bit **MPEG audio header** of frame



- It has
  - A 12-bit string of 1 for synchronization 同步
  - A MPEG ID bit (=1 implies an MPEG-I audio stream) and 2-bit layer identification
  - Error protection bit, mode (stereo, joint stereo, dual channel, single channel), and mode extension, etc.

# MPEG-I Audio Syntax

- The Figure shows the <sup>frame structure</sup> frame format of three layers of MPEG-I



MPEG-1 Audio frame format

- It contains 16 bit cyclic redundancy coding for error checking
- Bit allocation information
- Scale factor and scale factor select information
- ..., and ancillary data

# MPEG-I Audio-Scale Factor

- The scale factor is used to modify the quantization step size ensuring the quantization noise level is below the computed masked threshold.
- For MPEG-I layers II, one scaling factor (of 6 bits) is used for every 12 subband samples *layer 2 3 need scale factor*
- The maximum absolute value of the 12 samples is determined and mapped to a scale factor in a lookup table (see below)

*use biggest one to determine scale factor*

- The samples are divided by the scale factor to obtain a dynamic range of 120 dB
- The scale factor is transmitted only if the bit allocation for a band is non-zero.

MPEG Audio Layers I and II Scale Factors [ISO/IEC 11172-3]

Scale Factor Index	Scale Factor Value	Scale Factor Index	Scale Factor Value
0	2.00000000000000	32	0.00123039165029
1	1.58740105196820	33	0.00097656250000
2	1.25992104989487	34	0.00077509816991
3	1.00000000000000	35	0.00061519582514
4	0.79370052598410	36	0.00048828125000
5	0.62996052494744	37	0.00038754908495
6	0.50000000000000	38	0.00030759791257
29	0.00246078330058	61	0.00000151386361
30	0.00195312500000	62	0.00000120155435
31	0.00155019633981		

# Bit Allocation and Quantization

- The number of bits used to quantize sub-band samples in Layer I is between 0 and 15 (excluding the allocation of 1 bit due to the use of midread)
- The number of bits used to transmit **bit allocation information** is equal to  $4 \times 32 = 128$  bits for a single channel
- **Bit allocation routine:** to maintain smooth operation with the desired quality under the limitation of a possible given bit rate/second.
- Assuming a given bit rate:
  - Setting all bit allocation codes to be zero and assume no bits are needed to transmit scale factors
  - In each iteration, an additional bit is allocated to the subband that can still accept an additional bit (i.e., has less than 15 bits already)
  - Repeat until no more bit left for the next iteration.
  - Once bit allocations are determined, each sub-band sample is divided by its scale factor and quantized.
- Both the allocated bits and the associated scale factor are transmitted.

## ○ **Decoder :**

- Unpacks the information
- Scale and interpolate the quantized samples as instructed by the control information.
- Pass the 32 bands through a synthesis filter to generate the PCM samples - 16 bit representation for 96 dB SNR
- No psychoacoustic model is needed (complexity is reduced)
- Decoder standard is defined exactly by the MPEG standard.

## ○ **Encoder :**

- No fixed standard for psychoacoustic model for future expansion
- Just need to yield a valid bitstream

## ○ **MPEG-I layer II**

- Similar to layer I, but achieve higher audio quality at the same bitrate.
- 1024-point FFT (instead of 512) : higher frequency resolution
- Exploit similarity between adjacent scale factors (reduce control info bit)
- More accurate quantizer
- Achieve CD quality around 256 kbps stereo.
- Used in DAB

# Stereo Coding

立体声编码

## 4 modes in MPEG-I :

- Mono (No redundancy)
- Stereo (Coded together)
- Dual (Two separate channels, coded individually)
- Joint Stereo (combine two stereo channels)
- Redundancy found within channels can be removed to reduce bit rate.

sound source may be the same


## Matrix Stereo Coding: MS code

- Exploits the similarity between 2 stereo channels
- Code the sum (or middle) and difference (or side) signals :
- The transformation is lossless and reversible
- For correlated L and R signals, bitrate reduction is good since S is minimal.
- If signals are not correlated, no advantage

$$\begin{aligned}M &= \frac{L + R}{\sqrt{2}} \\S &= \frac{L - R}{\sqrt{2}} \\L &= \frac{M + S}{\sqrt{2}} \\R &= \frac{M - S}{\sqrt{2}}\end{aligned}$$



- MPEG-I layer III can use a combination of stereo techniques. Encoder can switch between independent stereo, matrix stereo, and/or intensity stereo. **Intensity stereo coding in all MPEG layers**

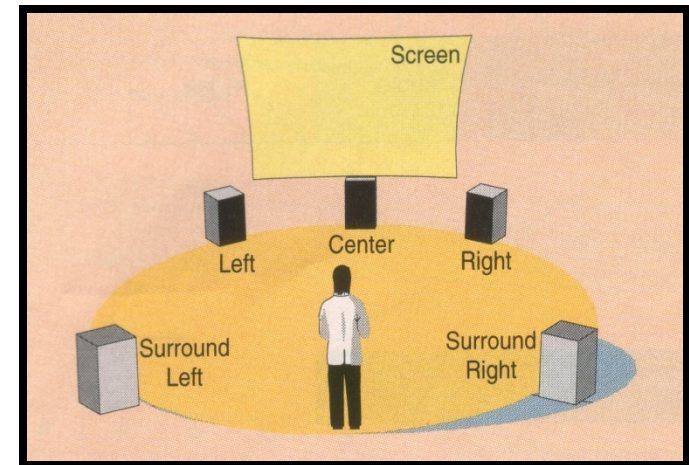
MPEG-I/Audio Coding	Approximate stereo bit rates	Compression ratio
Layer I	384 kbps stereo	4
Layer II	192 kbps joint stereo 	8
Layer III	128 kbps joint stereo (variable bit rate assumed)	12

