



Speech & Audio Coding

5



Introductions

- 人的耳朵在單一音源下，其敏感程度可以高達16個位元
- 然而若是在有其它聲音來源的情況下，我們耳朵的敏感程度將明顯降低
- 人的中耳在頻率4KHz左右時的共振效果最大
- 語音的遠近感受，除了在於聲音傳到左右耳的時間差之外，最主要的還是來自於左右耳對聲音的相位差的感覺



Introductions

- 我們聽得比說得好：
 - 一般人類說話的時候所產生的頻率範圍（大概在 50Hz 到 10kHz 之間）比人類所能聽到聲音的頻率範圍（大概在 15Hz 到 20kHz 之間，隨年齡不同而有所改變）還要窄
- 人不會持續不斷地說
 - 當我們說話的時候，一連串的音節中間必然存在一些安靜的小段落
 - 不僅如此，在兩個人的交談中，“傳輸線”的每一個方向平均只使用百分之四十左右的時間



Introductions

- Sample rate and precision

| 數位聲音種類 | 取樣率 (<i>KHz</i>) |
|-------------|-----------------------|
| 數位電話 | 8.00 |
| 個人電腦 | 22.05 |
| 數位電視音訊 | 32.00 |
| <i>CD</i> | 44.10 |
| <i>HDTV</i> | 48.00 |



Speech coding

- Historically, digital speech signals are sampled at rate of 8000 samples/sec
- Typically, each sample is represented by 8 bits (using μ -law)
- This corresponds to an uncompressed rate of 64 Kbps
- With current compression techniques, it is possible to reduce the rate to 8 Kbps with almost no perceptibly loss in quality

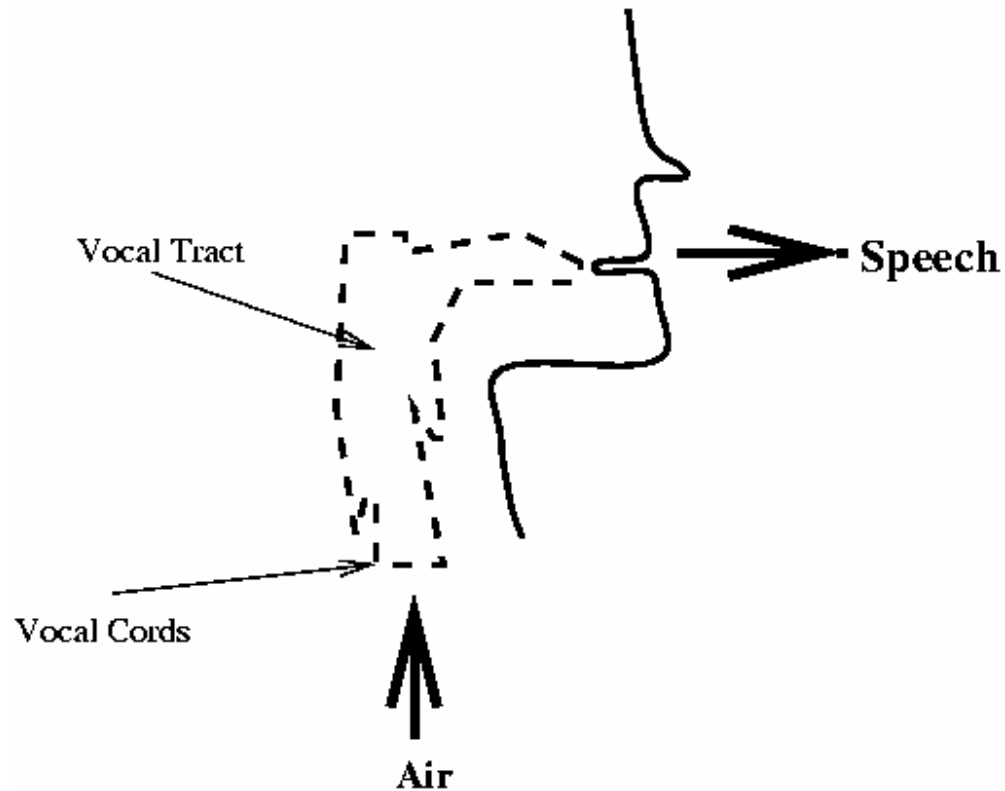


Speech coding

- All of the current low-rate speech coders are based on the principle of **linear predictive coding (LPC)**

LPC Modeling

- Physical Model





Physical model

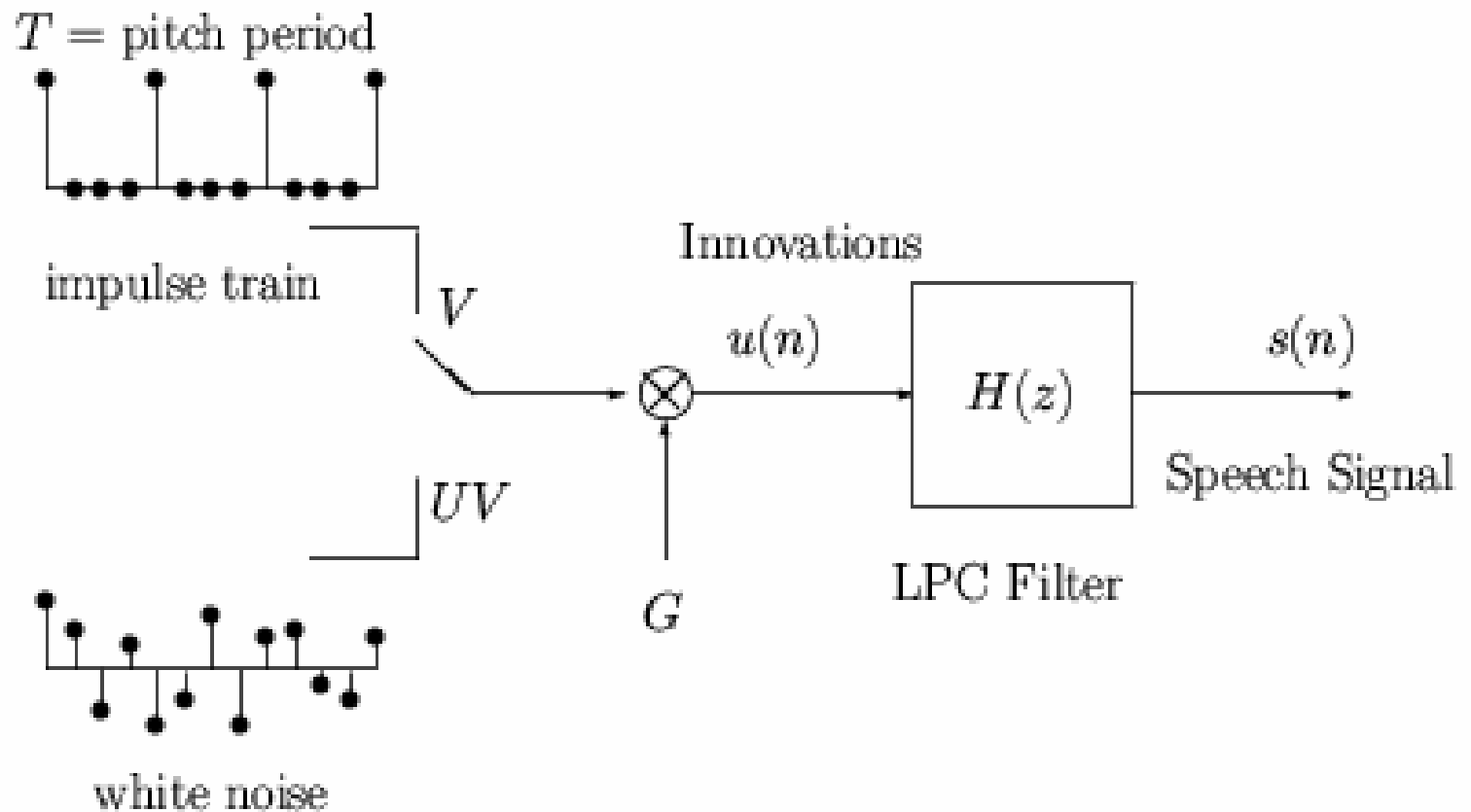
- When you speak
 - Air is pushed from your lung through your vocal tract and out of your mouth comes speech
 - For certain **voiced** sound, your vocal cords vibrate (open and close). The rate at which the vocal cords vibrate determines the **pitch** of your voice. Women and young children tend to have high pitch (fast vibration) while adult males tend to have low pitch (slow vibration)



Physical model

- For certain **fricatives and plosive (or unvoiced)** sound, your vocal cords do not vibrate but remain constantly opened.
- The shape of your vocal tract determines the sound that you make
- As you speak, your vocal tract changes its shape producing different sound
- The shape of the vocal tract changes relatively slowly (on the scale of 10 msec to 100 msec)
- The amount of air coming from your lung determines the loudness of your voice

Mathematical Model





Mathematical Model

- The above model is often called the LPC Model.
- The model says that the digital speech signal is the output of a digital filter (called the LPC filter) whose input is either a train of impulses or a white noise sequence.



Mathematical Model

- The **relationship** between the physical and the mathematical models

Vocal Tract \longleftrightarrow $H(z)$ (LPC Filter)

Air \longleftrightarrow $u(n)$ (Innovations)

Vocal Cord Vibration \longleftrightarrow V (voiced)

Vocal Cord Vibration Period \longleftrightarrow T (pitch period)

Fricatives and Plosives \longleftrightarrow UV (unvoiced)

Air Volume \longleftrightarrow G (gain)



Mathematical Model

- The LPC filter is given by:

$$H(z) = \frac{1}{1 + a_1z^{-1} + a_2z^{-2} + \dots + a_{10}z^{-10}}$$

which is equivalent to saying that the input-output relationship of the filter is given by the linear difference equation:

$$s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n)$$



Mathematical Model

- The LPC model can be represented in vector form as:

$$\mathbf{A} = (a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, G, V/UV, T)$$

- A changes every 20 msec or so. At a sampling rate of 8000 samples/sec, 20 msec is equivalent to 160 samples
- The digital speech signal is divided into **frames** of size 20 msec. There are 50 frames/second



Mathematical Model

- The model says that

$$\mathbf{A} = (a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, G, V/UV, T)$$

is equivalent to

$$\mathbf{S} = (s(0), s(1), \dots, s(159))$$

Thus the 160 values of S is compactly represented by the 13 values of A



Mathematical Model

- There's almost no perceptual difference in S if:
 - **For Voiced Sounds (V)** : the impulse train is shifted (insensitive to phase change).
 - **For Unvoiced Sounds (UV)** : a different white noise sequence is used
- **LPC Synthesis** : Given A , generate S (this is done using standard filtering techniques).
- **LPC Analysis** : Given S , find the best A



LPC Analysis

- Consider one frame of speech signal :

$$\mathbf{S} = (s(0), s(1), \dots, s(159))$$

- The signal $s(n)$ is related to the innovation $u(n)$ through the linear difference equation :

$$s(n) + \sum_{i=1}^{10} a_i s(n-i) = u(n)$$

- The ten LPC parameters $(a_1, a_2, \dots, a_{10})$ are chosen to minimize the energy of the innovation (air) :

$$f = \sum_{n=0}^{159} u^2(n)$$



LPC Analysis

- Using standard calculus, we take the derivative of f with respect to a_i and set it to zero :

$$\begin{aligned} df/da_1 &= 0 \\ df/da_2 &= 0 \\ &\dots \\ df/da_{10} &= 0 \end{aligned}$$

$$\begin{aligned} R(k) &= \sum_{n=0}^{159-k} s(n)s(n+k) \\ &= \text{autocorrelation of } s(n) \end{aligned}$$

- We now have 10 linear equations with 10 unknowns :

$$\begin{bmatrix} R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) & R(8) & R(9) \\ R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) & R(8) \\ R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) & R(7) \\ R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) & R(6) \\ R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) & R(5) \\ R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) & R(4) \\ R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) & R(3) \\ R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) & R(2) \\ R(8) & R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) & R(1) \\ R(9) & R(8) & R(7) & R(6) & R(5) & R(4) & R(3) & R(2) & R(1) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \\ a_9 \\ a_{10} \end{bmatrix} = \begin{bmatrix} -R(1) \\ -R(2) \\ -R(3) \\ -R(4) \\ -R(5) \\ -R(6) \\ -R(7) \\ -R(8) \\ -R(9) \\ -R(10) \end{bmatrix}$$



DPCM

- Optimal predictor

$$\hat{x}_m = \sum_{i=0}^{m-1} \alpha_i x_i = \alpha_{m-1} x_{m-1} + \alpha_{m-2} x_{m-2} + \dots + \alpha_0 x_0$$

We shall minimize the variance of $e_m = x_m - \hat{x}_m$,

i.e.,
$$\sigma_e^2 = E \left\{ (x_m - \hat{x}_m)^2 \right\} = E \left\{ \left(x_m - \sum_{i=0}^{m-1} \alpha_i x_i \right)^2 \right\}$$



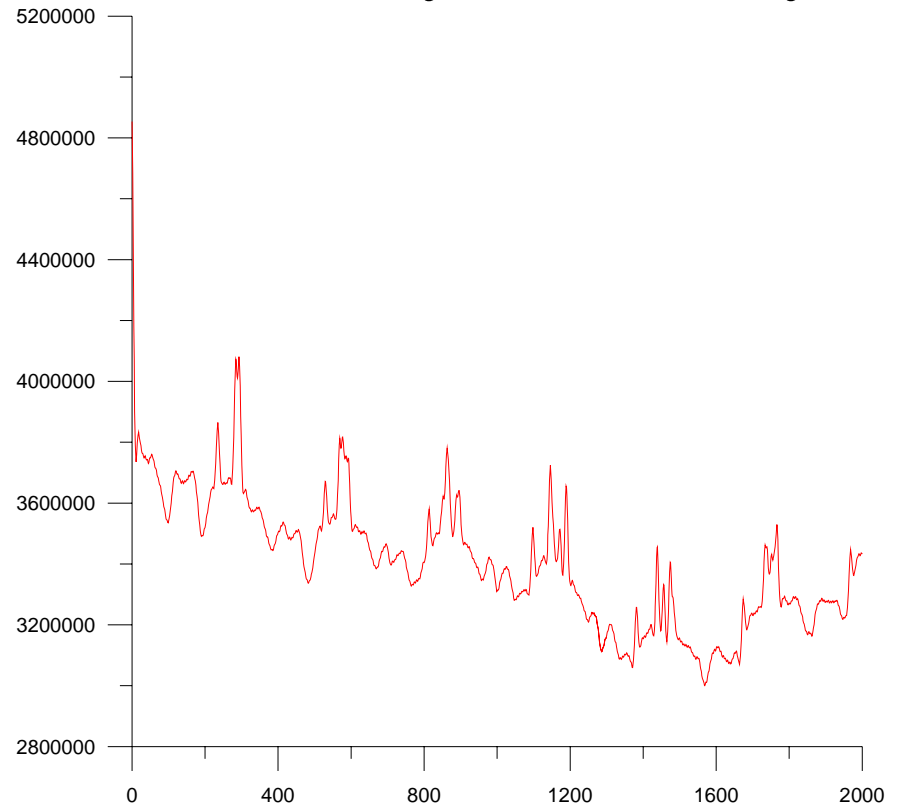
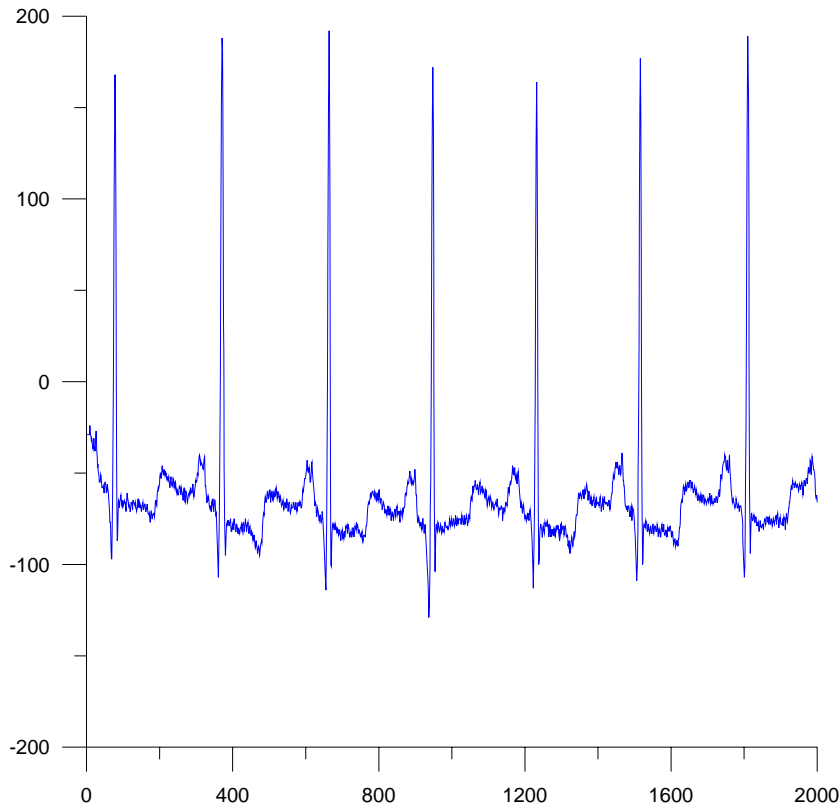
DPCM

- To find the minimal value, we partial differentiate the following and let it be zero :

$$\begin{aligned} & \frac{\partial E \left[(x_m - \hat{x}_m)^2 \right]}{\partial \alpha_i} \\ &= \frac{\partial E \left\{ \left[x_m - (\alpha_0 x_0 + \alpha_1 x_1 + \dots + \alpha_{m-1} x_{m-1}) \right]^2 \right\}}{\partial \alpha_i} \\ &= -2E \left[(x_m - \hat{x}_m) x_i \right] = 0 \quad i = 0, 1, \dots, m - 1 \quad (7.1) \end{aligned}$$

DPCM

Define the autocorrelation as $R_{ij} = E(x_i x_j)$





DPCM

Then by (7.1), we have :

$$E \{ \mathbf{x}_m \mathbf{x}_i \} = E \{ \hat{\mathbf{x}}_m \mathbf{x}_i \}$$

$$\begin{aligned} \mathbf{R}_{mi} &= E \{ \alpha_0 \mathbf{x}_0 \mathbf{x}_i + \alpha_1 \mathbf{x}_1 \mathbf{x}_i + \dots + \alpha_{m-1} \mathbf{x}_{m-1} \mathbf{x}_i \} \\ &= \alpha_0 \mathbf{R}_{0i} + \alpha_1 \mathbf{R}_{1i} + \dots + \alpha_{m-1} \mathbf{R}_{(m-1)i} \end{aligned}$$

$$i = 0, 1, \dots, m - 1 \quad (7.2)$$

$\alpha_i, i = 0, 1, \dots, m - 1$ can be obtained by solving the m equations in (7.2)

DPCM

If \hat{x}_m is obtained by the predictor that uses the optimal α_i thus obtained, then

$$\begin{aligned}\sigma_e^2 &= E \left\{ (x_m - \hat{x}_m)^2 \right\} \\ &= E \left\{ (x_m - \hat{x}_m) x_m \right\} - E \left\{ (x_m - \hat{x}_m) \hat{x}_m \right\}\end{aligned}$$

But $E \left\{ (x_m - \hat{x}_m) \hat{x}_m \right\} = 0$ from eq.(7.1) Thus,

$$\begin{aligned}\sigma_e^2 &= E \left\{ (x_m - \hat{x}_m) x_m \right\} \\ &= E \left[x_m^2 \right] - E \left[\hat{x}_m x_m \right] \\ &= R_{mm} - \left(\alpha_0 R_{0m} + \alpha_1 R_{1m} + \dots + \alpha_{m-1} R_{m-1m} \right)\end{aligned}$$

σ_e^2 : can be considered as the variance of the error signal.