# Applications for different layers

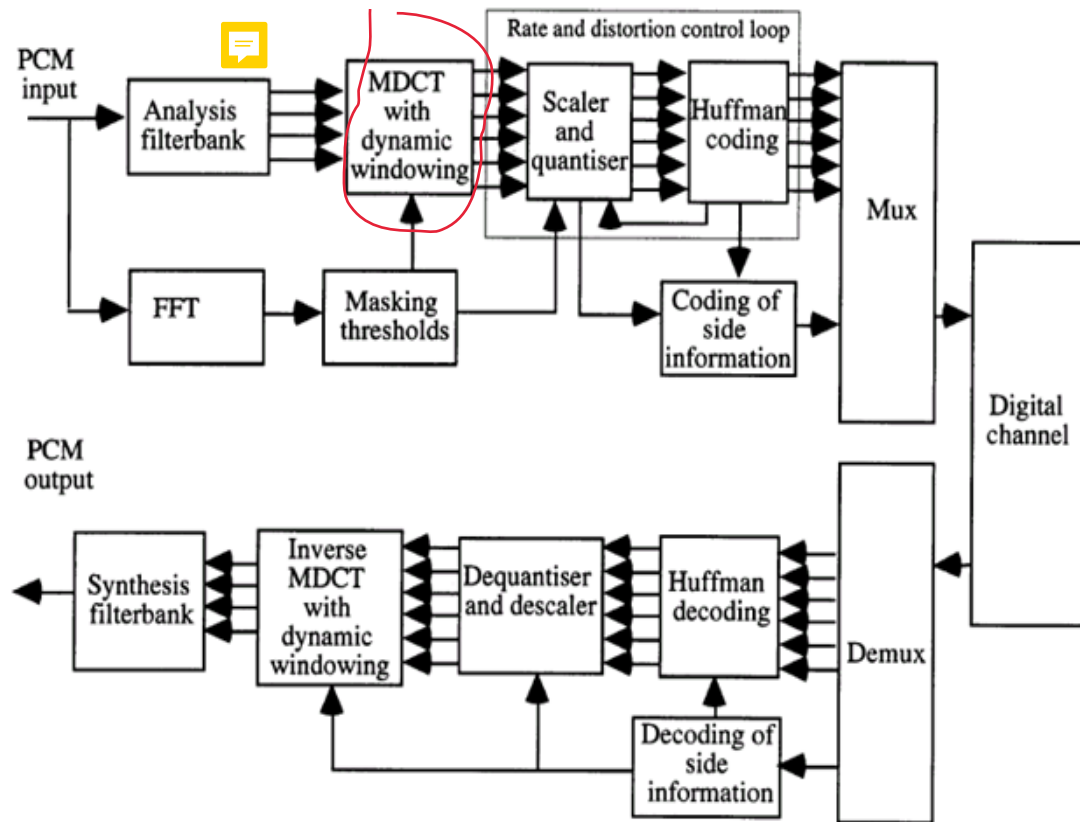
- Layer I: used in digital compact cassette at 192 kbps per channel  
磁带
- Layer II: used in digital audio broadcast (DAB) and DVB at 128 kbps per channel
- Layer III: used in transmission over ISDN and internet at 128 kbps per channel. Also is the well known MP3 file format.

# MPEG Multichannel Audio Coding

- In many applications, a realistic surround-sound field for :
  - Video conferencing
  - Videophony
  - Multimedia services
  - Electronic cinema
- Multichannels also for :
  - Multilingual channels
  - Additional channel for hearing impaired (enhanced intelligent)
- ITU-R 775 recommends 5-channel speaker configuration:
  - 3/2 stereo
  - 3 front (L,R,C)
  - 2 rear (LS and RS)



# MPEG-1 Layer III



MPEG layer 3 block diagram

- **Subband & transform coding**
- Further divide the 32 freq. bands by 6-pt or 18-pt MDCT.
- Finer resolution compared to Layers I and II.
- Allow switching between 2 possible MDCT lengths.
- Short block for encoding transient signal, Longer block for steady-state signals.

- The open standard also enables the use of variable bit-rate and fixed bit-rate coding
    - How many freq points are there at the output of the MDCT?
- There are 1024-pt FFT point used to derive the masking threshold.

# Block Switching in Layer III

- Layer III can switch between a low time resolution (32 x 36 = 1152 time-samples) or a high time resolution (32 x 12 = 384 time-samples).
- During transients, a sequence of 3 short windows replace the long window, thus maintaining the same total number of samples/frame

## Other differences in Layer III

- Use of **psychoacoustic model 2**
- Quantized samples are losslessly packed using **Huffman coding** variable length coding, used to further reduce the compression.
- It uses 16 different Huffman tables to best match to the signal statistics

# Constant Bitrate vs. Variable Bitrate

- Constant (or fixed) bitrate (CBR) 128 kbps encoding uses the same bitrate regardless of the dynamic of the audio signal. For example, the same number of bits are used in a frame whether audio is present or not.
- Variable bitrate (VBR) encoding varies the bitrate in accordance with the dynamic of the audio in the frame.
- MP3 files encoded with VBR may not play on VCD/DVD players.
- VBR file is often smaller than those encoded in constant bitrate, indicating that the VBR is more efficient. better quality

No. MP3 decoder must read the header regardless of variable bitrate (BR) or constant BR.  
Encoders and decoders are typically more complex.

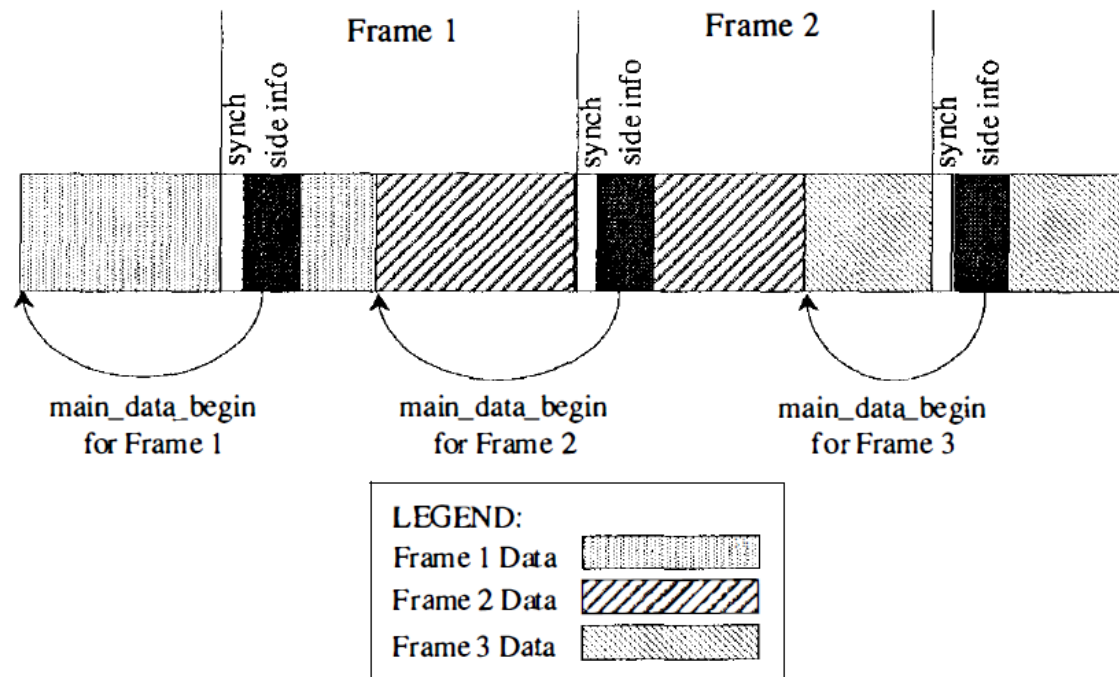
Yes. When encoding the same file using VBR and CBR will result in different sized files. better quality-to-space ratio compared to an equivalent CBR.

# Bit Reservoir

- We hope the transmission of audio to be at a constant rate
- Allowing for donating bits to the bit reservoir when the analyzed signal requires less than average number of bits. These bits can be taken from the reservoir for peak demand
- This coding scheme supports **locally-variable bit rate** operation, and obtains **a constant bitrate operation in the long run.**
- The maximum channel capacity is never exceeded.
- The maximum deviation from the target is fixed by the size of the maximum allowed delay in the decoder, i.e., 7680 bits, which corresponds to the data rate of 320 kb/s at 48 kHz sampling rate.
- The data reservation is only possible for applications that require a smaller data rate.

# Bit Reservoir

- The first element in Layer III audio data stream is a 9-bit pointer to provide the starting location the data frame
- The Figure shows the arrangement that the current frame data is started from previous frame by using the pointer mechanism.





# Overview MPEG Standards

# Before MPEG

- 1963, Phillips introduced the audio **cassette tape format** that eventually became popular among home audio enthusiasts.
- 1978, Sony developed the first **digital audio recording** devices to be used by professional studios
- 1979, Sony produced first walkman portable **audio cassette player**
- 1988, first **compact discs**

**File sharing and media distribution** becomes important

- 1987 - The Fraunhofer Institut in Germany began research code-named EUREKA project EU147, Digital Audio Broadcasting (DAB).
- January 1988 - Moving Picture Experts Group or MPEG was established as a subcommittee of the International Standards Organization/International Electrotechnical Commission or ISO/IEC.
- April 1989 - Fraunhofer received a German patent for MP3.
- 1992 - Fraunhofer's and Dieter Seitzer's audio coding algorithm was integrated into MPEG-I.

# The Beginning of MPEG

- 1993 - MPEG-1 standard published.
- 1994 - MPEG-2 developed and published a year later.
- November 26, 1996 - United States patent issued for MP3.
- September 1998 - Fraunhofer started to enforce their patent rights. All developers of MP3 encoders or rippers and decoders/players now have to pay a licensing fee to Fraunhofer.
- February 1999 - A record company called SubPop is the first to distribute music tracks in the MP3 format.
- 1999 - Portable MP3 players appear.

# MPEG Standards

- MPEG is the acronym for *Moving Pictures Experts Group* that forms a workgroup (WG-11) of ISO/IEC JTC-1 subcommittee (SC-29).
- The main functions of MPEG are:
  - a) to publish technical results and reports related to audio/video compression techniques;
  - b) to define means to multiplex (combine) video, audio, and information bit streams into a single bit stream, and
  - c) to provide descriptions and syntax for low bit rate audio/video coding tools for Internet and bandwidth restricted communications applications.
- MPEG standards do not characterize or provide any rigid encoder specifications, but rather standardizes the type of information that an encoder has to produce as well as the way in which the decoder has to decompress this information.

## An overview of the MPEG audio standards.

Standard	Standardization details	Bit rates (kb/s)	Sampling rates (kHz)	Channels	Related information
MPEG-1	ISO/IEC 11172-3 1992	32–448 (Layer I) 32–384 (Layer II) 32–320 (Layer III)	32, 44.1, 48	<u>Mono (1), stereo (2)</u>	A <u>generic compression</u> standard that targets primarily multimedia storage and retrieval.
MPEG-2 BC/LSF	ISO/IEC 13818-3 1994	32–256 (Layer I) 8–160 (Layers II, III)	16, 22.05, 24	<u>Multichannel Surround sound (5.1)</u>	First digital television standard that enables lower frequencies and multichannel audio coding.
MPEG-2 NBC/AAC	ISO/IEC 13818-7 1997	8–160	8–96	Multichannel	Advanced audio coding scheme that incorporates new coding methodologies (e.g., prediction, noise shaping, etc.).
MPEG-4 (Version 1)	ISO/IEC 14496-3 Oct, 1998	0.2–384	8–96	Multichannel	The <u>first content-based</u> multimedia standard, allowing <u>universal-ity/interactivity</u> and a combination of natural and synthetic material, coded in the form of <u>objects</u> .
MPEG-4 (Version 2)	ISO/IEC 14496-3/AMD-1, Dec, 1999	0.2–384 (finer levels of increment possible)	8–96	Multichannel	
MPEG-7	ISO/IEC 15938-4 Sept, 2001	–	–	–	A normative metadata standard that provides a <u>Multimedia Content Description Interface</u> .
MPEG-21	ISO/IEC-21000	–	–	–	A multimedia framework that provides interoperability in content-access and distribution.