R_{mm} : can be considered as the variance of original signal.



DPCM

$$\sigma_e^2 = R_{mm} - \left(\alpha_{m-1} R_{(m-1)m} + \alpha_{m-2} R_{(m-2)m} + \dots + \alpha_0 R_{0m} \right)$$

$$\hat{x}_m = \sum_{i=0}^{m-1} \alpha_i x_i = \alpha_{m-1} x_{m-1} + \alpha_{m-2} x_{m-2} + \dots + \alpha_0 x_0$$



LPC Analysis

- The above matrix equation could be solved using:
 - The Gaussian elimination method
 - Any matrix inversion method (MATLAB)
 - The Levinson-Durbin recursion (described below)



Levinson-Durbin Recursion

$$E^{(0)} = R(0)$$

$$k_i = [R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j)] / E^{(i-1)} \quad i = 1, 2, \dots, 10$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \quad j = 1, 2, \dots, i-1$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}$$

- Solve the above for $i=1, 2, \dots, 10$, and then set

$$a_i = -\alpha_i^{(10)}$$



Levinson-Durbin Recursion

- To get the other three parameters: $(V/UV, G, T)$, we solve for the innovation:

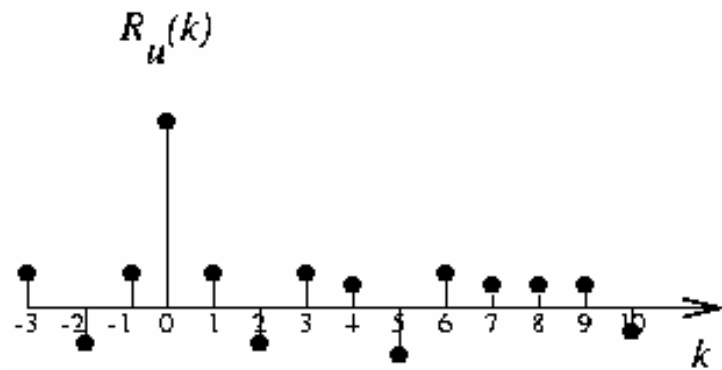
$$u(n) = s(n) + \sum_{i=1}^{10} a_i s(n-i)$$

- Then calculate the autocorrelation of $u(n)$

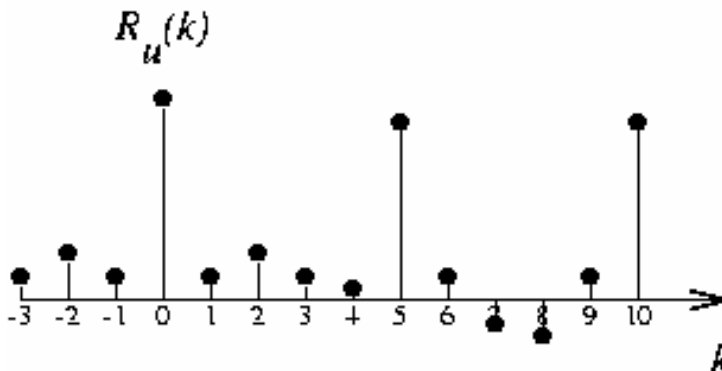
$$R_u(k) = \sum_{n=0}^{159-k} u(n)u(n+k)$$

Levinson-Durbin Recursion

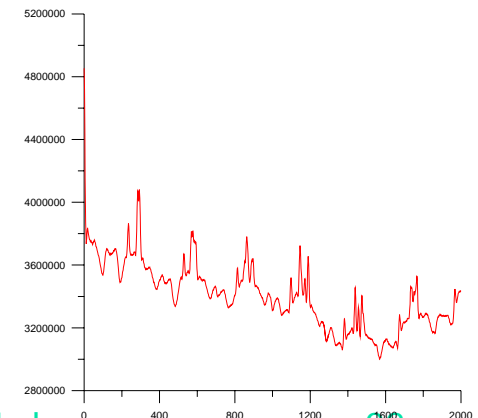
- Then make a decision based on the autocorrelation:



$\Rightarrow UV$



$\Rightarrow V, T=5$





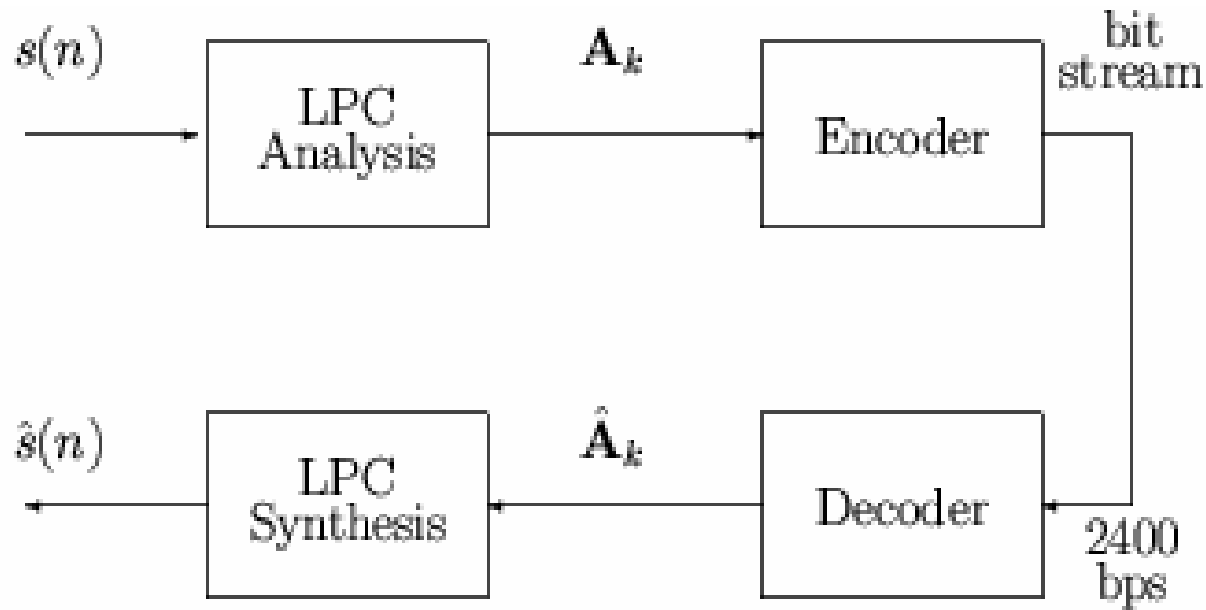
Levinson-Durbin Recursion

- Gain G can be calculated by

$$\begin{aligned} G &= E\left\{ \left[s(n) - \sum_{i=1}^{10} a_i s(n-i) \right]^2 \right\} \\ &= E\left\{ \left[s(n) - \sum_{i=1}^{10} a_i s(n-i) \right] s(n) \right\} \\ &= \phi(0,0) - \sum_{i=1}^{10} a_i \phi(0,i) \\ &= R(0) - \sum_{i=1}^{10} a_i R(i) \end{aligned}$$

2.4 Kbps LPC Vocoder

- The following is a block diagram of a 2.4 kbps LPC Vocoder :





2.4 Kbps LPC Vocoder

- The LPC coefficients are represented as *line spectrum pair* (LSP) parameters
- LSP are mathematically equivalent (one-to-one) to LPC
- LSP are more amenable to quantization
- LSP are calculated as follows:

$$\begin{aligned}P(z) &= 1 + (a_1 - a_{10})z^{-1} + (a_2 - a_9)z^{-2} + \dots + (a_{10} - a_1)z^{-10} - z^{-11} \\Q(z) &= 1 + (a_1 + a_{10})z^{-1} + (a_2 + a_9)z^{-2} + \dots + (a_{10} + a_1)z^{-10} + z^{-11}\end{aligned}$$



2.4 Kbps LPC Vocoder

- Factoring the above equations, we get :

$$P(z) = (1 - z^{-1}) \prod_{k=2,4,\dots,10} (1 - 2 \cos \omega_k z^{-1} + z^{-2})$$

$$Q(z) = (1 + z^{-1}) \prod_{k=1,3,\dots,9} (1 - 2 \cos \omega_k z^{-1} + z^{-2})$$

$\{\omega_k\}_{k=1}^{10}$ are called the LSP parameters.

- LSP are ordered and bounded :

$$0 < \omega_1 < \omega_2 < \dots < \omega_{10} < \pi$$



2.4 Kbps LPC Vocoder

- LSP are more correlated from one frame to the next than LPC.
- The frame size is 20 msec. There are 50 frames/sec. 2400 bps is equivalent to 48 bits/frame. These bits are allocated as follows:

| Parameter Name | Parameter Notation | Rate (bits/frame) |
|--------------------------|--|-------------------|
| LPC (LSP) | $\{a_k\}_{k=1}^{10}$ ($\{\omega_k\}_{k=1}^{10}$) | 34 |
| Gain | G | 7 |
| Voiced/Unvoiced & Period | $V/UV, T$ | 7 |
| Total | | 48 |



2.4 Kbps LPC Vocoder

- The 34 bits for the LSP are allocated as follows:

| LSP | No. of Bits |
|---------------|-------------|
| ω_1 | 3 |
| ω_2 | 4 |
| ω_3 | 4 |
| ω_4 | 4 |
| ω_5 | 4 |
| ω_6 | 3 |
| ω_7 | 3 |
| ω_8 | 3 |
| ω_9 | 3 |
| ω_{10} | 3 |
| Total | 34 |



2.4 Kbps LPC Vocoder

- The gain, G , is encoded using a 7-bit non-uniform scalar quantizer
- For voiced speech, values of T ranges from 20 to 146. V/UV , T are jointly encoded as follows:

| V/UV | T | Encoded Value |
|--------|-----|---------------|
| UV | — | 0 |
| V | 20 | 1 |
| V | 21 | 2 |
| V | 22 | 3 |
| V | 23 | 4 |
| ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ |
| V | 146 | 127 |

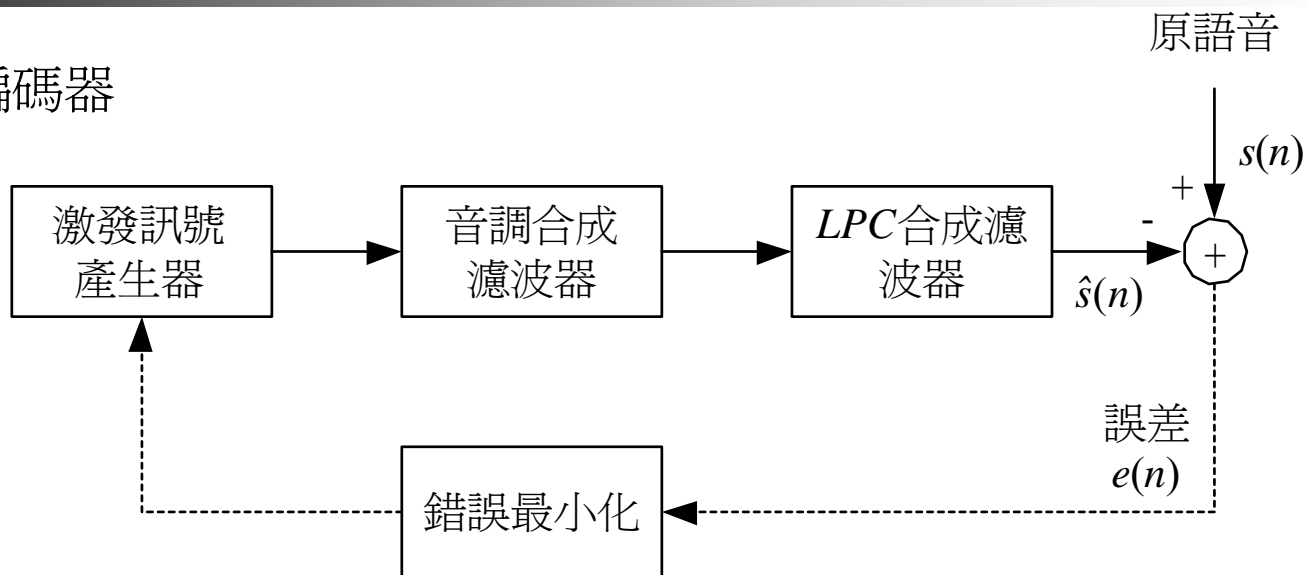


AbS-LPC

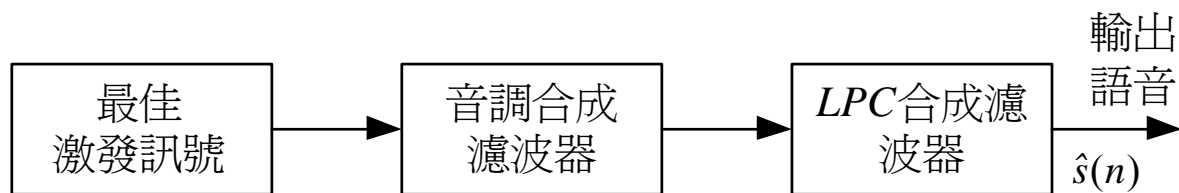
- Algorithm that is most suitable for low bit rate speech coding
 - AbS-LPC
 - Analysis by Synthesis LPC

AbS-LPC

編碼器



解碼器





AbS-LPC

- Better compression performance can be obtained by careful design of
 - LPC filter : STP (Short-Term Predictor)
 - Pitch filter : LTP (Long-Term Predictor)
 - Perceptually based error minimization procedure
 - Excitation signal
 - Voiced : periodic pulse
 - Unvoiced : white noise

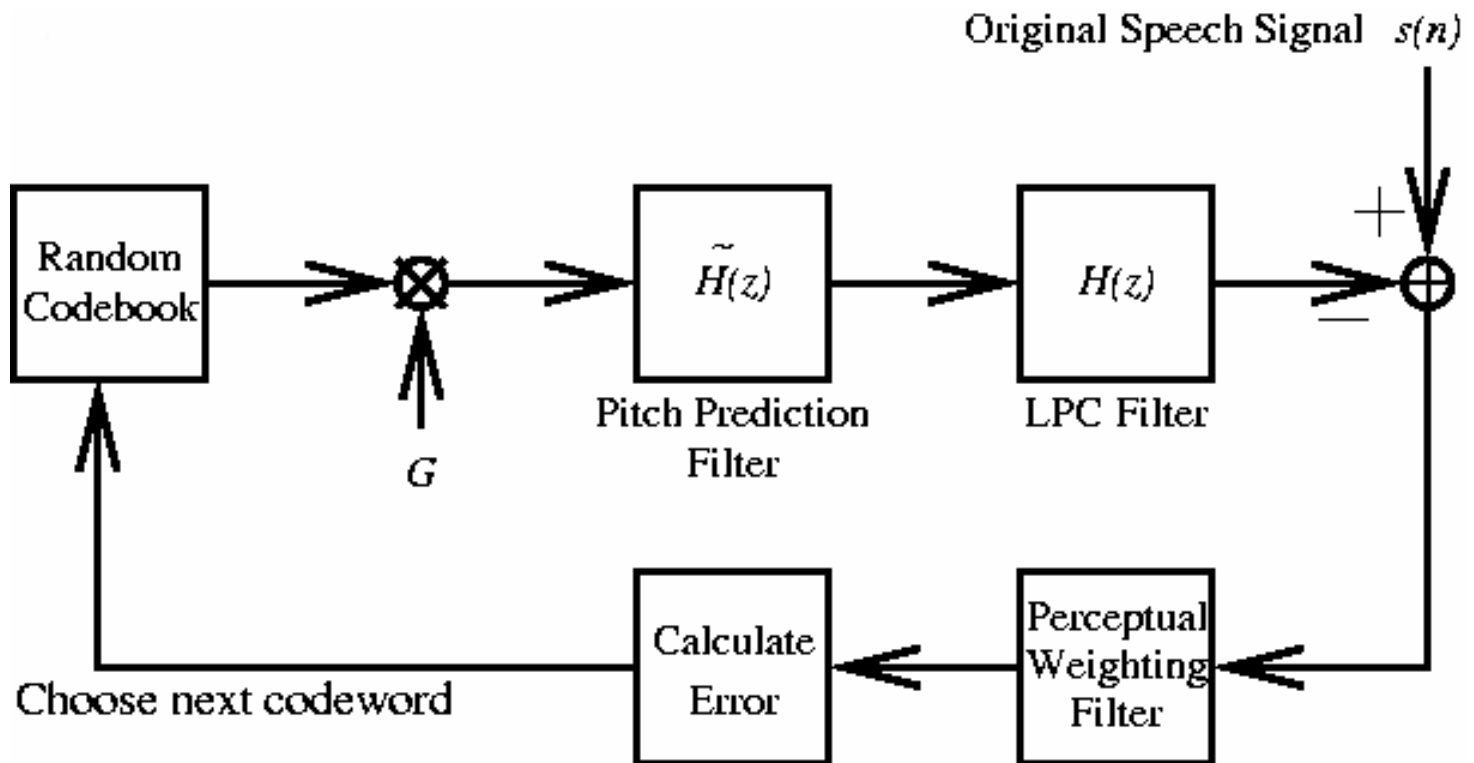


4.8 Kbps CELP Coder

- CELP=Code-Excited Linear Prediction
- The principle is similar to the LPC Vocoder except:
 - Frame size is 30 msec (240 samples)
 - $u(n)$ is coded directly
 - More bits are need
 - Computationally more complex
 - A pitch prediction filter is included
 - Vector quantization concept is used

4.8 Kbps CELP Coder

- A block diagram of the CELP encoder





4.8 Kbps CELP Coder

- The pitch prediction filter is given by:

$$\tilde{H}(z) = \frac{1}{1 + bz^{-T}}$$

where T could be an integer or a fraction thereof

- The perceptual weighting filter is given by:

$$W(z) = \frac{H(z/\gamma_2)}{H(z/\gamma_1)}$$

where $\gamma_1 = 0.9$, $\gamma_2 = 0.5$ have been determined to be good choices.



4.8 Kbps CELP Coder

- Each frame is divided into 4 subframes. In each subframe, the codebook contains 512 codevectors.
- The gain is quantized using 5 bits per subframe.
- The LSP parameters are quantized using 34 bits similar to the LPC Vocoder.
- At 30 msec per frame, 4.8 kbps is equivalent to 144 bits/frame. These 144 bits are allocated as follows:

| Parameters | No. of Bits |
|-------------------------|-------------|
| LSP | 34 |
| Pitch Prediction Filter | 48 |
| Codebook Indices | 36 |
| Gains | 20 |
| Synchronization | 1 |
| FEC | 4 |
| Future Expansion | 1 |
| Total | 144 |



8.0 Kbps CS-ACELP

- Conjugate-Structured Algebraic CELP
- The principle is similar to the 4.8 kbps CELP Coder except:
 - Frame size is 10 msec (80 samples)
 - There are only two subframes, each of which is 5 msec (40 samples)
 - The LSP parameters are encoded using two-stage vector quantization
 - The gains are also encoded using vector quantization.