# MPEG Standards

- The MPEG-I audio standard (ISO/IEC 11172-3) [ISO192] comprises a flexible hybrid coding technique that incorporates several methods including subband decomposition, filter-bank analysis, transform coding, entropy coding, dynamic bit allocation, nonuniform quantization, adaptive segmentation, and psychoacoustic analysis.
- MPEG-I audio codec operates on 16-bit PCM input data at sample rates of 32, 44.1, and 48 kHz.
- MPEG-I offers separate modes for mono, stereo, dual independent mono, and joint stereo.
- Available bit rates are 32–192 kb/s for mono and 64–384 kb/s for stereo.  
from 32 to 192
- The MPEG-I architecture contains three layers of increasing complexity, delay, and output quality. Each higher layer incorporates functional blocks from the lower layers.

layer 3 has most delay



# MPEG Standards

- For layer I encoding, decimated subband sequences are quantized and transmitted to the receiver in conjunction with side information, including quantized scale factors and quantizer selections.
- Layer II improves three portions of layer I in order to realize enhanced output quality and reduce bit rates at the expense of greater complexity and increased delay.
  - First, the layer II perceptual model relies upon a higher-resolution FFT (1024 points) than does layer I (512 points).
  - Second, the maximum subband quantizer resolution is increased from 15 to 16 bits. Also, a lower overall bit rate is achieved by decreasing the number of available quantizers with increasing subband index.
  - Finally, scale factor side information is reduced while exploiting temporal masking by considering properties of three adjacent 12-sample blocks and optionally transmitting one, two, or three scale factors plus a 2-bit side parameter to indicate the scale factor mode.
- Average mean opinion scores (MOS) of 4.7 and 4.8 were reported for monaural layer I and layer II codecs operating at 192 and 128 kb/s, respectively.

# MPEG Standards

- The layer III MPEG architecture achieves performance improvements by adding several important mechanisms on top of the layer I/II foundation.
- The MPEG layer-III algorithm operates on consecutive frames of data. Each frame consists of 1152 audio samples and is further split into two subframes of 576 samples each. At the decoder, every subframe can be decoded independently.
- A hybrid filter bank is introduced to increase frequency resolution and thereby better approximate critical band behaviour.
  - The hybrid filter bank includes adaptive segmentation to improve pre-echo control. Sophisticated bit allocation and quantization strategies that rely upon nonuniform quantization, analysis-by-synthesis, and entropy coding are introduced to allow reduced bit rates and improved quality.
  - The hybrid filter bank is constructed by following each subband filter with an adaptive MDCT. This allows for higher-frequency resolution and pre-echo control. Use of an 18-point MDCT, for example, improves frequency resolution to 41.67 Hz per spectral line.

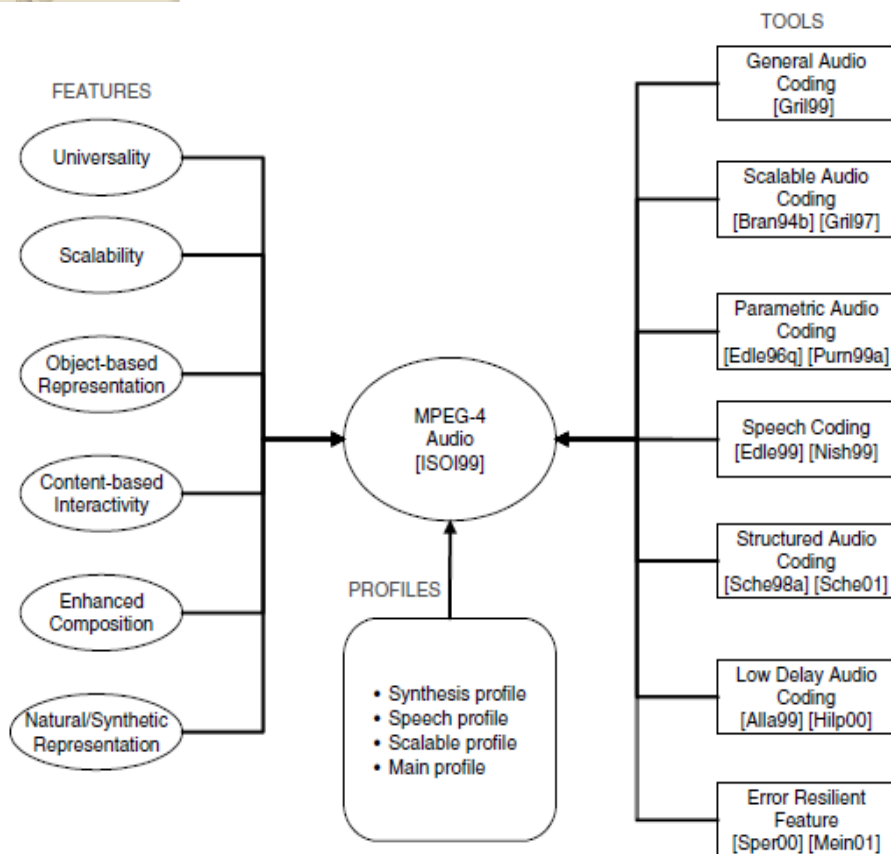
remove echo




# MPEG Standards

- The adaptive MDCT switches between 6 and 18 points to allow improved pre-echo control.
  - shorter blocks (4 ms) provide for temporal pre-masking of pre-echoes during transients;
  - longer blocks during steady-state periods improve coding gain by reducing side information and hence bit rates.
- Bit allocation and quantization of the spectral lines are realized in a nested loop procedure that uses both nonuniform quantization and Huffman coding. The inner loop adjusts the nonuniform quantizer step sizes for each block until the number of bits required to encode the transform components falls within the bit budget.
- The outer loop evaluates the quality of the coded signal (analysis-by-synthesis) in terms of quantization noise relative to the just noticeable distortion (JND) thresholds. Average mean opinion score (MOS) of 3.1 and 3.7 were reported for monaural layer II and layer III codecs operating at 64 kb/s.