8.0 Kbps CS-ACELP

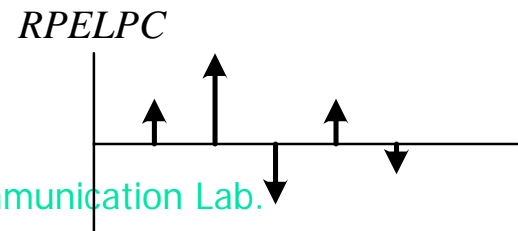
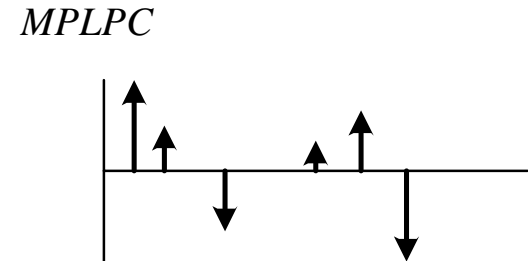
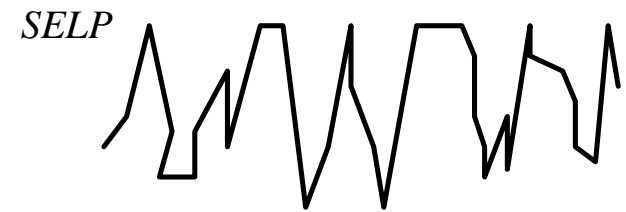
- At 10 msec per frame, 8 Kbps is equivalent to 80 bits/frame. These 80 bits are allocated as follows:

| Parameters | No. of Bits |
|-------------------------|-------------|
| LSP | 18 |
| Pitch Prediction Filter | 14 |
| Codebook Indices | 34 |
| Gains | 14 |
| Total | 80 |

Excitation signal

- $U_i = g_i X_i$
 - U_i : i -th L -dimensional excitation signal
 - g_i : L -dimensional gain
 - X_i : unit shape vector of $M \times L$
 - CELP : codebook
 - SELP : self-excitation
 - MPLPC : multi-pulse
 - RPELPC : regular pulse

0 L-1





CELP

- Codebook LPC
- The codebook contains C representative excitation vectors and gain vectors.
- Gain vectors are usually scalar
- Gain vectors can be vectors in order to weight differently each component of an excitation vector



SELP

- Self excitation
- Unit shape vectors are generated directly from coded, previous excitation functions



MPLPC

- Multi pulse LPC
- The first AbS-LPC
- Voiced or unvoiced classification not used any more
- Excitation signals are defined by a set of various amplitude, unequally spaced pulses
- Coding works : find out the pulse positions and amplitudes that minimize the error



RPELPC

- Regular pulse LPC
- Equally spaced pulses
- Only position of the pulse first pulse need to be coded

Speech coding standards

| 推出年份 | 位元率 (Kb/s) | 描述 | MOS |
|------|------------|--------------|------|
| 1972 | 64 | PCM | 4.4 |
| 1976 | 2.4 | LPC-10 | 2.7 |
| 1984 | 32 | G.721 | 4.1 |
| 1988 | 48, 56, 64 | G.722 | ≈4.0 |
| 1990 | 4.15 | INMARSAT | ≈3.2 |
| 1991 | 13 | GSM | 3.6 |
| 1991 | 4.8 | CELP | 3.2 |
| 1992 | 16 | G.728 | 4.0 |
| 1992 | 8 | VSELP | 3.5 |
| 1993 | 1~8 | QCELP | ≈3.4 |
| 1993 | 6.8 | VSELP | ≈3.3 |
| 1995 | 8 | G.729 | ≈4.2 |
| 1995 | 6.3 | G.723.1 | 3.98 |
| 1995 | 5~6 | 半位元率 GSM | ≈3.4 |
| 1996 | 2.4 | MELP | ≈3.3 |
| 1999 | 1.4~23.8 | MPEG-4 audio | -- |



G.72x

- G.711
 - PCM
 - 64 Kbps
- G.721
 - 3003 ~ 4000 Hz
 - ADPCM
 - 32 Kbps
- G.721 was replaced by G.726 in 1990
 - 16, 24, 32, and 40 Kbps
 - ADPCM



G.72x

- G.722
 - Wideband telephony : 0~8KHz
 - Enhanced G.726
 - Subband + ADPCM
 - 0~4 KHz and 4~8KHz subbands
 - Low freq. Subband : 48 Kbps
 - High freq. Subband : 16 Kbps
 - Total bit rate : 64 Kbps



G.72x

- G.728
 - Quality better than G.721
 - 16 Kbps, however
 - LD-CELP (Low-Delay CELP)
 - Index of the excitation codevector that produces the minimal signal error is transmitted

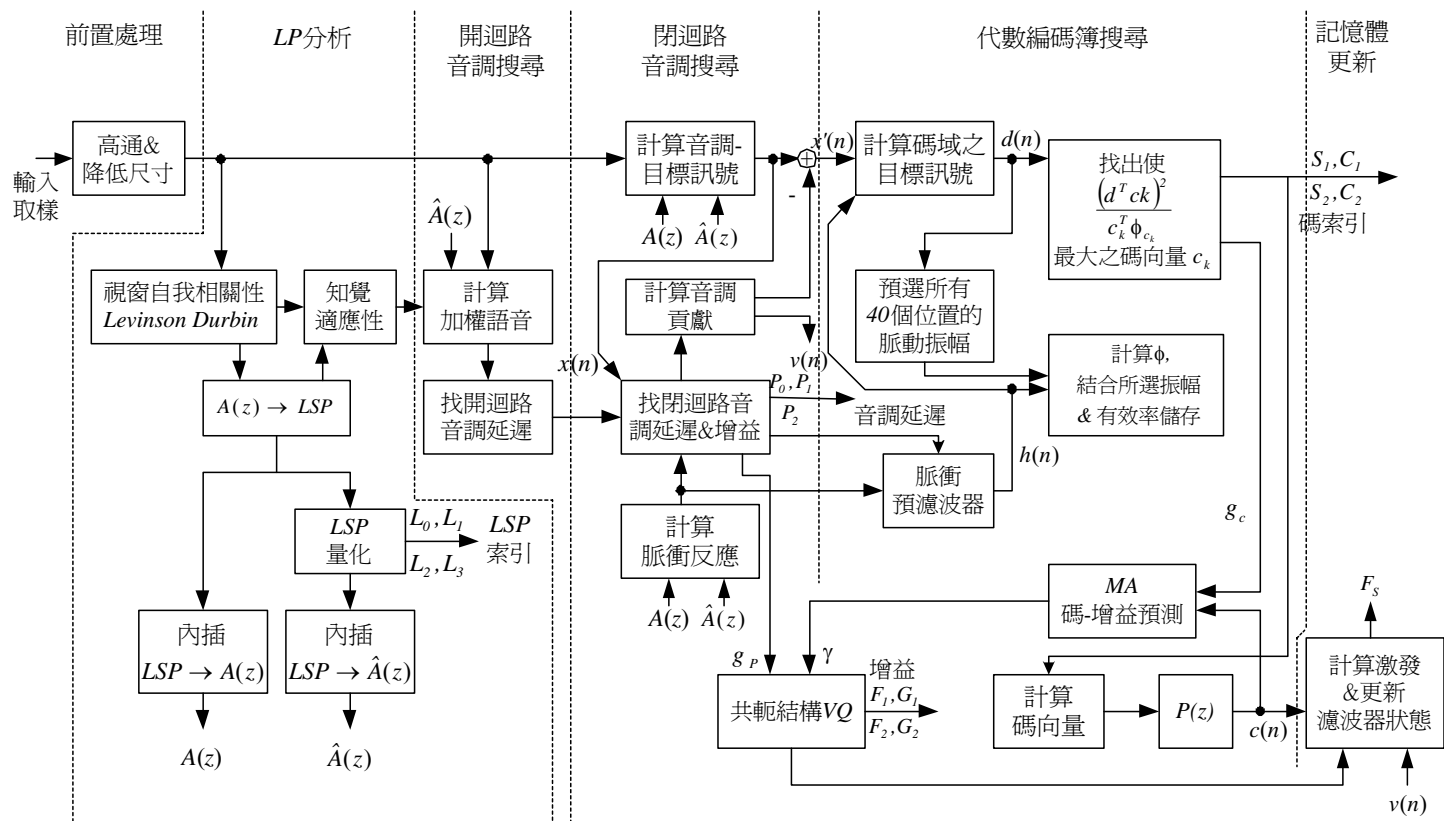


G.729

- The first CODEC that operates at 8 Kbps
- Quality not worse than G.726 (32 Kbps)
- Delay : less than 16 ms
- Robust under noisy environment
- CS-ACELP
 - Conjugate Structure-Algebraic CELP
 - Special codebook structure to speed up codebook search

G.729

← 每一個資料框 → → 每一個子資料框 →





G.729

- Parameters (bit numbers) for each 10 ms frame

| 參數 | 碼向量 | 位元數 |
|-----------------------|----------------------|--------|
| 線性頻譜對 | L_0, L_1, L_2, L_3 | 18 |
| 適應性編碼簿之索引 | P_1, P_2 | 8, 5 |
| P_1 之 <i>parity</i> | P_0 | 1 |
| 固定編碼簿之索引 | C_1, C_2 | 13, 13 |
| 固定編碼簿之脈衝正負號 | S_1, S_2 | 4, 4 |
| 編碼簿增益 (第一階) | F_1, F_2 | 3, 3 |
| 編碼簿增益 (第二階) | G_1, G_2 | 4, 4 |
| 總共 | | 80 |



G.729

- G.729 was later improved as G.729A and G.729B, respectively

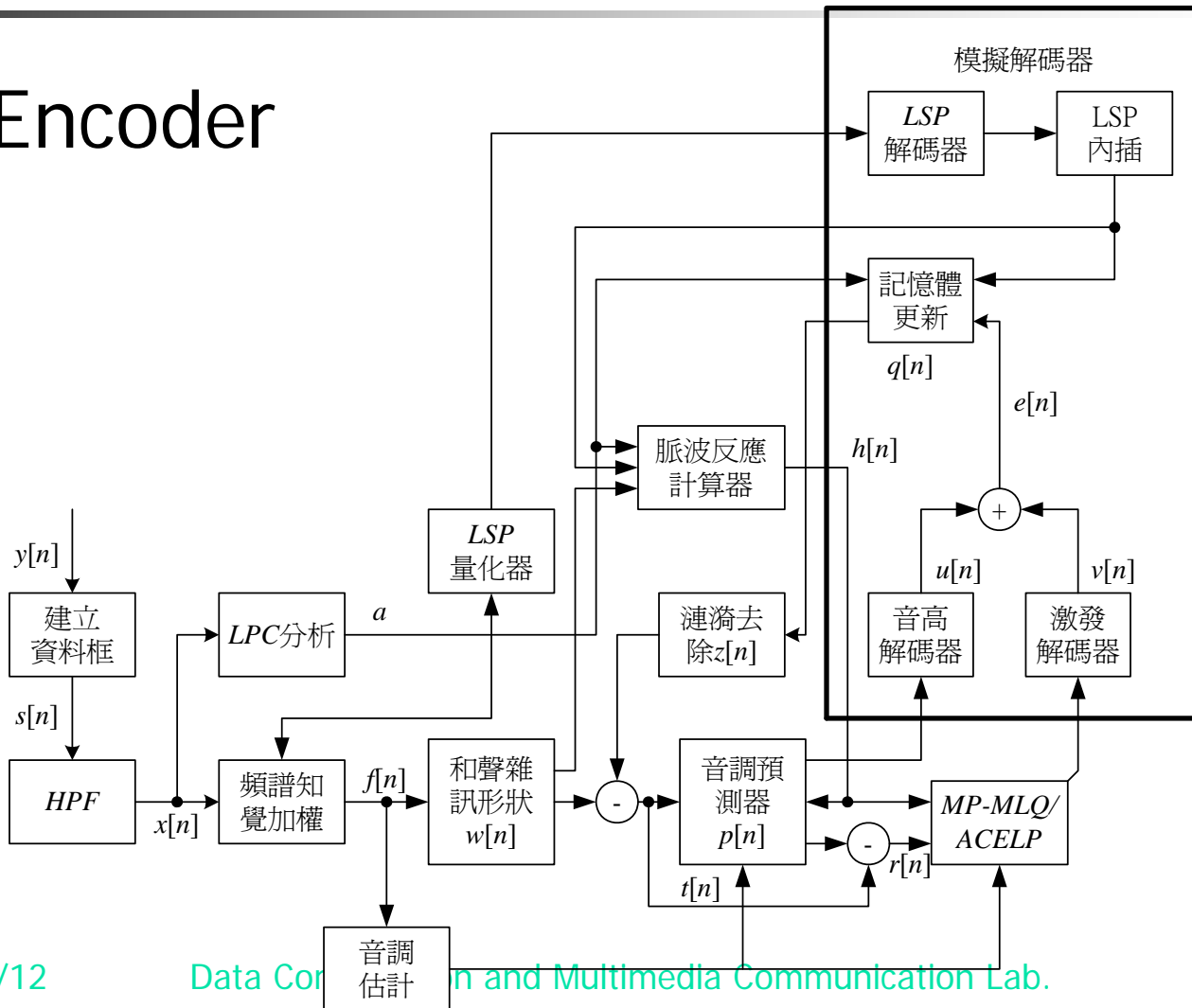


G.723.1

- Dual bit rates
 - 6.3 Kbps, 5.3 Kbps
 - Uses different excitation codebook
- 6.3 Kpps
 - MPLPC
 - Even frames use 6 non-zero pulses
 - Odd frames use 5
 - Positions should either be all odd or all even
 - Single gain for all excitation pulses

G.723.1

Encoder





G.723.1

- Parameters (bit numbers) for each 30 ms frame – 6.3 Kbps

| 參數 | 子資料框 0 | 子資料框 1 | 子資料框 2 | 子資料框 3 | 總共 |
|---------------|--------|--------|--------|--------|-----|
| <i>LPC</i> 索引 | | | | | 24 |
| 適應性編碼簿延遲 | 7 | 2 | 7 | 2 | 18 |
| 所有結合的增益 | 12 | 12 | 12 | 12 | 48 |
| 脈衝位置 | 20 | 18 | 20 | 18 | 76 |
| 脈衝正負號 | 6 | 5 | 6 | 5 | 22 |
| 格子索引 | 1 | 1 | 1 | 1 | 4 |
| 總共 | | | | | 192 |

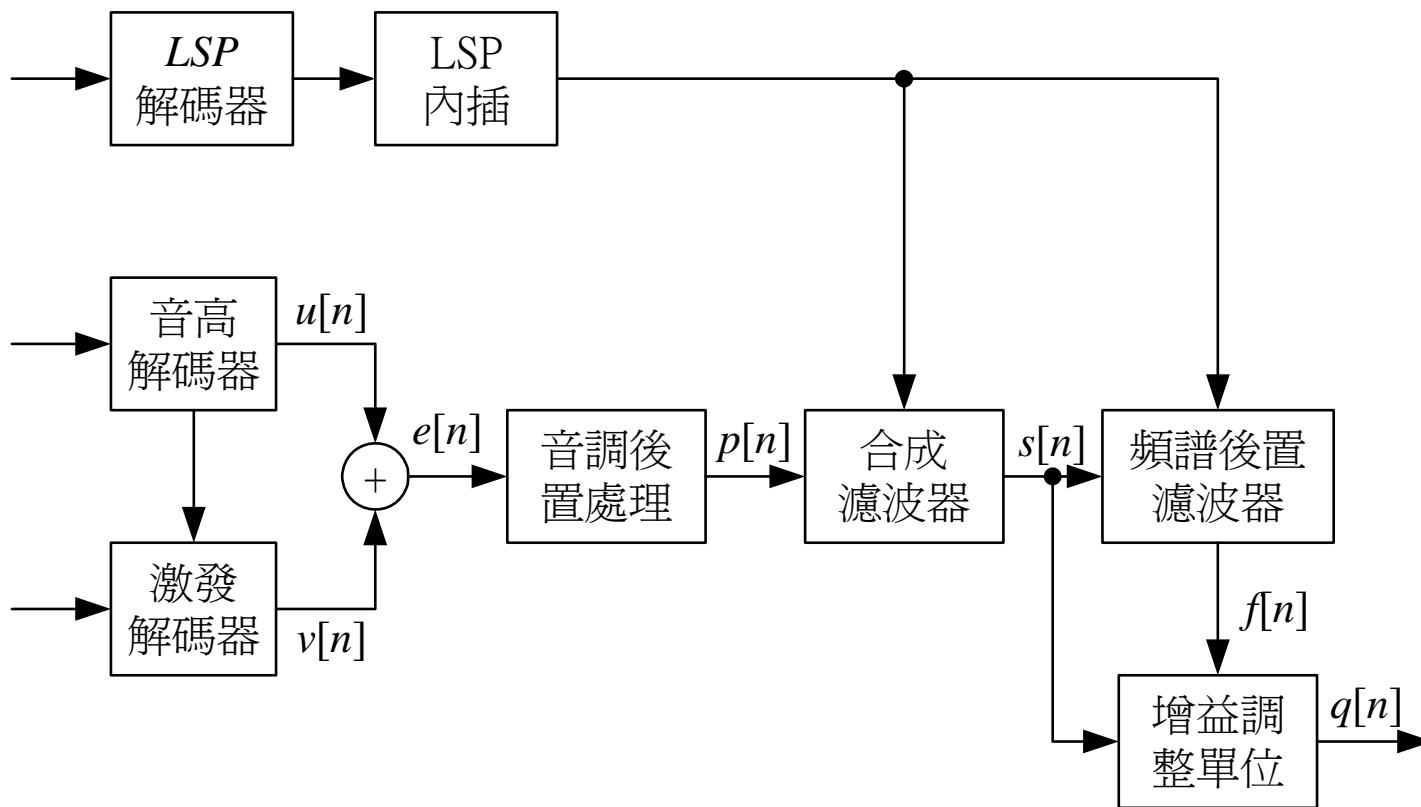


G.723.1

- Parameters (bit numbers) for each 30 ms frame – 5.27 Kbps

| 參數 | 子資料框 0 | 子資料框 1 | 子資料框 2 | 子資料框 3 | 總共 |
|---------------|--------|--------|--------|--------|-----|
| <i>LPC</i> 索引 | | | | | 24 |
| 適應性編碼簿延遲 | 7 | 2 | 7 | 2 | 18 |
| 所有結合的增益 | 12 | 12 | 12 | 12 | 48 |
| 脈衝位置 | 12 | 12 | 12 | 12 | 48 |
| 脈衝正負號 | 4 | 4 | 4 | 4 | 16 |
| 格子索引 | 1 | 1 | 1 | 1 | 4 |
| 總共 | | | | | 158 |

G.723.1



MPEG-4 speech CODEC

- Three mode : same LPC coefficients coder, but different excitation modules

| 架構 | 激發訊號模組 | 位元率範圍 (Kb/s) | 取樣率 (KHz) | 可調性 |
|----------------|-------------|-----------------|--------------|-------------|
| <i>HVXC</i> | <i>HVXC</i> | 1.4~4 | 8 | 位元率 |
| <i>CELP I</i> | <i>RPE</i> | 14.4~22.5 | 16 | ----- |
| <i>CELP II</i> | <i>MPE</i> | 3.85~23.8 | 8 & 16 | 位元率 & 頻寬 |

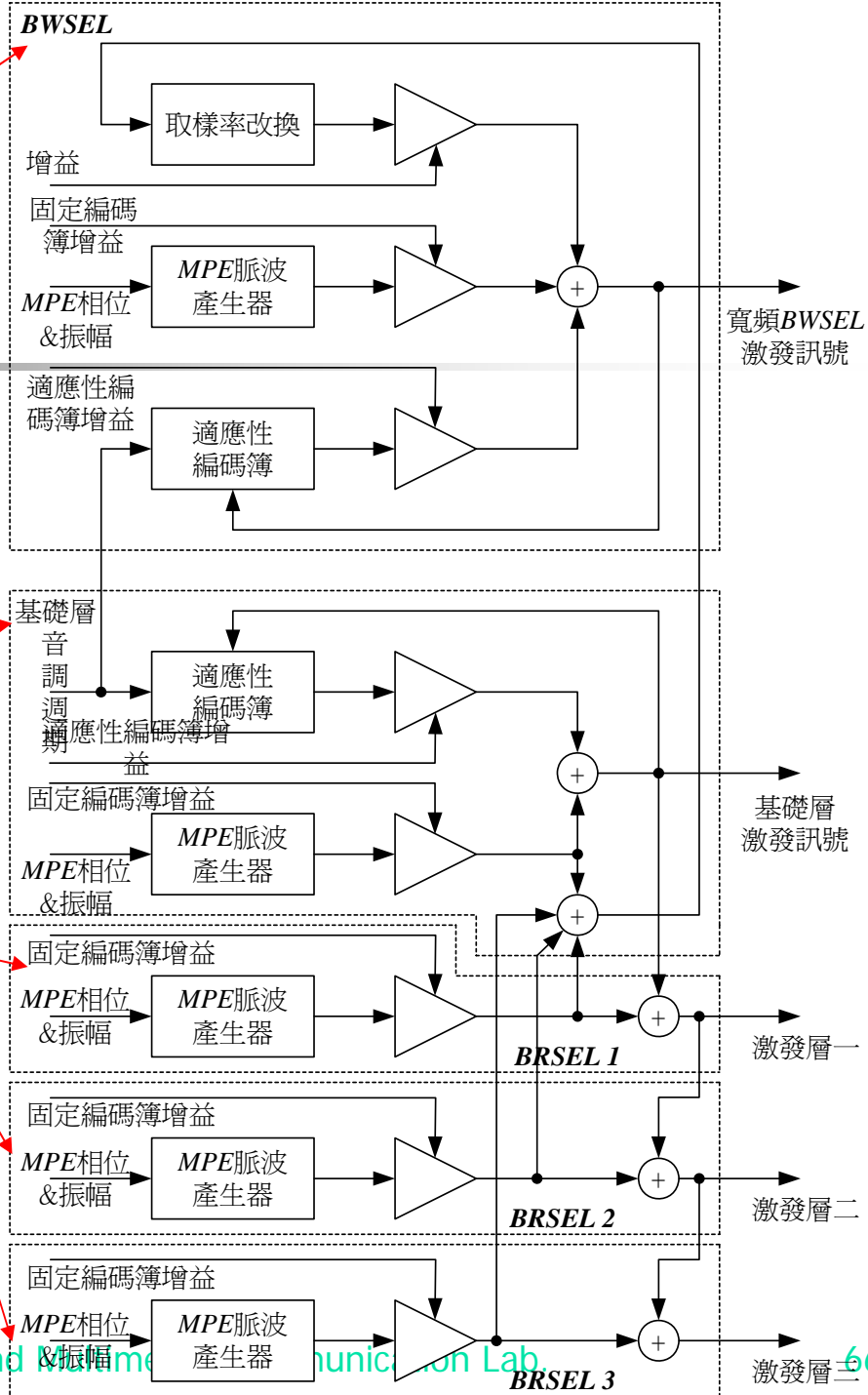


MPEG-4 speech CODEC

- Coding of LPC coefficients
 - All three structures use two-stage split-VQ
 - Quantizing and coding of LPC coefficients in LSP (line spectrum pairs) domain
- Coding of MPE excitation signals
 - Multi-mode multi-pulse excitation
 - Narrow band and broad band coding
 - Bit rate and bandwidth scalability

MPE arch.

- BWSEL (bandwidth scalable enhancement layer)
- Base layer
 - Same as ordinary CELP encoder
- Three BRSEL (bit rate scalable layer)





MPE

- Base layer operates with sampling rates 8 KHz or 16 KHz
- Above base layer, either of two enhancement layers is possible.
- If BRSEL is used, extra excitation pulses are added.
 - Sample rate is kept the same as base layer
- If BWSEL is used, with or without BRSEL(s), the output by BRSEL is always 16 KHz
 - Sample rate for base layer and BRSEL is 8 KHz



RPE

- Regular pulse excitation
- Computation complexity for wide band coding is about $\frac{1}{2}$ of MPE
- Least work in decoding of excitation signals

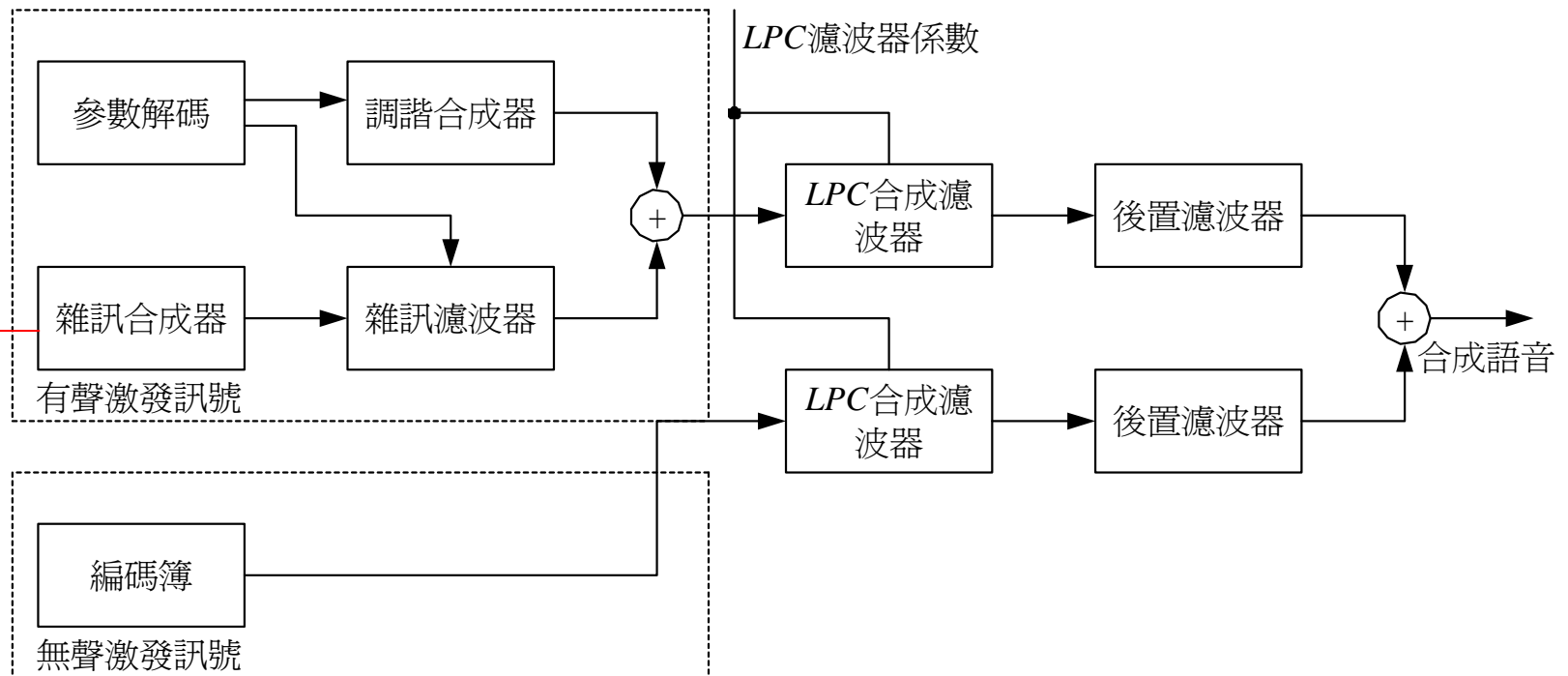


HVXC

- Harmonic vector excitation coding
- Very low bit rate coding
 - 2 Kbps
- Represent input speech by parameters
- In decoding, speed and tone can be changed easily

HVXC - decoder

- Noise is added to make speech more nature





Demo



Audio coding

- Basics
 - Psychoacoustic
 - Subband coding
- MPEG audio
- AC-3



Digital audio

| | Frequency Band (Hz) | Sampling Rate (kHz) | Bits per Sample | Raw Bitrate (kbits/s) |
|------------------|---------------------|---------------------|-----------------|-----------------------|
| Telephone Speech | 300~3400 | 8 | 8 | 64 |
| Wideband Speech | 50~7000 | 16 | 8 | 128 |
| Mediumband Audio | 10~11000 | 24 | 16 | 384 |
| Wideband Audio | 10~22000 | 48 | 16 | 768 |

– CD: $44.1 \text{ kHz} \times 16 \text{ bits} \times 2 \text{ channels} = 1.411 \text{ Mbits/s}$



Psychoacoustics

- The range of human hearing is about 20 Hz to about 20 kHz
- The frequency range of the voice is typically only from about 500 Hz to 4 kHz
- The dynamic range, the ratio of the maximum sound amplitude to the quietest sound that humans can hear, is on the order of about 120 dB



Equal-Loudness Relations

- **Fletcher-Munson Curves**

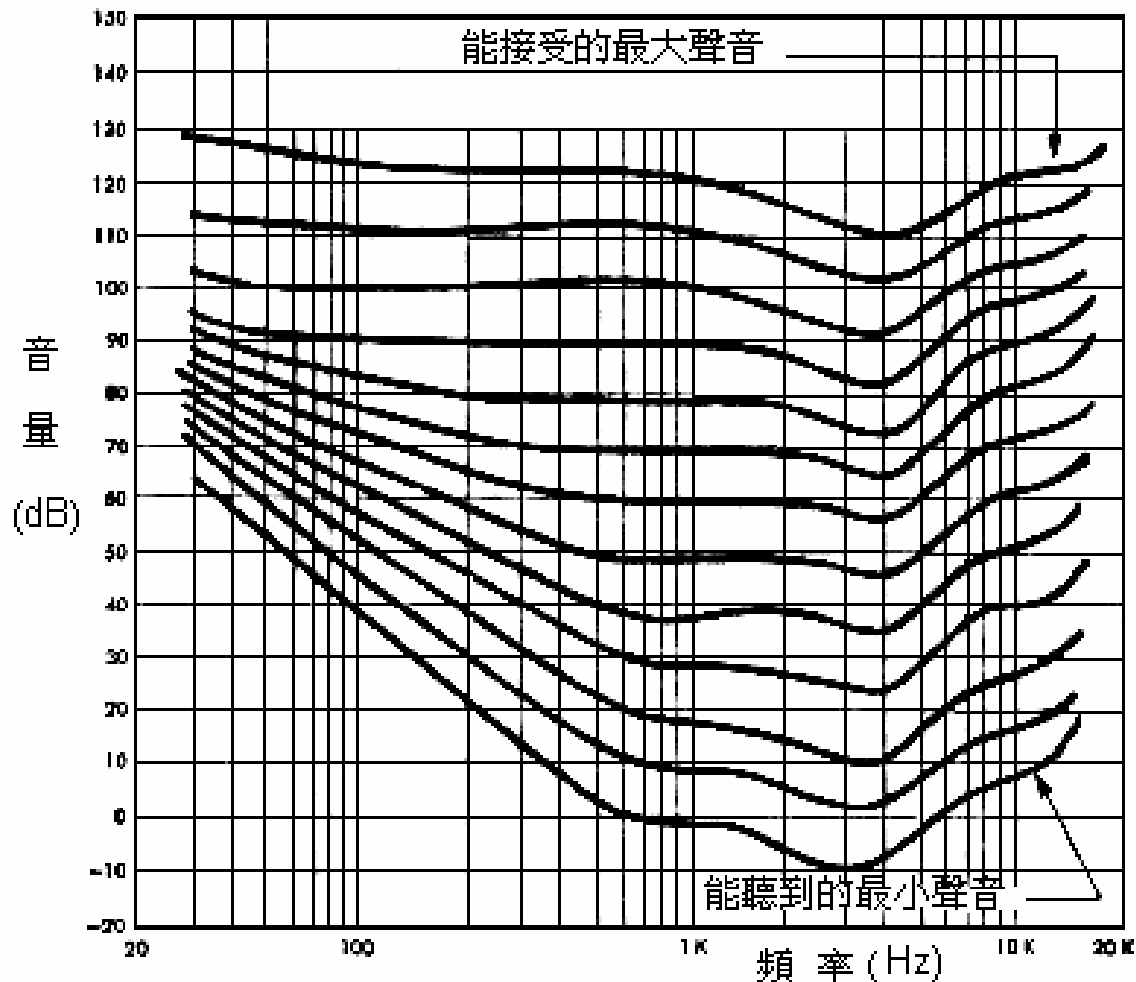
- Equal loudness curves that display the relationship between perceived loudness ("Phons", in dB) for a given stimulus sound volume ("Sound Pressure Level", also in dB), as a function of frequency



Equal-Loudness Relations (cont'd)

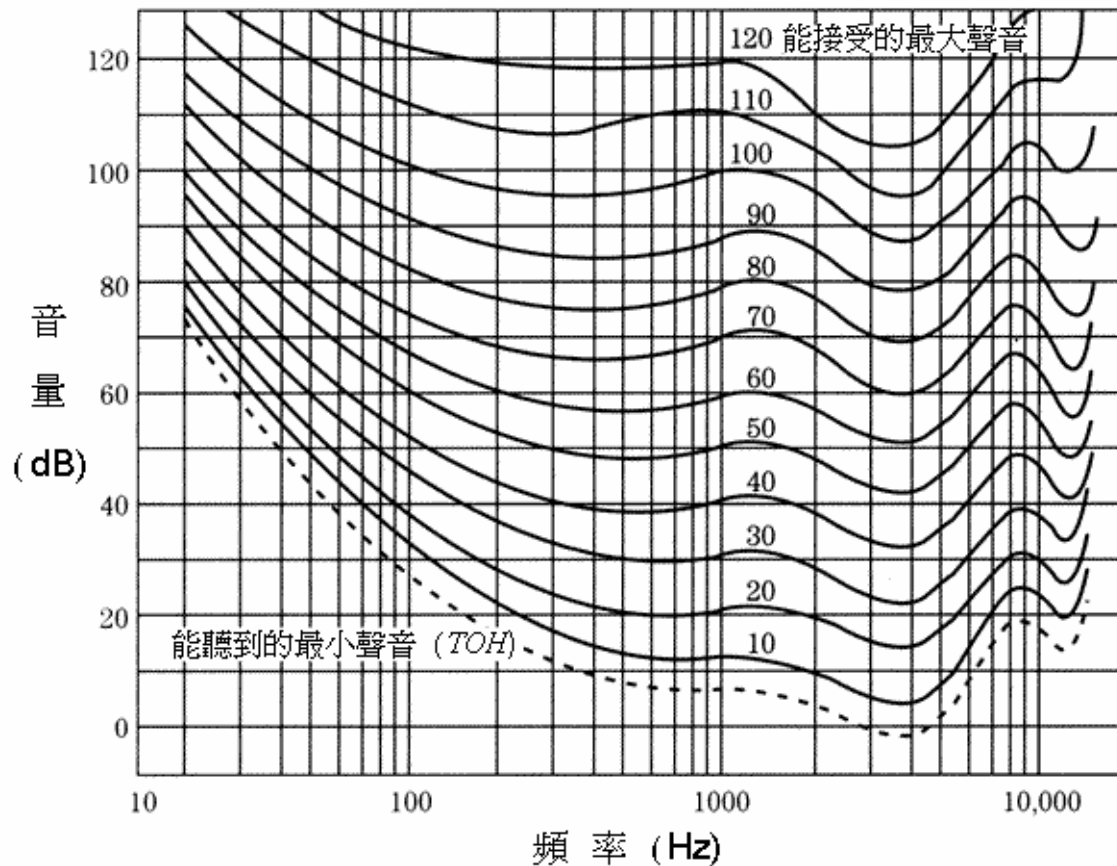
- Next slide shows the ear's perception of equal loudness:
 - The bottom curve shows what level of pure tone stimulus is required to produce the perception of a 10 dB sound
 - All the curves are arranged so that the perceived loudness level gives the same loudness as for that loudness level of a pure tone at 1 kHz

Fletcher-Munson Curves



Robinson-Dadson Curves

- ISO 226 in 1986





Frequency Masking

- Lossy audio data compression methods, such as MPEG/Audio encoding, remove some sounds which are masked anyway

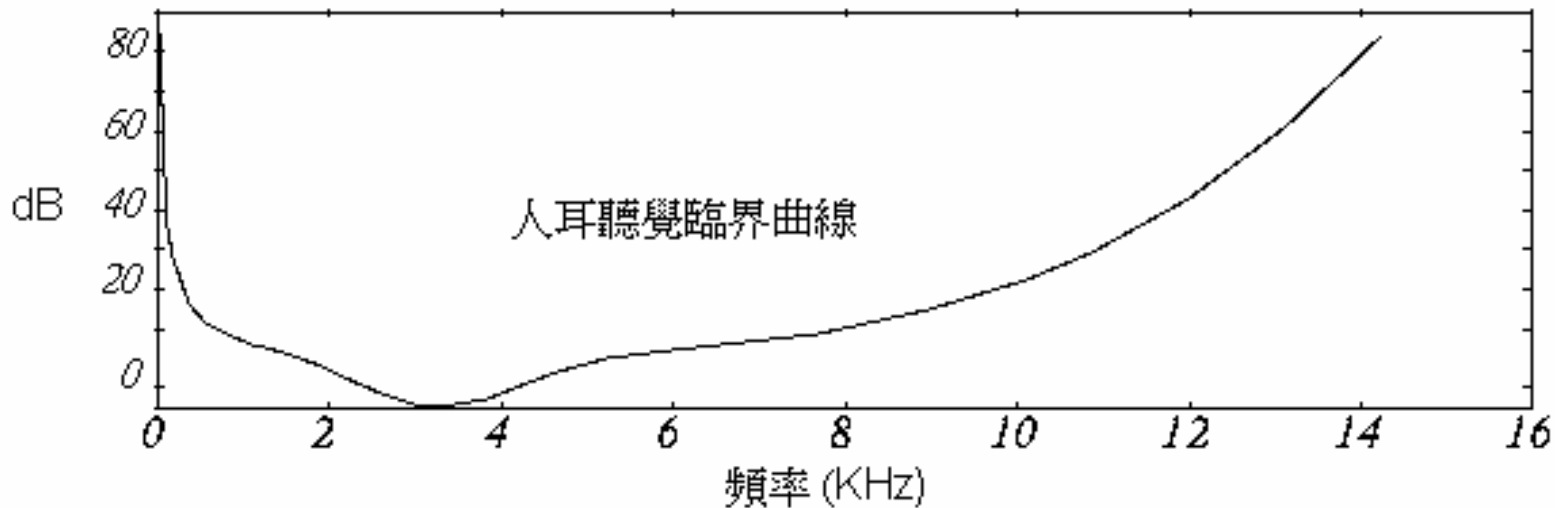


Frequency Masking_(cont'd)