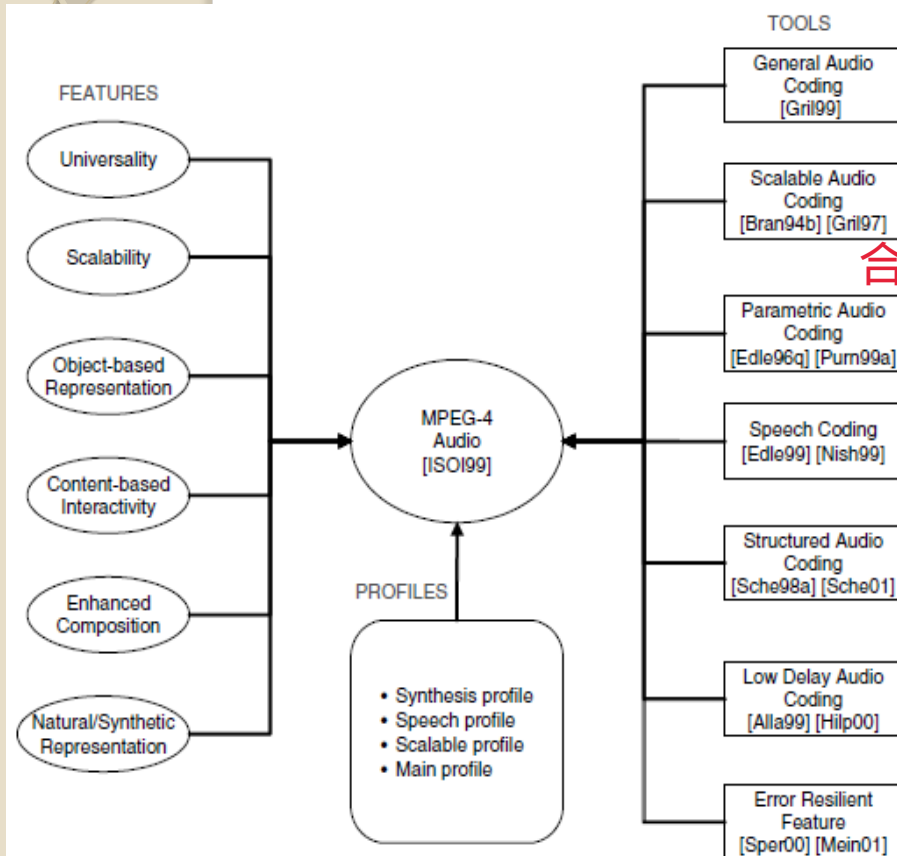
# MPEG-4 Standards



An overview of the MPEG-4 audio coder

- MPEG-4 (1998) comprises an integrated family of algorithms with wide-ranging provision of tools than just perceptive coding
- For **scalable, object-based** speech and audio coding at bit rates from **200 b/s up to 60 kb/s** per channel. 
- Because it is **object-based representation**, possible to achieve user **interactivity/object manipulation** with a **set of coding tools** to accommodate trade-offs between bit rates, complexity and quality.
- **Scalability** allows the creation and manipulation of data streams that can be decoded at varying bit rates to support Internet applications with various environments.

# MPEG-4 – supporting a wider range of services

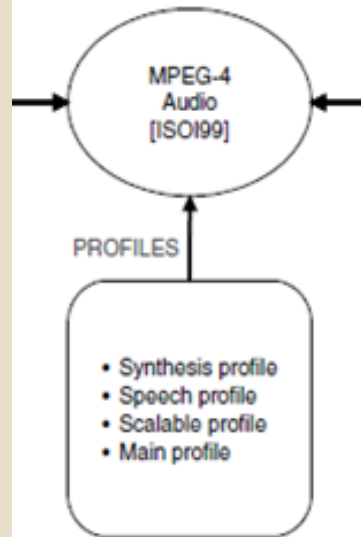


An overview of the MPEG-4 audio coder

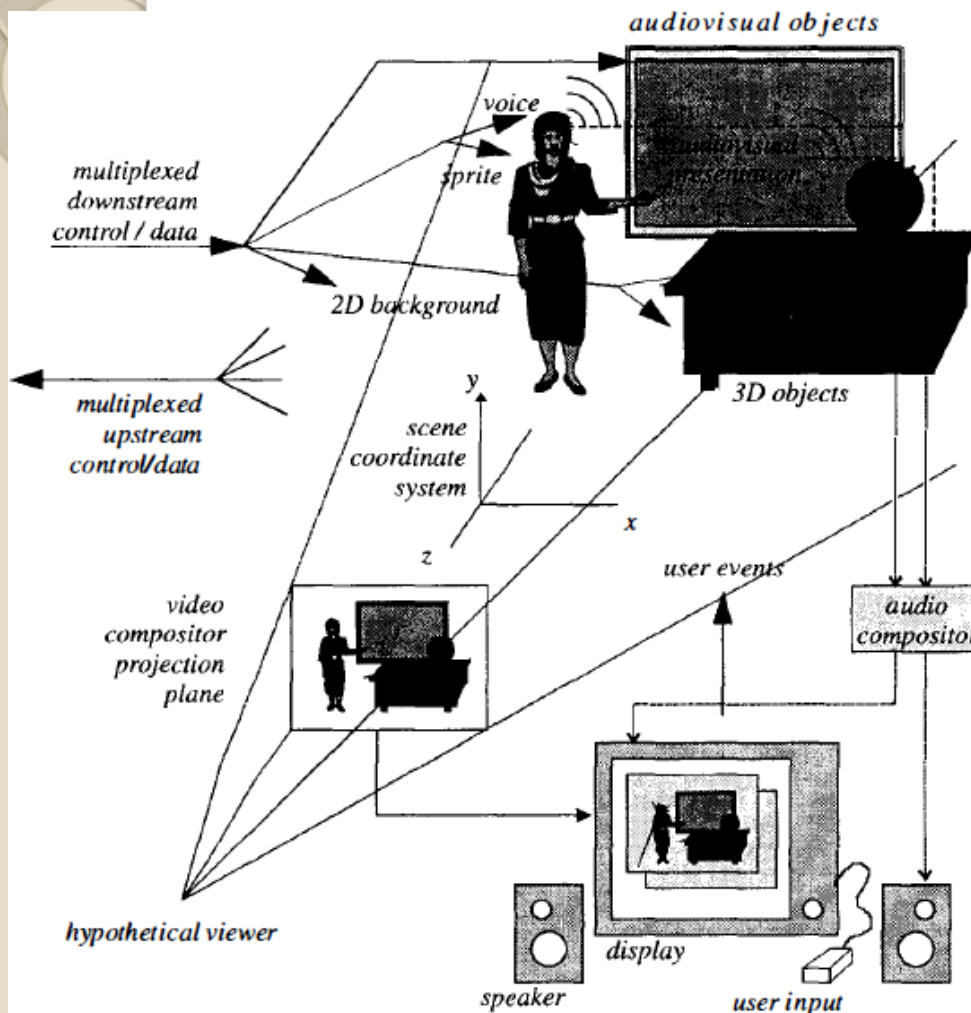
- Efficient and flexible coding of different content (objects), such as natural audio/speech and synthetic audio/speech, for some multimedia applications.
- **MPEG-4 audio provides coding and *合成的* composition of natural and synthetic audio/speech content at various bit rates.**
  - Very low rates are achieved through the use of structured representations for synthetic speech and music, such as text-to-speech and music instrument digital interface (MIDI).
  - For higher bit rates and “natural audio” speech and music, the standard provides integrated coding tools making use of different signal models, depending on desired bit rate, bandwidth, complexity, and quality.
- Coding tools are also specified in terms of MPEG-4 “profiles” that essentially recommend tool sets for a given level of functionality and complexity.