- The general situation in regard to masking is as follows:
 - A lower tone can effectively mask (make us unable to hear) a higher tone
 - The reverse is not true - a higher tone does not mask a lower tone well
 - The greater the power in the masking tone, the wider is its influence - the broader the range of frequencies it can mask.
 - As a consequence, if two tones are widely separated in frequency then little masking occurs

Threshold of Hearing

- Pure tone





Threshold of Hearing (cont'd)

- The threshold of hearing curve: if a sound is above the dB level shown then the sound is audible
- Turning up a tone so that it equals or surpasses the curve means that we can then distinguish the sound



Threshold of Hearing (cont'd)

- An approximate formula exists for this curve:

$$TOH(f) = 3.64(f / 1000)^{-0.8} - 6.5e^{(f / 1000 - 3.3)^2} + 10^{-3} (f / 1000)^4 \quad (5.1)$$

- The threshold units are dB; the frequency for the origin (0,0) in formula (5.1) is 2,000 Hz: $TOH(\hat{f}) = 0$ at $f = 2$ kHz

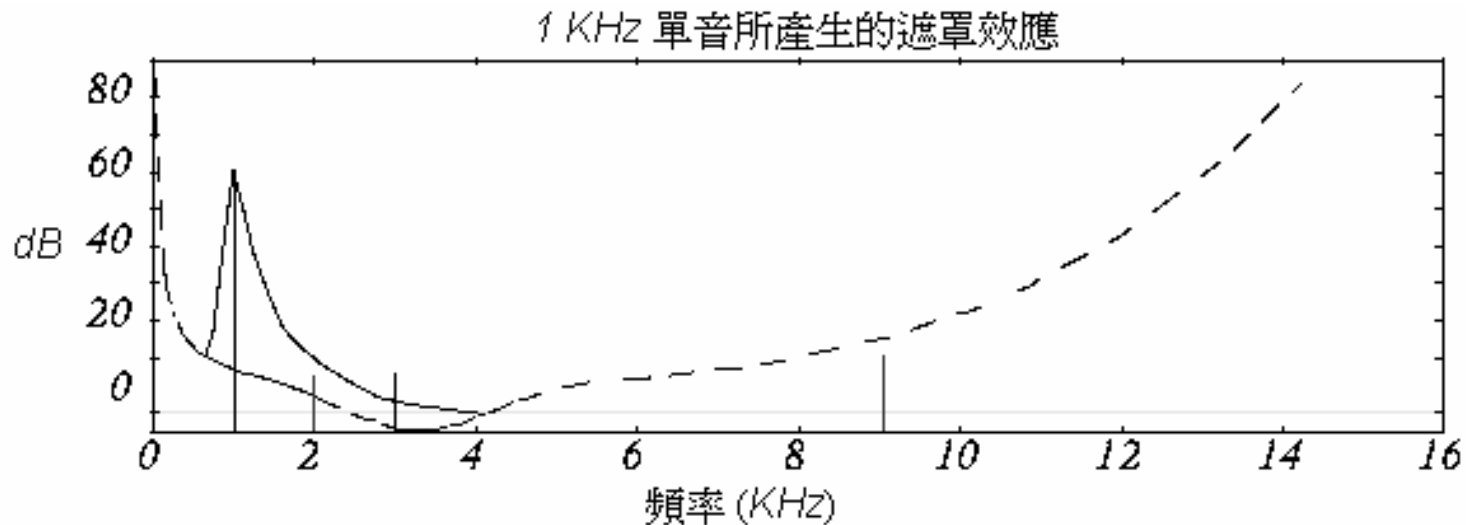


Frequency Masking Curves

- Frequency masking is studied by playing a particular pure tone, say 1 kHz again, at a loud volume, and determining how this tone affects our ability to hear tones nearby in frequency
 - one would generate a 1 kHz **masking tone**, at a fixed sound level of 60 dB, and then raise the level of a nearby tone, e.g., 1.1 kHz, until it is just audible

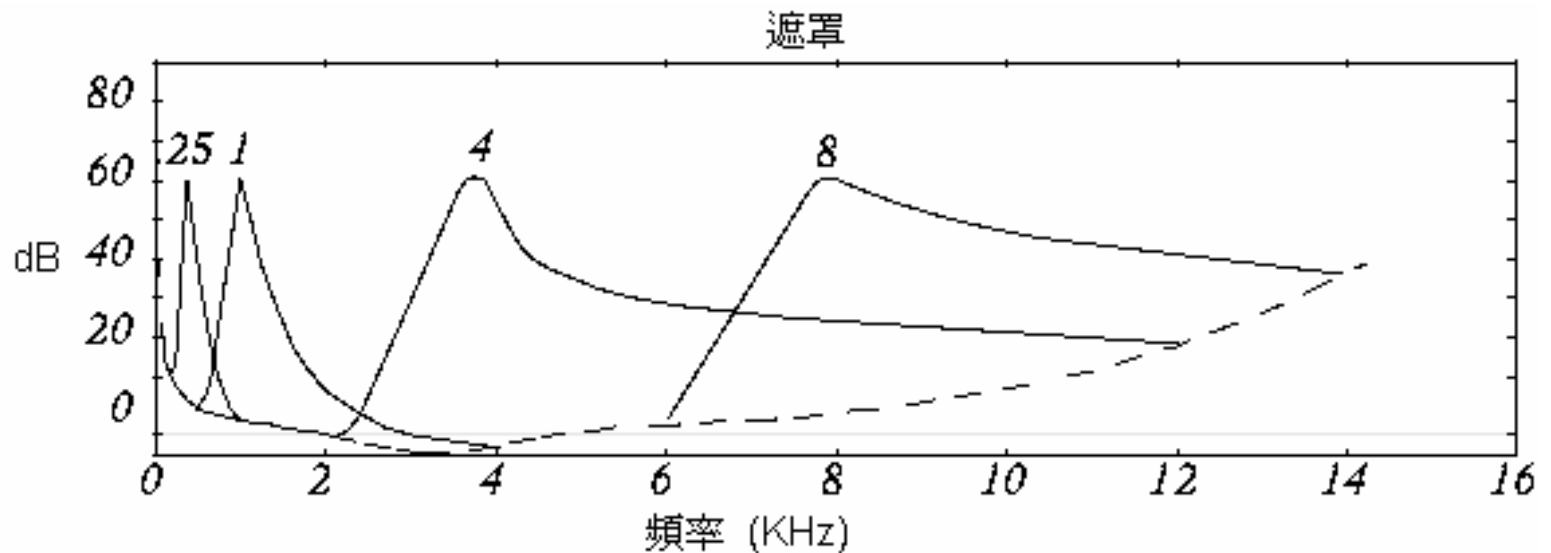
Frequency Masking Curves

- The audible level for a single masking tone (1 kHz)



Frequency Masking Curves

- The plot changes if other masking tones are used





Critical Bands

- **Critical bandwidth** represents the ear's resolving power for simultaneous tones or partials
 - At the low-frequency end, a critical band is less than 100 Hz wide, while for high frequencies the width can be greater than 4 kHz



Critical Bands (cont'd)

- Experiments indicate that the critical bandwidth:
 - for masking frequencies < 500 Hz: remains approximately constant in width (about 100 Hz)
 - for masking frequencies > 500 Hz: increases approximately linearly with frequency



25-Critical Bands and Bandwidth

| Band # | Lower Bound (Hz) | Center (Hz) | Upper Bound (Hz) | Bandwidth (Hz) |
|--------|------------------|-------------|------------------|----------------|
| 1 | - | 50 | 100 | - |
| 2 | 100 | 150 | 200 | 100 |
| 3 | 200 | 250 | 300 | 100 |
| 4 | 300 | 350 | 400 | 100 |
| 5 | 400 | 450 | 510 | 110 |
| 6 | 510 | 570 | 630 | 120 |
| 7 | 630 | 700 | 770 | 140 |
| 8 | 770 | 840 | 920 | 150 |
| 9 | 920 | 1000 | 1080 | 160 |
| 10 | 1080 | 1170 | 1270 | 190 |
| 11 | 1270 | 1370 | 1480 | 210 |
| 12 | 1480 | 1600 | 1720 | 240 |

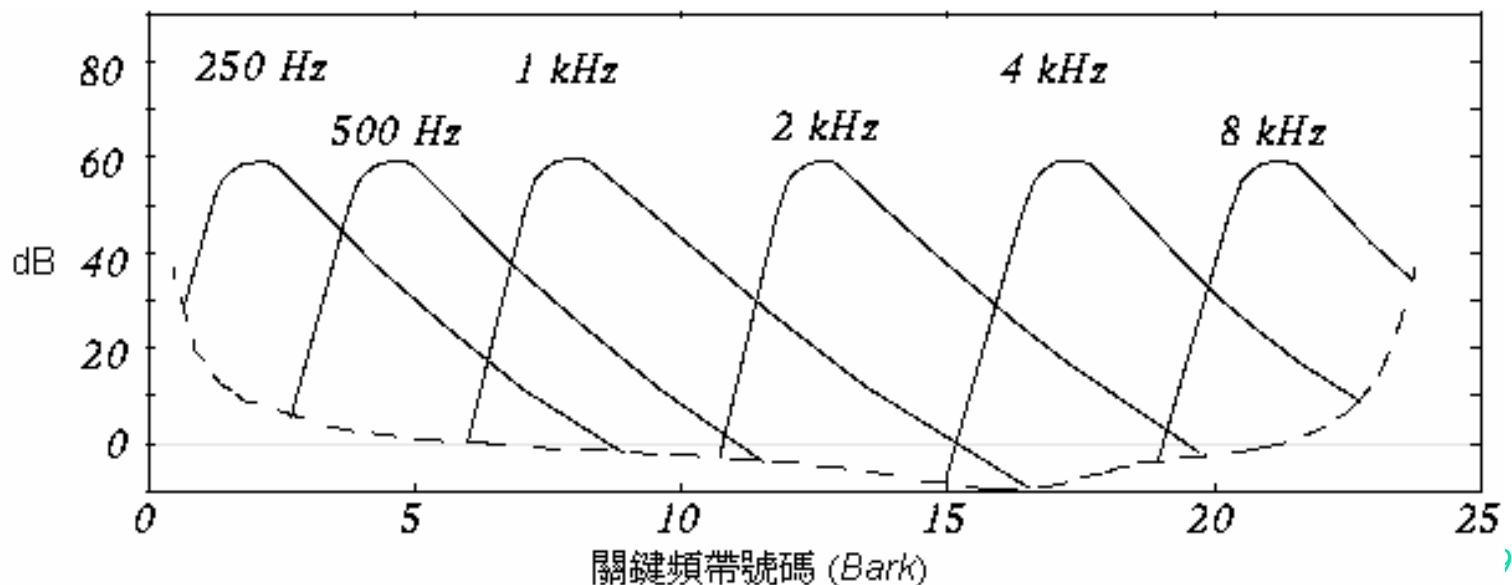


25-Critical Bands and Bandwidth

| Band # | Lower Bound (Hz) | Center (Hz) | Upper Bound (Hz) | Bandwidth (Hz) |
|--------|------------------|-------------|------------------|----------------|
| 13 | 1720 | 1850 | 2000 | 280 |
| 14 | 2000 | 2150 | 2320 | 320 |
| 15 | 2320 | 2500 | 2700 | 380 |
| 16 | 2700 | 2900 | 3150 | 450 |
| 17 | 3150 | 3400 | 3700 | 550 |
| 18 | 3700 | 4000 | 4400 | 700 |
| 19 | 4400 | 4800 | 5300 | 900 |
| 20 | 5300 | 5800 | 6400 | 1100 |
| 21 | 6400 | 7000 | 7700 | 1300 |
| 22 | 7700 | 8500 | 9500 | 1800 |
| 23 | 9500 | 10500 | 12000 | 2500 |
| 24 | 12000 | 13500 | 15500 | 3500 |
| 25 | 15500 | 18775 | 22050 | 6550 |

Bark Unit

- **Bark unit** is defined as the width of one critical band, for any masking frequency
- The idea of the Bark unit : every critical band width is roughly equal in terms of Barks





Conversion: Frequency & Critical Band Number

- Conversion expressed in the Bark unit:
- Critical band number (Bark) =

$$\begin{cases} f / 100, & f < 500 \\ 9 + 4 \log_2(f / 1000), & f \geq 500 \end{cases}$$

- Another formula used for the Bark scale:

$$b = 13.0 \tan^{-1}(0.76f) + 3.5 \tan^{-1}(f^2 / 56.25)$$

where f is in KHz and b is in Bark number



Conversion: Frequency & Critical Band Number

- The inverse equation:

$$f = \left[\frac{e^{0.219 \times b}}{352} + 0.1 \right] \times b - 0.032 \times e^{-0.15 \times (b-5)^2}$$

- The critical bandwidth (df) for a given center frequency f can also be approximated by:

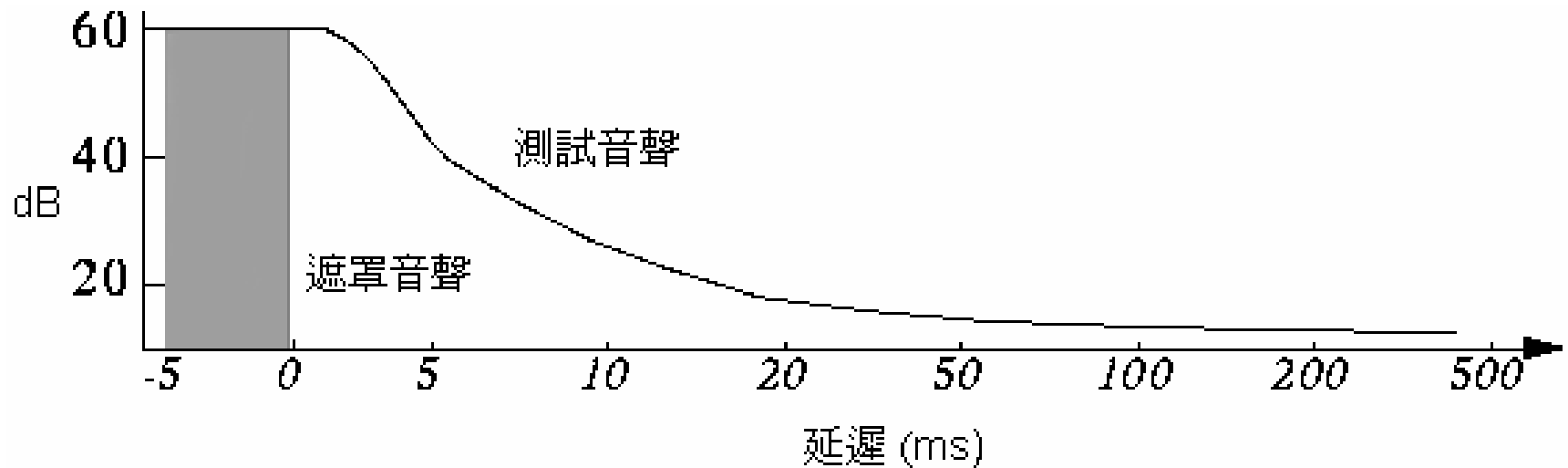
$$df = 25 + 75 \times [1 + 1.4 \times f^2]^{0.69}$$



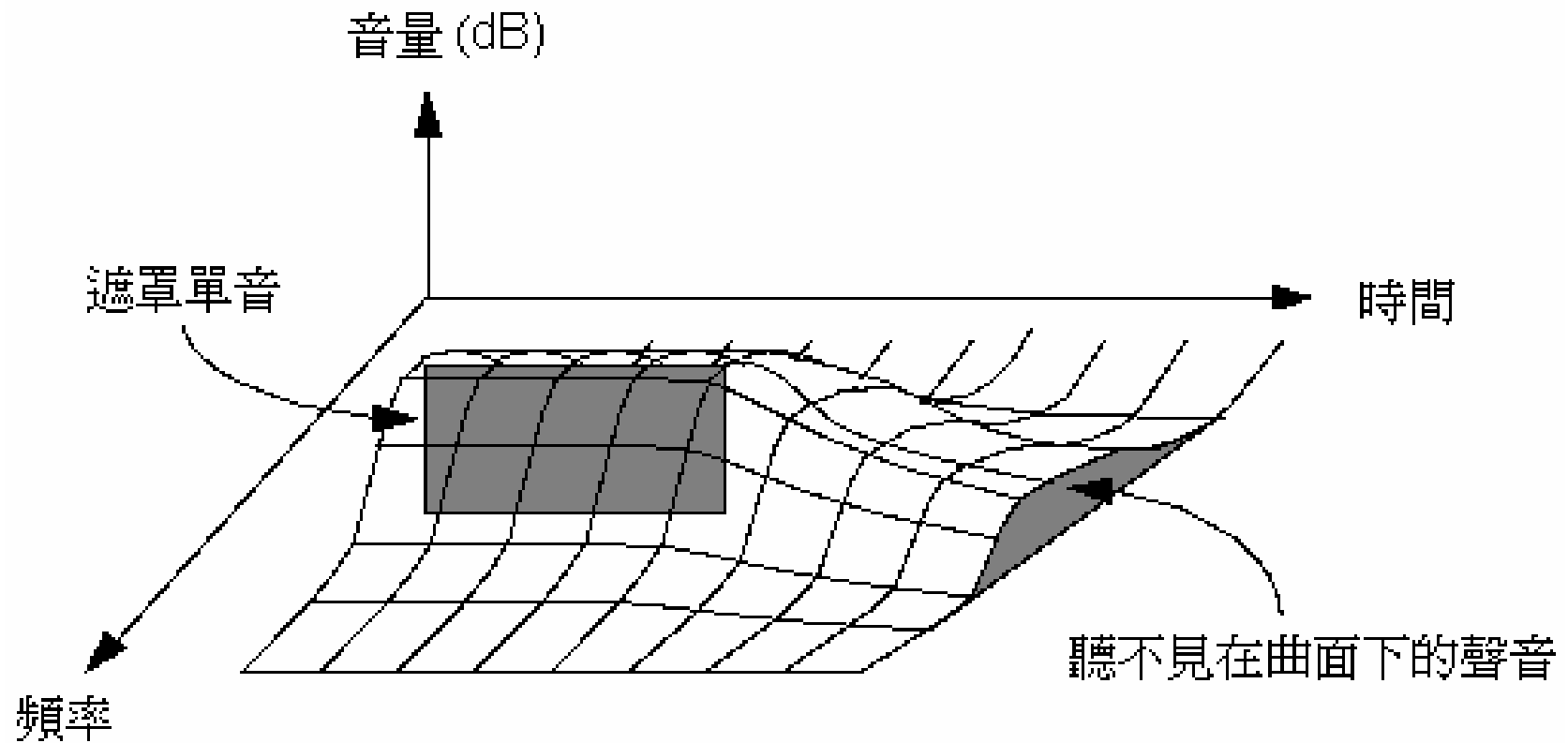
Temporal Masking

- **Phenomenon**: any loud tone will cause the hearing receptors in the inner ear to become **saturated** and require time to recover
- The following figures show the results of Masking experiments:

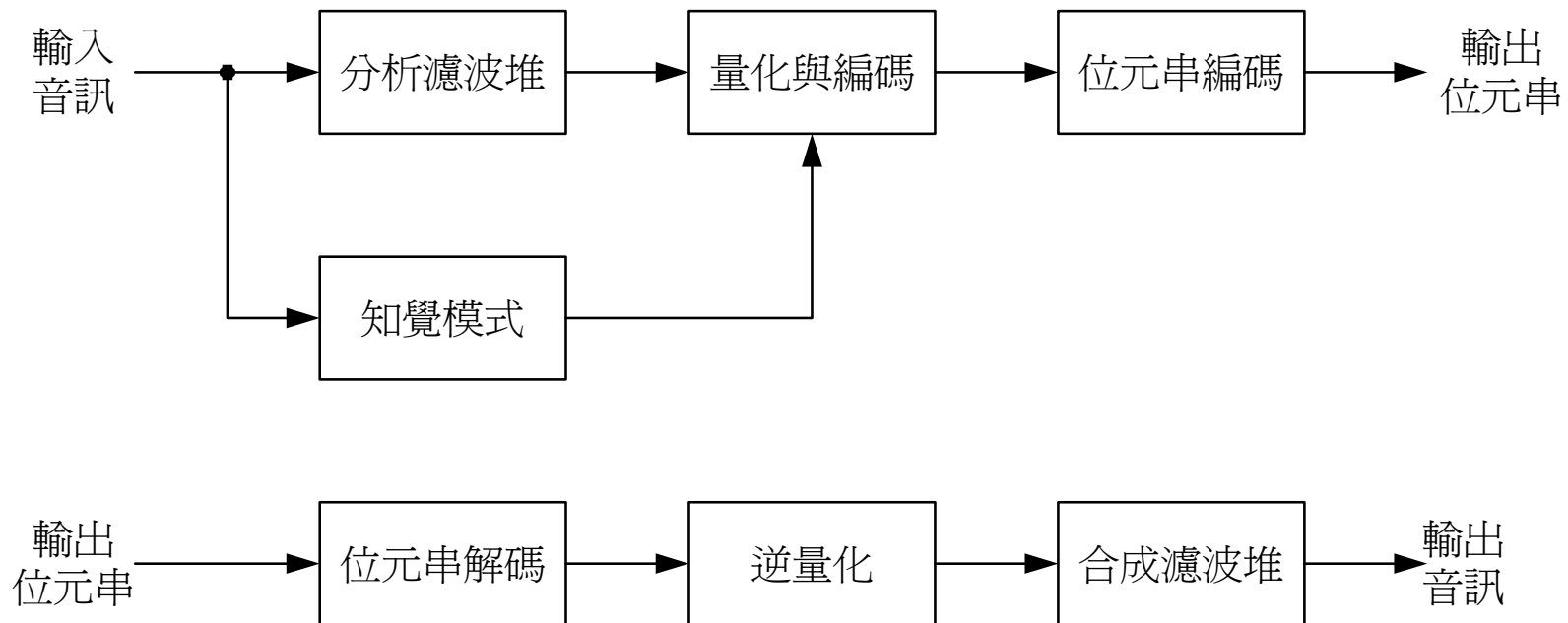
Temporal Masking



Temporal & Frequency Masking



Subband coding and audio coding





Audio coding standards

| 推出 年份 | 位元率 (<i>Kb/s</i>) | 描述 | 音質極優 之位元率 (<i>Kb/s</i>) |
|----------|------------------------|----------------------|------------------------------|
| 1991 | 32~448 | <i>MPEG-1 Layer1</i> | 每聲道 192 <i>Kb/s</i> |
| 1991 | 32~384 | <i>MPEG-1 Layer2</i> | 每聲道 128 <i>Kb/s</i> |
| 1993 | 32~320 | <i>MPEG-1 Layer3</i> | 每聲道 96 <i>Kb/s</i> |
| 1991 | 32~640 | <i>AC-3</i> | 5.1 聲道 384 <i>Kb/s</i> |
| 1997 | 384 | <i>MPEG-2 AAC</i> | 每聲道 64 <i>Kb/s</i> |
| 1999 | 64 | <i>MPEG-4</i> | 每聲道 64 <i>Kb/s</i> |



MPEG-1 audio

- ISO/IEC 11172-3
- First high quality audio compression standards
- Sampling rates : 32, 44.1, or 48 KHz
- CD quality two-channel audio at ~256 Kbps
- Quality demonstration (Layer III)



MPEG-1 audio

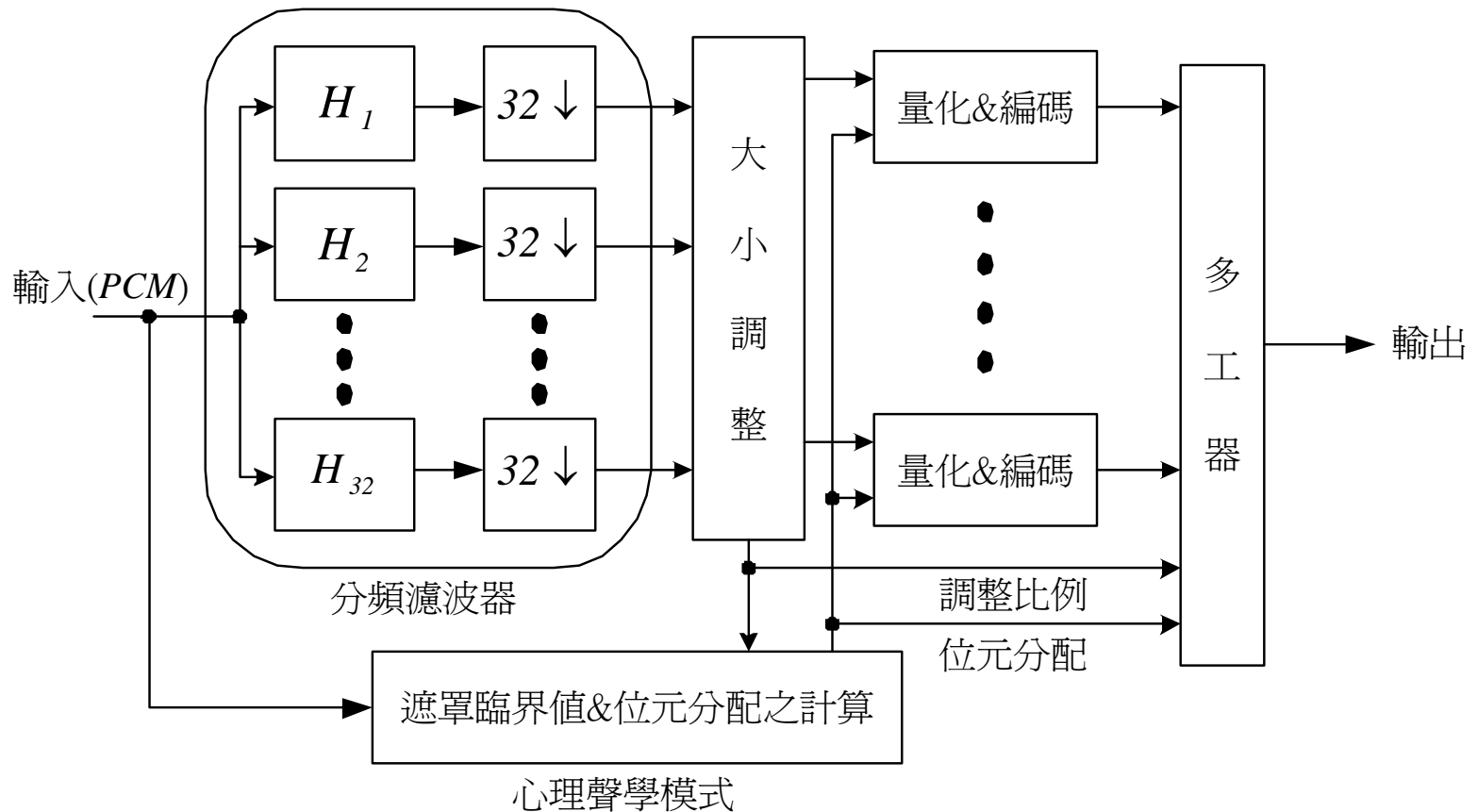
- **MPEG audio compression** takes advantage of psychoacoustic models, constructing a large multi-dimensional lookup table to transmit masked frequency components using fewer bits



MPEG-1 audio

- **MPEG Audio Overview**
 - Applies a filter bank to the input to break it into its frequency components
 - In parallel, a psychoacoustic model is applied to the data for bit allocation block
 - The number of bits allocated are used to quantize the info from the filter bank - providing the compression

Encoder block diagram





Filter bank

- $H_i(\omega)$'s are all obtained from $H_1(\omega)$, which is the impulse response of the low pass filter, by shifting in frequency domain
- 32 samples in, 32 samples out
- It was shown that $s(i) = \sum_{n=0}^{511} x(t-n)H_i(n)$, where

$$H_i(n) = h(n) \cos \left[\frac{(2i+1)(n-16)\pi}{64} \right]$$

can be calculated by 32-point IDCT.

- See text for details



Scaling

- Normalizing each subband output
- Every 12 output values are normalized by the maximal absolute value in them

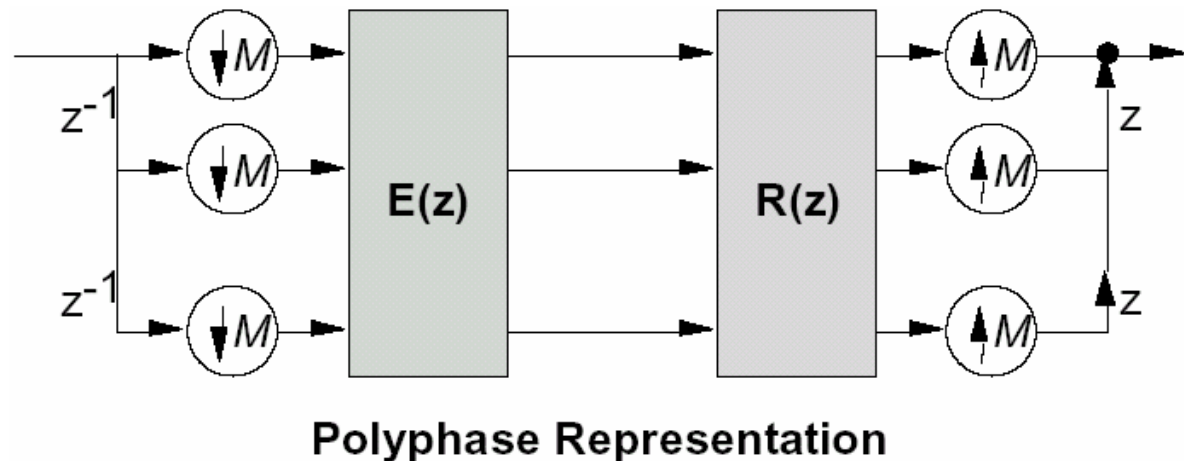
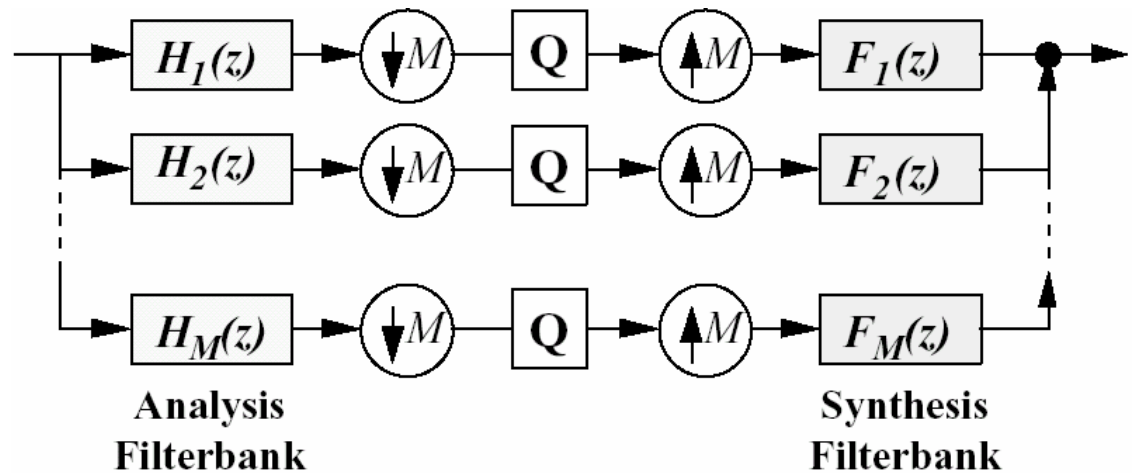


Quantization & coding

- Psychoacoustics
 - Masking thresholds
 - 512- or 1024- points FFT
 - Bit allocation
- Note that Psychoacoustics model is not part of the standard,
 - So, ...
 - So, bit allocation information should be sent to the decoder

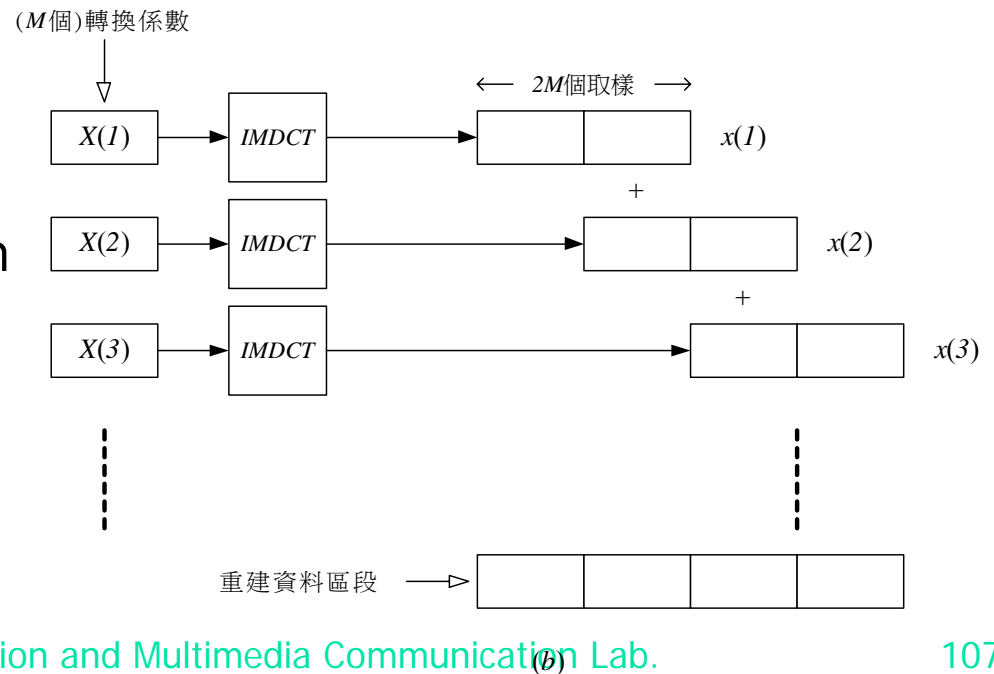
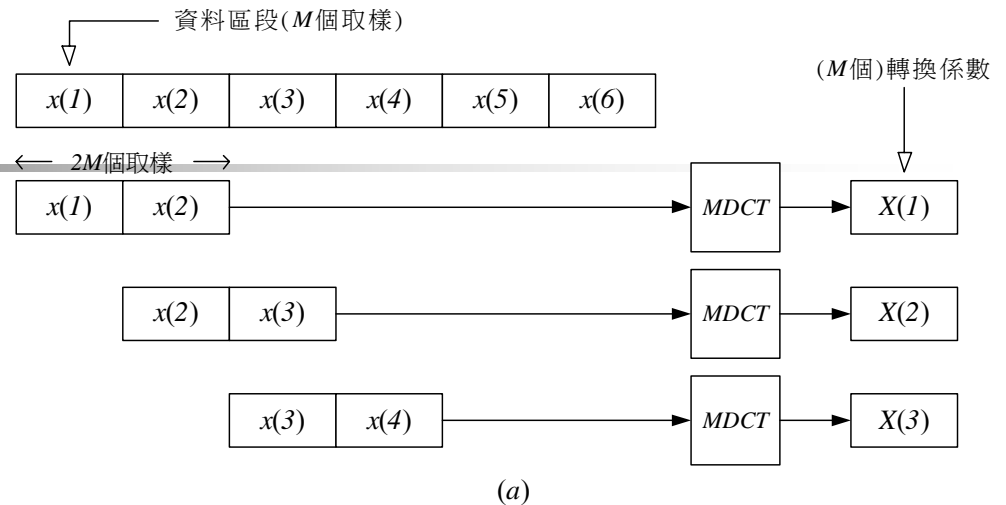
Subband coding and DCT

- When $E(z) = \text{DCT}$ matrix, this becomes DCT
 - No overlap
 - Blocking artifacts



Implemented by MDCT

- QMF cannot divide input into 32 subbands directly
- MDCT (modified DCT)
 - Overlapped, windowed
 - Sine window
- Layer III
 - (a) overlapped transform
 - Analysis
 - (b) inverse transform
 - IMDCT
 - Synthesis
- 50% overlap : less blocking artifacts





Layers

- Increasing complexity, delay, quality
 - Layer 1 : 384 Kbps for perceptually lossless quality (4:1)
 - Layer 2 : 192 Kbps for perceptually lossless quality (8:1)
 - Layer 3 : 128 Kbps for perceptually lossless quality (12:1)
(for two channels)



MPEG Layers

- MPEG audio covers three compatible **layers** :
 - Each succeeding layer able to understand the lower layers
 - Each succeeding layer offering more complexity in the psychoacoustic model and better compression for a given level of audio quality
 - each succeeding layer, with increased compression effectiveness, accompanied by extra delay



MPEG Layers (cont'd)

- The objective of MPEG layers: a good tradeoff between quality and bit-rate
- Layer 1 quality can be quite good provided a comparatively high bit-rate is available
 - Digital Audio Tape typically uses Layer 1 at around 192 kbps



MPEG Layers (cont'd)

- Layer 2 has more complexity; was proposed for use in Digital Audio Broadcasting
- Layer 3 (MP3) is most complex, and was originally aimed at audio transmission over ISDN lines
- Most of the complexity increase is at the encoder, not the decoder - accounting for the popularity of MP3 players



MPEG Audio Strategy

- **MPEG approach to compression** relies on:
 - Quantization
 - Human auditory system is not accurate within the width of a critical band (perceived loudness and audibility of a frequency)



MPEG Audio Strategy (cont'd)

- **MPEG encoder** employs a bank of filters to:
 - Analyze the frequency ("spectral") components of the audio signal by calculating a frequency transform of a window of signal values
 - Decompose the signal into subbands by using a bank of filters
 - Layer 1 & 2: "quadrature-mirror";
 - Layer 3: adds a DCT;
 - psychoacoustic model: Fourier transform