# MPEG-4 – supporting a wider range of services

- In order of bit rate, the profiles are as follows.
  - The **low rate synthesis audio profile** provides only wave table based synthesis and a text-to-speech (TTS) interface.
  - For natural audio processing capabilities, the **speech audio profile** provides a very-low-rate speech coder and a CELP speech coder.
  - The **scalable audio profile** offers a superset of the first two profiles.
    - With bit rates ranging from 6 to 24 kb/s and bandwidths from 3.5 to 9 kHz, this profile is suitable for scalable coding of speech, music, and synthetic music for applications such as Internet streaming or narrow-band audio digital broadcasting (NADIB).
  - The **main audio profile** is a superset of all other profiles, and it contains tools for both natural and synthetic audio.



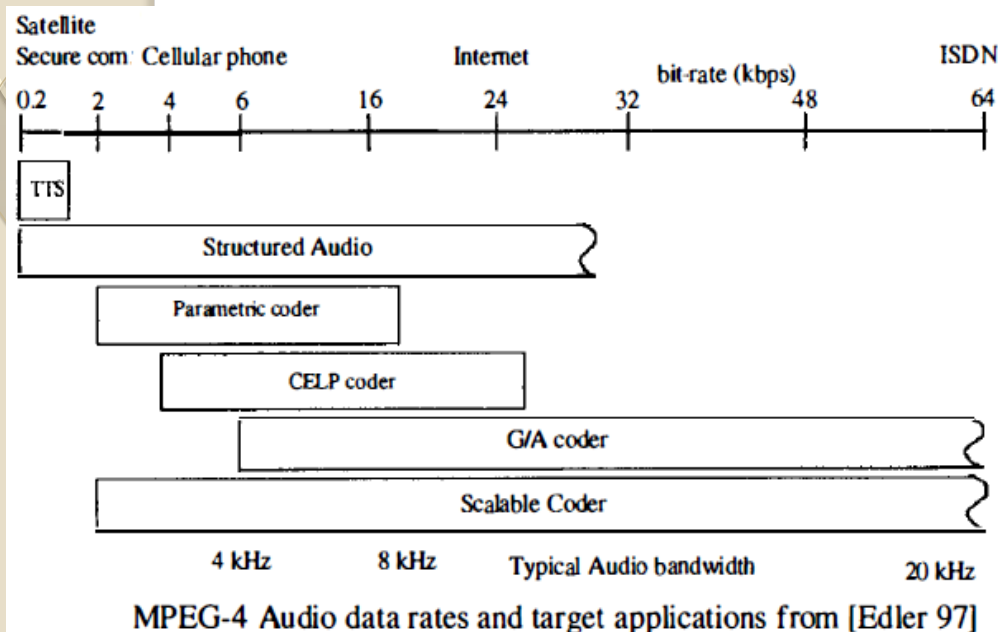
# MPEG-4 – a wider range of tools based on Object Oriented Concept



An example of an MPEG-4 scene from [ISO/IEC MPEG N4668]

- MPEG-4 standard describes the composition of objects to create groups of media objects that collectively describes the audiovisual scene, for example
  - A talking person without background (video object)
  - The person's voice plus background noise (audio object)
  - Allow the creation of objects for describing more complex audiovisual scenes.
  - The background image (still image object)
- Facilities for multiplexing and synchronization of the data associated with media objects for transportation over media channels.
- Tools for interaction with audiovisual scenes at the receiver's end
- Also support controlled access for intellectual property management

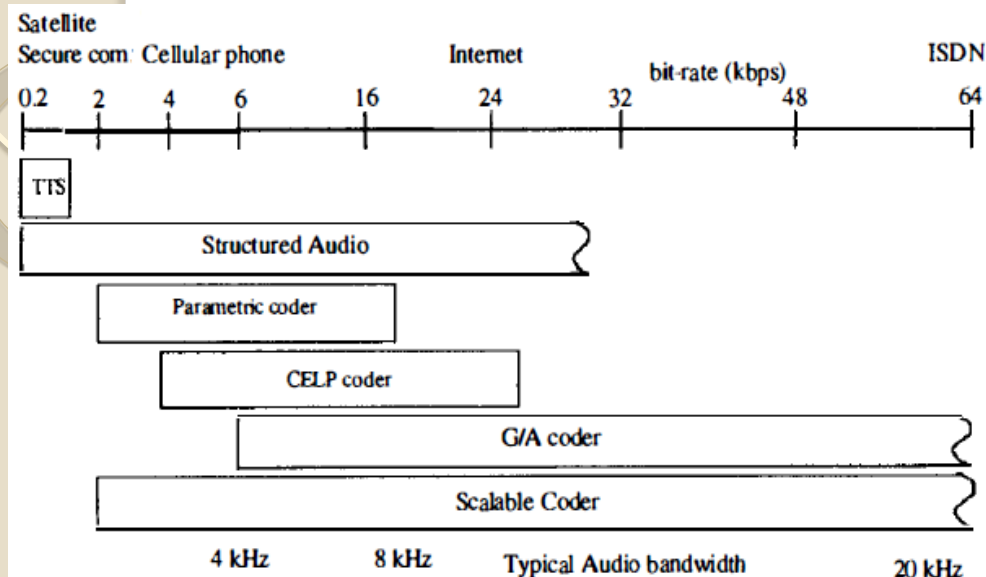
# MPEG-4 – Audio Goals



- Addressing various applications such as telephony and mobile comm., digital broadcasting, Internet works, interactive multimedia, etc..
- Requiring a high degree of coding efficiency together with flexible coded data for protection against transmission errors.
- Figure shows the typical data rate requirements for different applications versus the bandwidth of the coded signal
- The goals and functions include efficient audio coding, and the provision of speech coding for telephone service,
- Universal access by scalability of the coded data to address different channel requirements and robustness in error prone environment.
- Content based interactivity through flexible access and manipulation of coded data
- Support to synthetic audio and speech through the structured audio facility



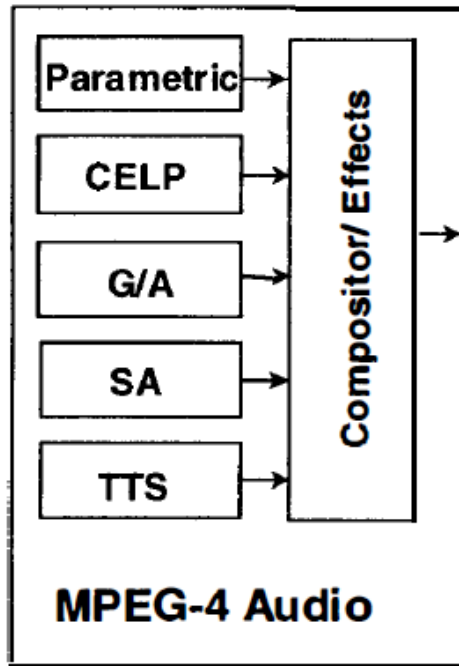
# MPEG-4 Standards – Audio Goals



MPEG-4 Audio data rates and target applications from [Edler 97]

- Test-to-speech (TTS): translates text information into speech for visually impaired, automatic response systems,... E.g., **voiced SMS** (200 and 1.2kb/s)
- Structured Audio (SA): Specifies a set of synthesis algorithms to create sounds with a set of synthesis control parameters such as MIDI, E.g., karaoke applications.
- Possible to generate sounds of musical instruments with SA orchestra language
- Code-excited linear prediction (CELP) is for speech signals with a good quality at 2 to 4 kb/s
- General audio (G/A) is for the data rates between 6 kb/s for audio signals with bandwidth of 4 kHz and 300 kb/s per channel for signals with bandwidths over 20 kHz for mono to multichannel audio.
- Scalability allows data to be parsed into bit streams of lower data rates that can still be decoded into a meaningful signal.

# MPEG-4 Standards



MPEG-4 Audio structure from [Grill 97a]

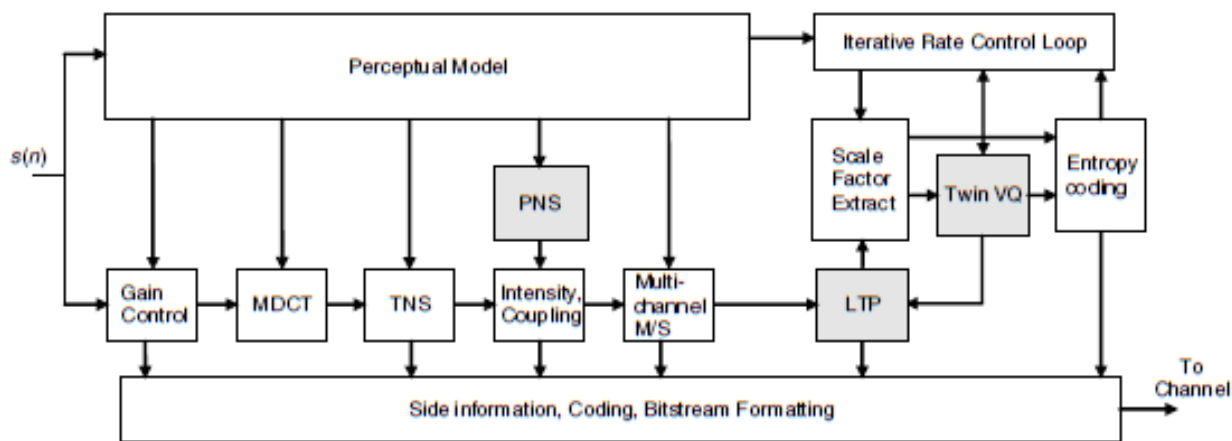
- This figure shows audio objects, from different algorithms, are combined to form a data stream by the compositor.
- Also multiplexed into video streams
- The compositor provides tools for manipulating the objects to produce a mixed or individual data stream
- A speaking voice and background music can be coded separately.
- Different data streams are obtained by the decoder for playing, edition and scene composition

**Example:** For a speaker's voice and the background music.

- Directly using the general audio (G/A) coding tool to produce a 64 b/s stream.
- Or the voice is coded using CELP at 16 kb/s per channel, and the music is produced by using SA tools at 2 kb/s per channel, being combined by the compositor;



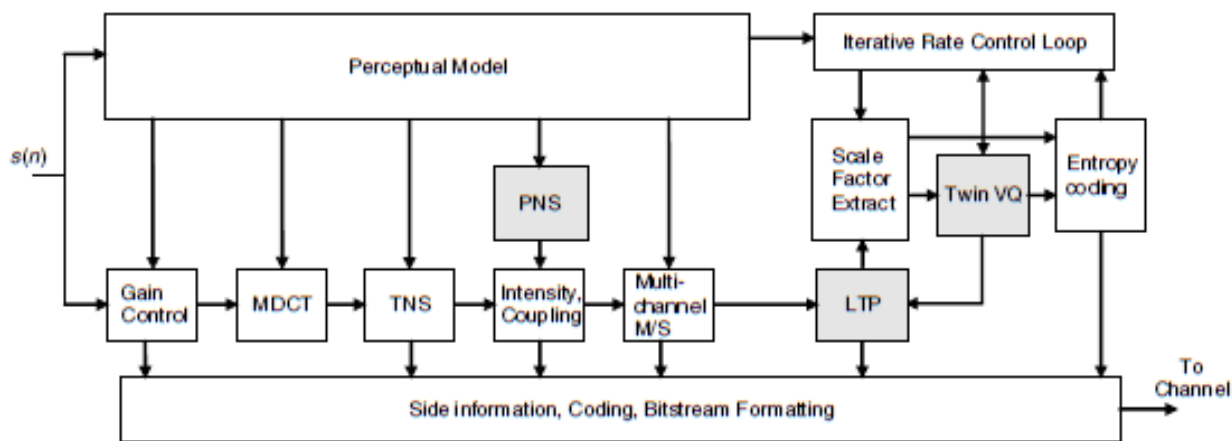
# MPEG-4 Standards –G/A coder



MPEG-4 GA coder.