MPEG Audio Strategy (cont'd)

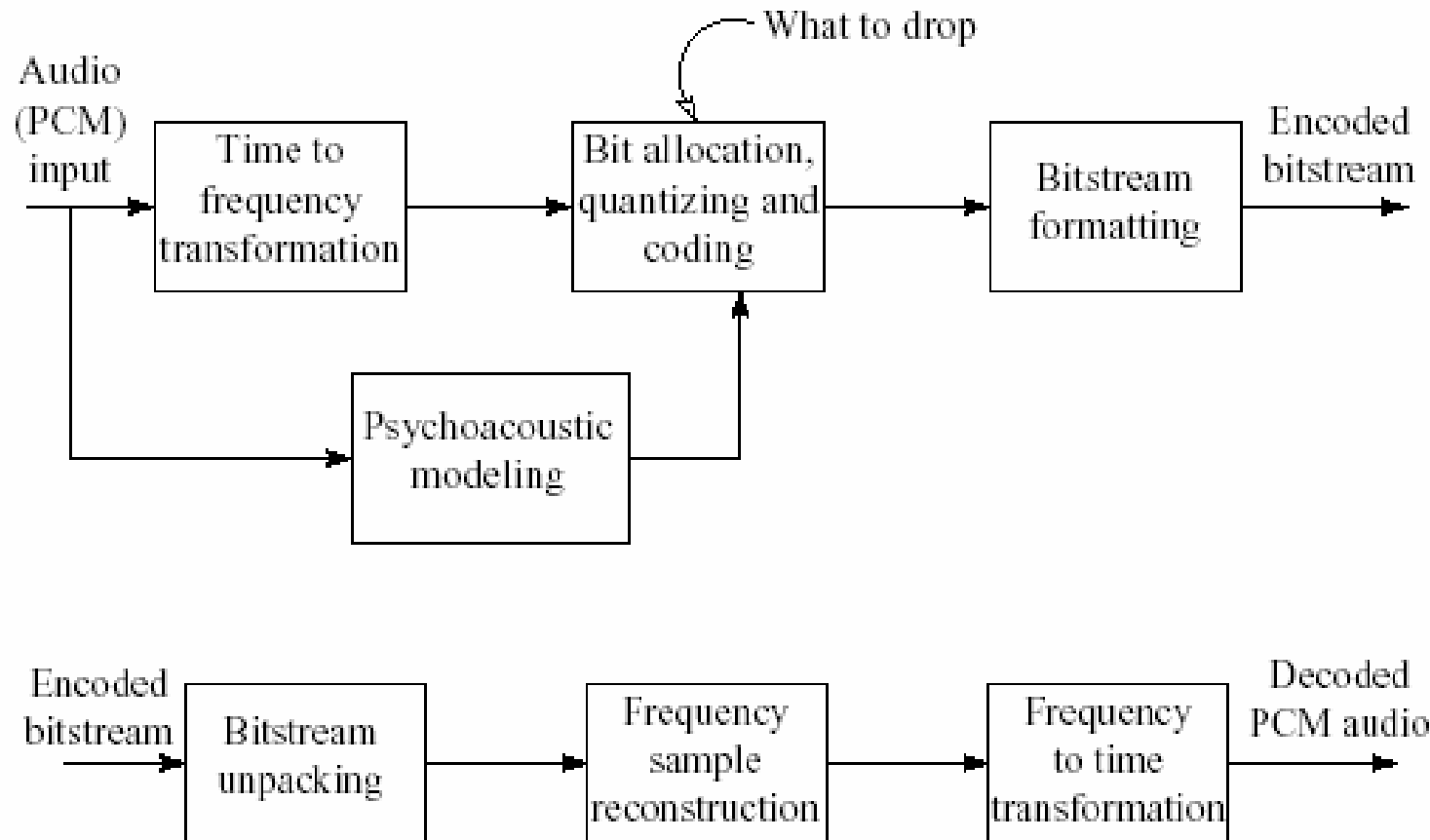
- **Frequency masking**: by using a psychoacoustic model to estimate the just noticeable noise level:
 - Encoder balances the masking behavior and the available number of bits by discarding inaudible frequencies
 - Scaling quantization according to the sound level that is left over, above masking levels



MPEG Audio Strategy (cont'd)

- May take into account the actual width of the critical bands:
 - For practical purposes, audible frequencies are divided into 25 main critical bands
 - To keep simplicity, adopts a *uniform width* for all frequency analysis filters, using 32 overlapping subbands

MPEG Audio Strategy (cont'd)



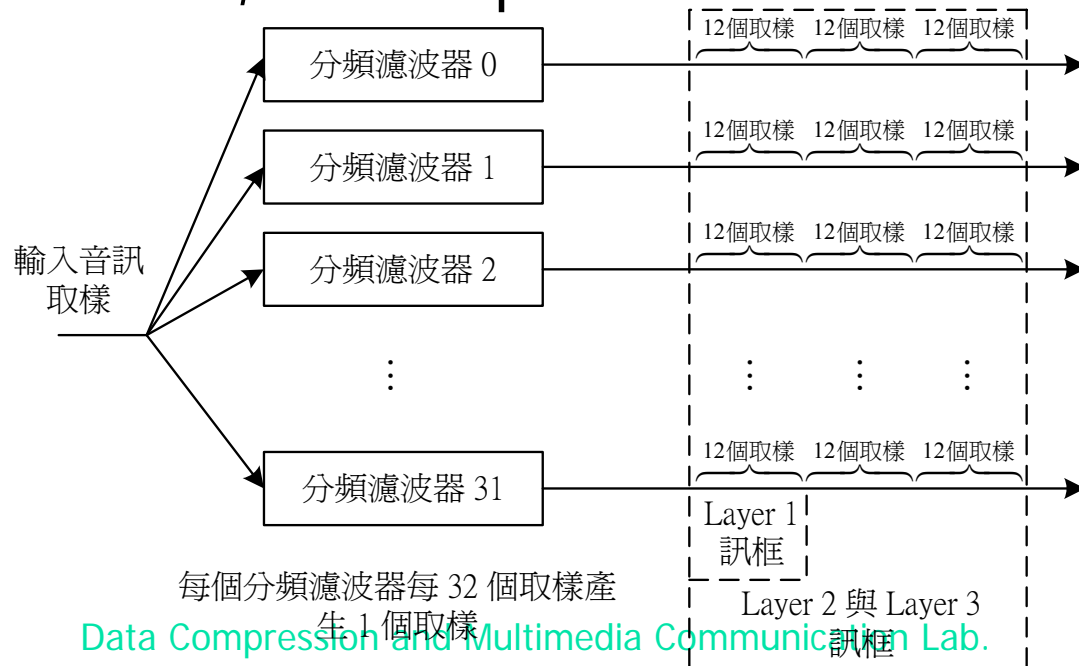


Basic Algorithm (cont'd)

- The algorithm proceeds by dividing the input into 32 frequency subbands, via a filter bank
 - A linear operation taking 32 PCM samples, sampled in time; output is 32 frequency coefficients
- In the Layer 1 encoder, the sets of 32 PCM values are first assembled into a set of 12 groups of 32s
 - an inherent time lag in the coder, equal to the time to accumulate 384 (i.e., 12×32) samples

Basic Algorithm (cont'd)

- How samples are organized
 - A Layer 2 or Layer 3, frame actually accumulates more than 12 samples for each subband: a frame includes 1,152 samples





Bit Allocation Algorithm

- **Aim**: ensure that all of the quantization noise is below the masking thresholds
- **One common scheme**
 - For each subband, the psychoacoustic model calculates the *Signal-to-Mask Ratio* (SMR) in dB
 - Then the "Mask-to-Noise Ratio" (MNR) is defined as the difference

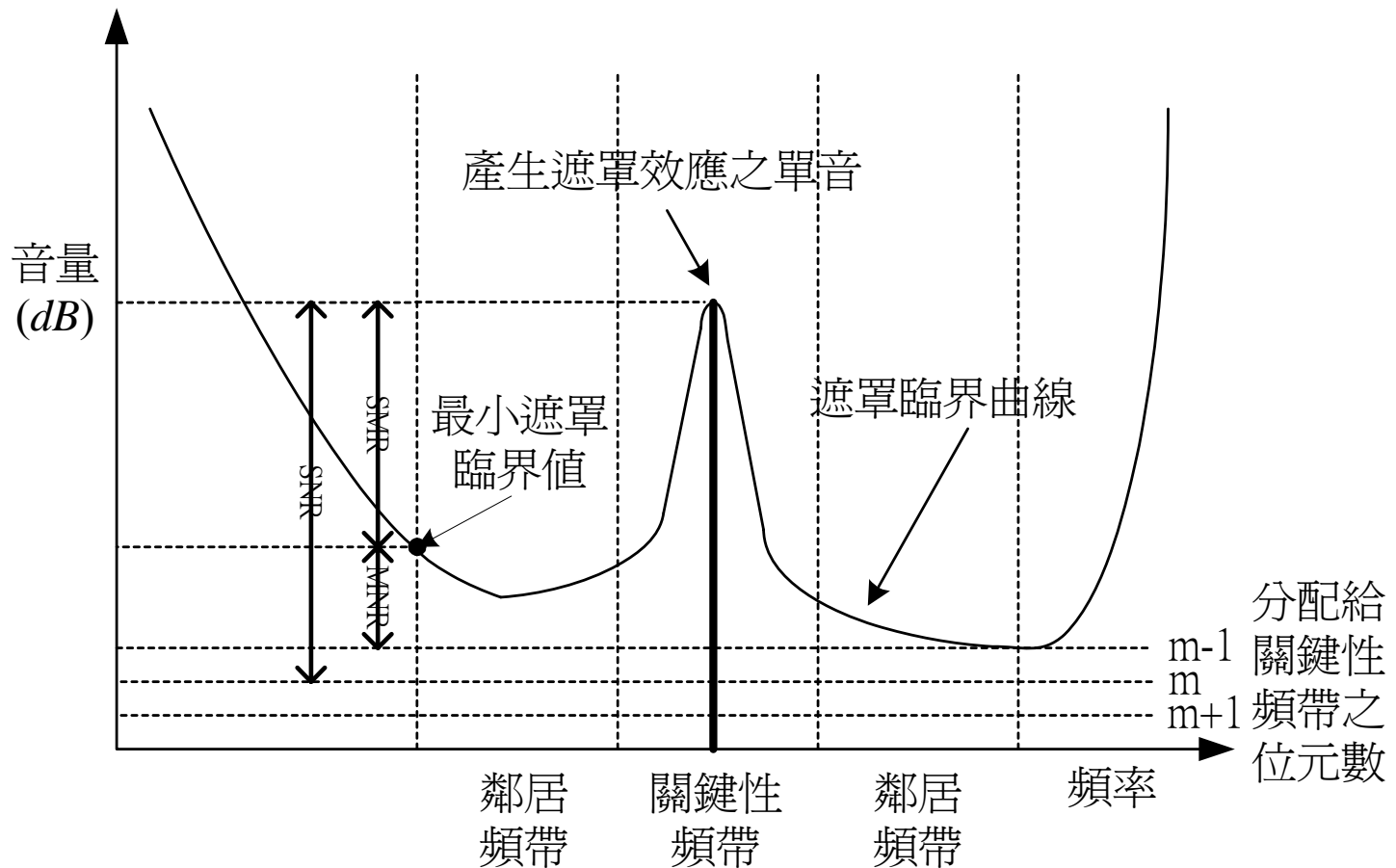
$$\text{MNR}_{\text{dB}} \equiv \text{SNR}_{\text{dB}} - \text{SMR}_{\text{dB}} \quad (14.6)$$



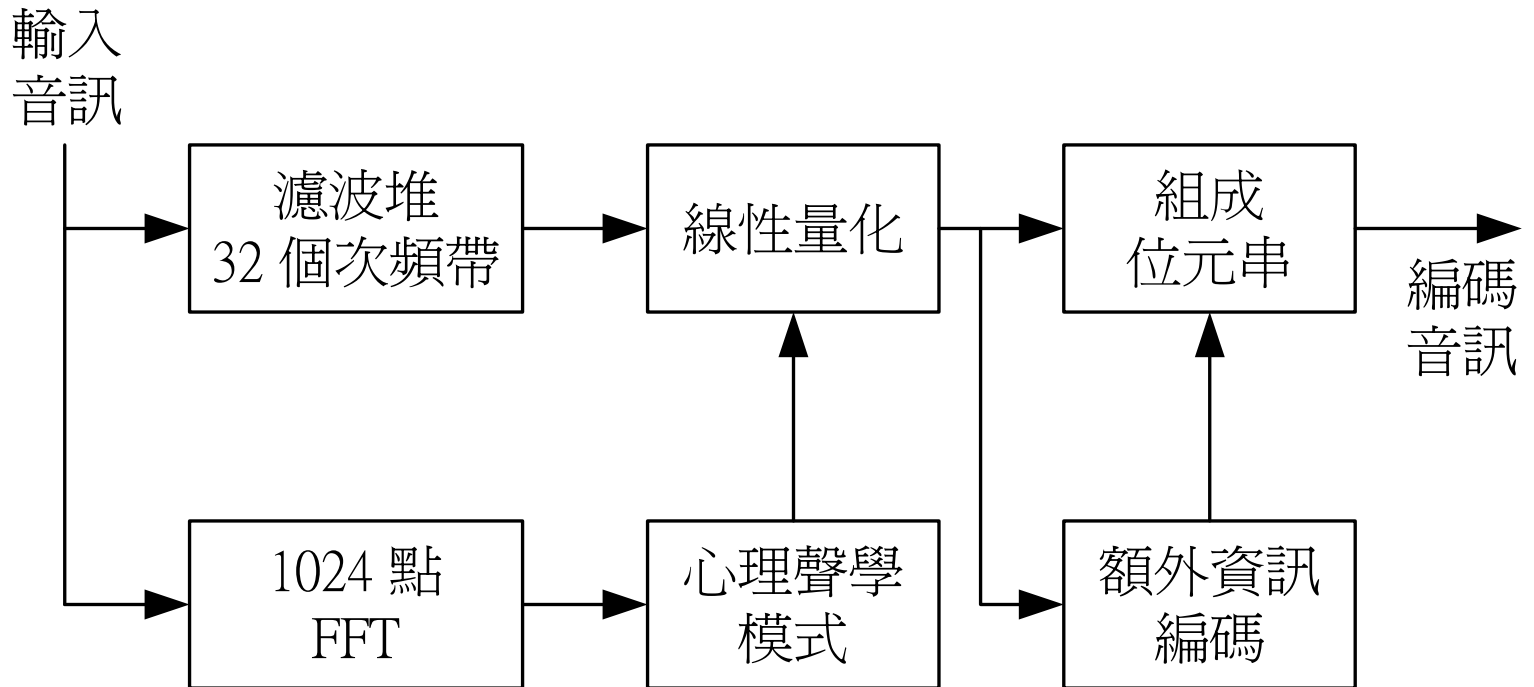
Bit Allocation Algorithm

- **One common scheme** (cont'd)
 - The lowest MNR is determined, and the number of code-bits allocated to this subband is incremented
 - Then a new estimate of the SNR is made, and the process iterates until there are no more bits to allocate

Bit Allocation Algorithm



Layer 1 & 2





Layer 2 of MPEG-1 Audio

- **Main difference:**

- Three groups of 12 samples are encoded in each frame and temporal masking is brought into play, as well as frequency masking
 - Layer 1 uses frequency masking only
- Bit allocation is applied to window lengths of 36 samples instead of 12
- The resolution of the quantizers is increased from 15 bits to 16

- **Advantage:**

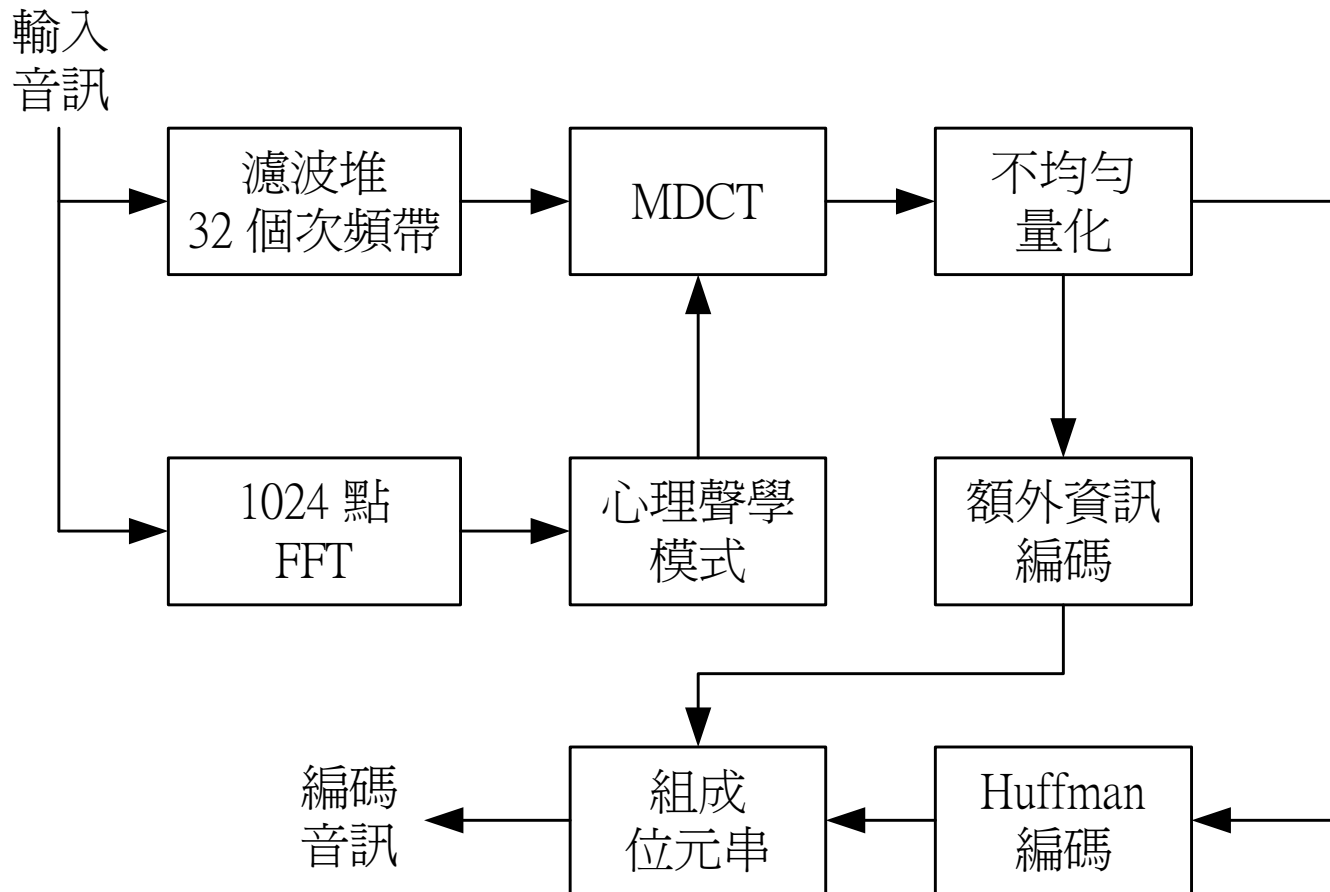
- a single scaling factor can be used for all three groups