- G/A has the most vital and versatile function that deals all coding systems
- Operates at bit rate from 6 to 300 k b/s at the sampling rates between 7.35 and 96 kHz
- It includes features given by perceptual noise substitution (PNS), long term prediction (LTP) and twin VQ and scalability.
- Perceptual noise substitution (PNS) allows frequency selective parametric encoding of noise-like components. These parameters are used at decoder to approximate these components.
- Reduce the no. of samples for full quantization and coding each sub-band transform coefficients.

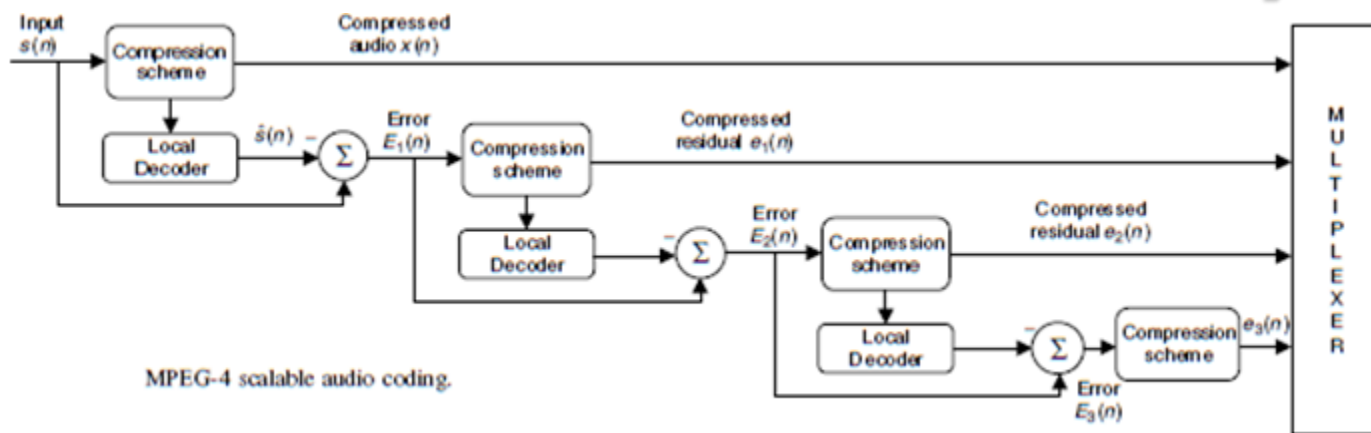
# MPEG-4 Standards –G/A coder



MPEG-4 GA coder.

- Long term prediction (LTP) deals with tonal signals with higher coding precision
  - Transform the input into frequency domain with analysis filter and temporal noise shaping (TNS)
  - Quantize and code the spectral coefficients
- It requires low complexity replacement for prediction tool to provide comparable performance with 50% saving in terms of memory and processing utilization.
- **Transform Domain Weighted Interleave Vector Quantization (Twin VQ)** performs vector quantization of the transformed spectrum coefficients based on a perceptually weighted model.
  - Provide high coding efficiencies even for music and tonal signals at extremely low bit rates (6-8 kb/s)

# MPEG-4 Standards - Scalability



- Supports variable rate encoding/decoding bit streams at a dynamic rates matching the varying transmission channel capacity.
- Encoding/decoding of a subset of the total bit stream results in a valid signal at a lower rate, i.e., decoding at 16, 32, or 64 kb/s from a 64 kb/s data stream according to the channel capacity, receiver complexity and quality requirements
- The core layer encodes the main audio stream guaranteeing reconstruction with minimum artifacts, while the enhancement layers provide further resolution and scalability for better quality.
- Successive error signals,  $E_1(n)$ ,  $E_2(n)$  and  $E_3(n)$  are used for the enhancement layers.
- Applications include digital audio broadcasting, mobile multimedia comm., and streaming audio

# MPEG-7 Standards

- It targets **content-based** multimedia applications and supports a broad range of applications including **indexing/searching, multimedia editing, broadcast media selection, and multi-media digital library sorting.**
- Provide the ways for efficient audio file retrieval and support for both text-based and context-based queries.
- Provide complementary functionality to MPEG-1, 2 and 4, as the first content-based standard incorporating multimedia interfaces based on descriptions.
- For example, spoken content search, musical instrument timbre search, sound recognition/indexing and audio identification/fingerprinting.

# MPEG-2I Framework

- The need for multimedia content access and distribution, MPEG-2I addresses the issues of **interoperability and automation**.
- Creating a platform that encompasses many functionalities for both content-users and content-creators/providers.
- Including applications of multimedia resource delivery to various networks and terminals such as PCs, PDAs, mobile devices, digital audio/video broadcasting, HDTV and home entertainment systems.
- Together MPEG-7 and MPEG-2I standards provide an open framework for building application-oriented interfaces or tools that satisfying a specific criterion such as a query or an audio file indexing.
- MPEG-7 standards provide an interface for indexing, accessing and distribution of multimedia content
- MPEG-2I defines an interoperable framework to access the multimedia content.

# MPEG-2I Framework

- MPEG-I and –II for compressions with good quality
- MPEG-4 for more applications with tools to support audio creations and management
- MPEG-7 standards provide an interface for indexing, accessing and distribution of multimedia content, more for friendly use by both creators and end users
- MPEG-2I defines an interoperable framework to access the multimedia content, for all kinds of users on various platforms.