Layer 3 of MPEG-1 Audio

- **Main difference:**
 - Employs a similar filter bank to that used in Layer 2, except using a set of filters with non-equal frequencies
 - Takes into account stereo redundancy
 - Uses Modified Discrete Cosine Transform (MDCT) —addresses problems that the DCT has at boundaries of the window used by overlapping frames by 50%

Layer 3 of MPEG-1 Audio





Layer 3 of MPEG-1 Audio

- 在500 Hz以下關鍵性頻帶的頻寬大約都維持在100 Hz左右的固定值；但是，在500 Hz以上的關鍵性頻帶其頻寬大抵隨著頻率的提高線性地變寬。
- 因此，在比較低頻的次頻帶我們需要更高的頻率解析度。
- Layer 3的做法是在最低頻的兩個次頻帶使用18點的MDCT，而在其餘的次頻帶使用6點的MDCT。



Layer 3 of MPEG-1 Audio

- 假設取樣率是 $f_s = 48 \text{ KHz}$ （每秒 48,000 個取樣）。那麼，根據 Nyquist 取樣定理，它所對應的音訊之最高頻為 $f_s/2$ 。
- 因此，對應頻寬會被切分成 32 個等頻寬的次頻帶，也就是每一個次頻帶的頻寬為 $f_s/64$ ，例如 $48,000/64 = 750 \text{ Hz}$ 。
 - 如果使用的是 18 點的 MDCT，那麼原本頻寬為 750 Hz 的次頻帶將再被切分成 18 個次頻帶
 - 這相當於我們的頻率解析度由原本的 750 Hz 變成 $750/18 = 41.67 \text{ Hz}$ 。

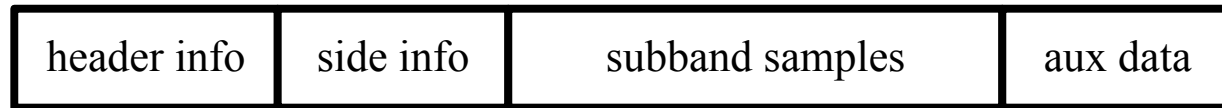


MP3 compression performance

| Sound Quality | Bandwidth | Mode | Compression Ratio |
|------------------------|-----------|--------|-------------------|
| Telephony | 3.0 kHz | Mono | 96:1 |
| Better than Short-wave | 4.5 kHz | Mono | 48:1 |
| Better than AM radio | 7.5 kHz | Mono | 24:1 |
| Similar to FM radio | 11 kHz | Stereo | 26 - 24:1 |
| Near-CD | 15 kHz | Stereo | 16:1 |
| CD | > 15 kHz | Stereo | 14 - 12:1 |



Frame structure



- Header info : Sync bits, system info, CRC
- Side info : bit allocation, scale factor
- Subband samples : 32×12 for layer I and 32×36 for layer II and III
- Packetization : 4-byte header, 184-byte payload



Stereo redundancy coding

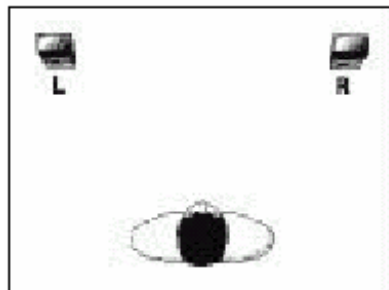
- Four modes : mono, stereo, dual with two separate channel, joint stereo
- Joint stereo mode
 - Human stereo perception $> 2\text{KHz}$ is based on envelop
 - Intensity stereo coding $> 2\text{KHz}$
 - Encode (L+R)
 - Assign independent left- and right- scale factors
- Layer III support (L+R) and (L-R) coding



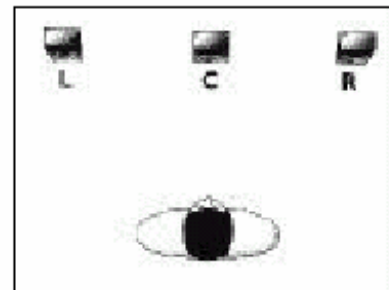
MPEG-2 audio

- ISO/IEC 13818-3
 - Allows lower sampling rates
 - 16, 22.05, and 24 KHz : about half of MPEG-1
 - From wideband speech to mediumband audio
 - Higher frequency resolution
 - Layer I, II, and III
- Multichannel coding
 - 2~5 channels
- Backward compatibility and non-backward compatible coding (13818-7 : MPEG-2 AAC)

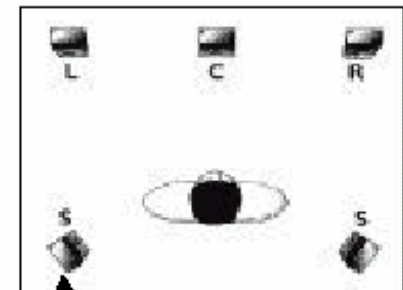
Multichannel audio



2/0-stereo

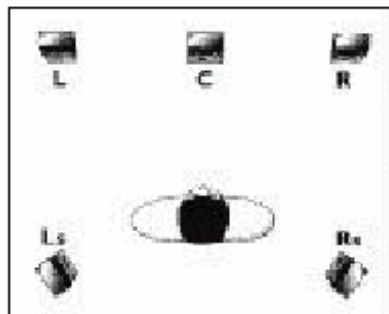


3/0

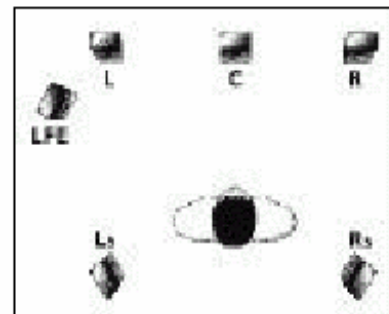


3/1

Surround



3/2



**3/2 with woofer
(5.1 system)**

LFE: Low-frequency enhancement (woofer)

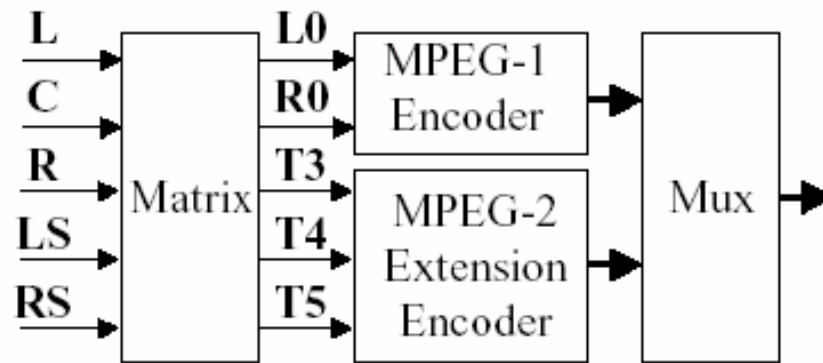
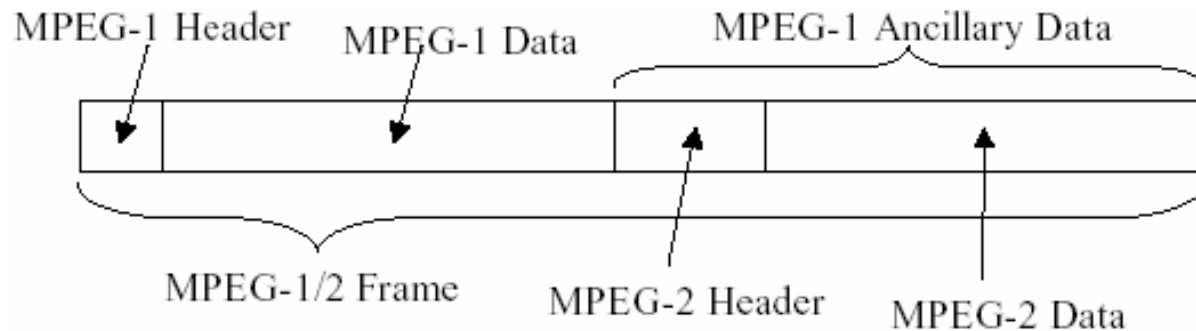
- 15~120 Hz
- Can be anywhere



Compatibility

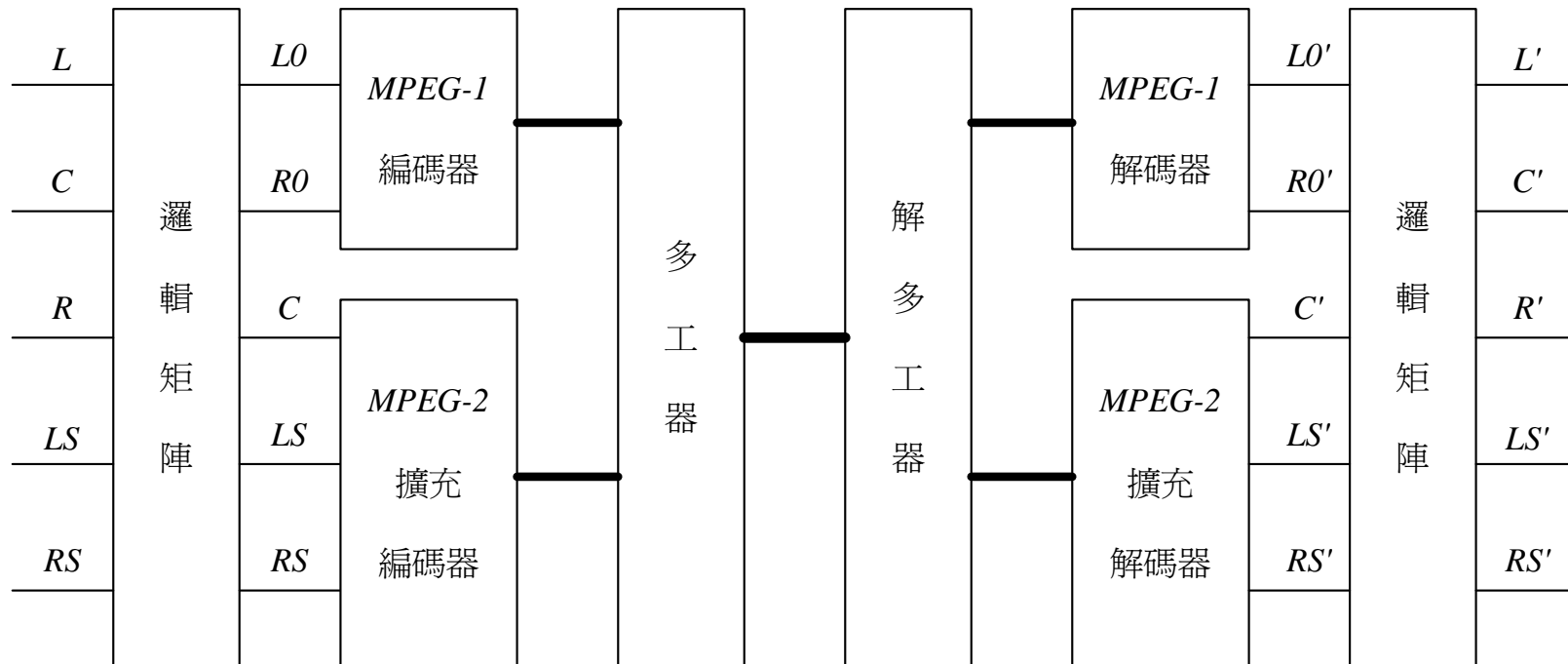
- Forward compatibility
 - A new decoder can decode an old bitstream
 - Usually simple to achieve
- Backward compatibility
 - An old decoder can decode a new bitstream, at least partially
 - Usually limits the coding efficiency

MPEG-2 backward compatible audio coding



$$\begin{cases} L0 = \alpha(L + \beta \cdot C + \delta \cdot LS) \\ R0 = \alpha(R + \beta \cdot C + \delta \cdot RS) \end{cases} \quad \alpha = \frac{1}{1+\sqrt{2}}; \beta = \delta = \frac{1}{\sqrt{2}} \text{ or } \alpha = 1; \beta = \delta = 0$$

Backward compatible audio coding (cont.)





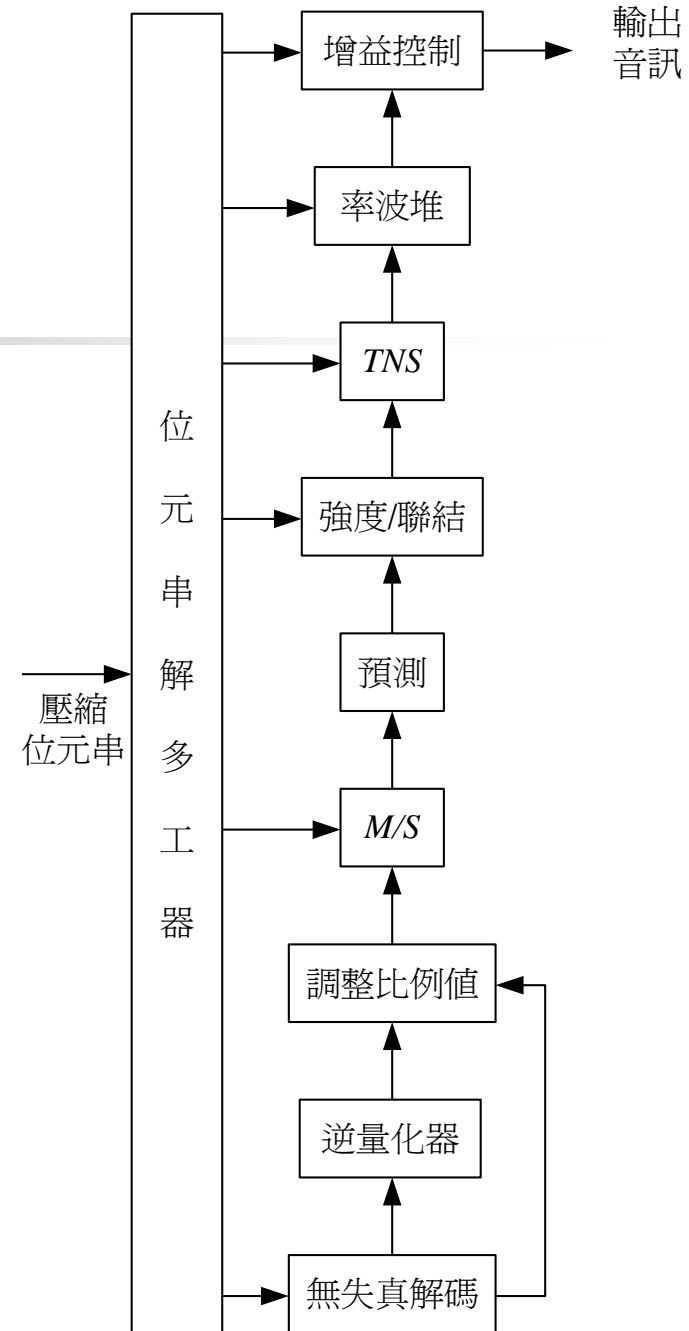
Non-backward compatible (NBC) audio coding

- MPEG-2 advanced audio coding (AAC)
 - ISO/IEC 13818-7 (April 1997)
 - 320~384 Kbps for 5 channels, 64 Kbps/channel
 - NBC at 320 Kbps as good as BC at 640 Kbps
 - Same framework (perceptually subband coding) as MPEG-1, with some enhancement

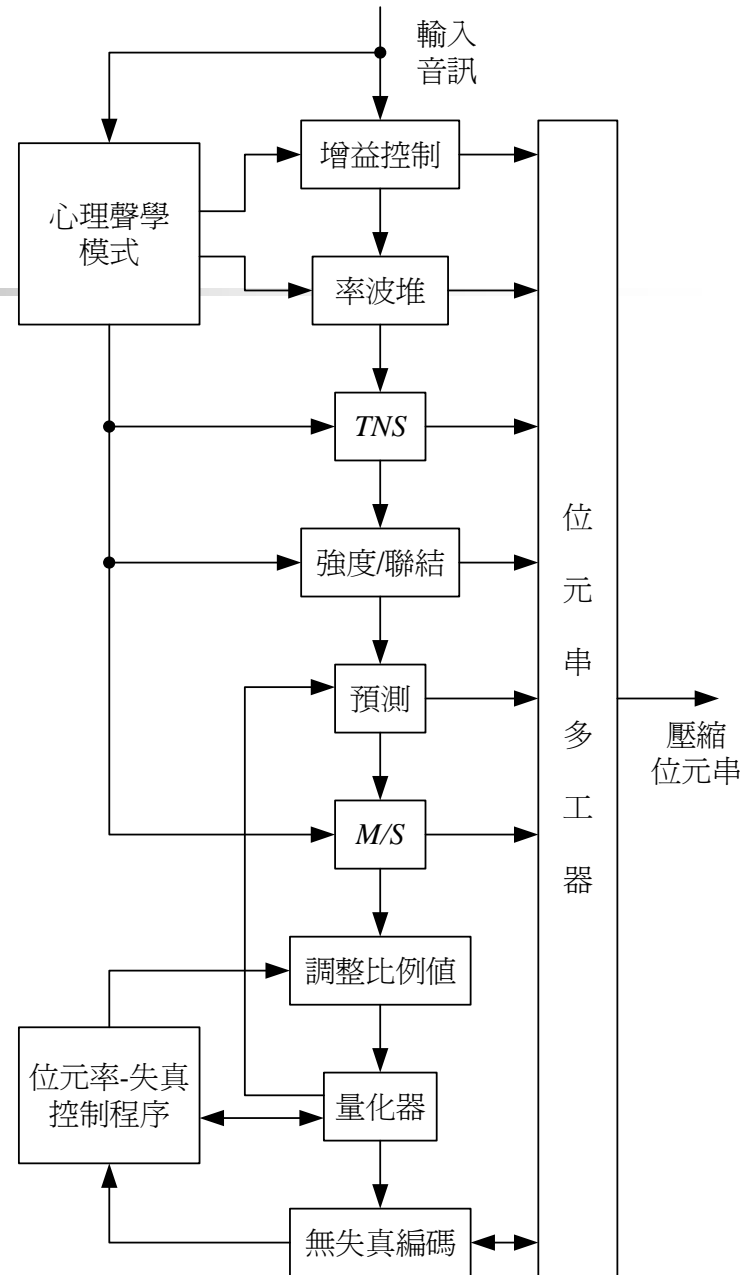
MPEG-2 AAC

■ Enhancements

- Preprocessing
- High resolution filterbanks (128~1024- points MDCT)
- Temporal noise shaping (TNS) : time-dependent quantization noise
- Coupling channel
 - Intensity multiple channel coding
- Backward adaptive prediction in subbands
- M/S stereo coding : $M=(R+L)/2$, $Side=(L-R)/2$
- Noiseless coding : Huffman coding

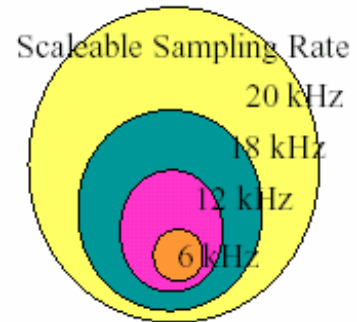
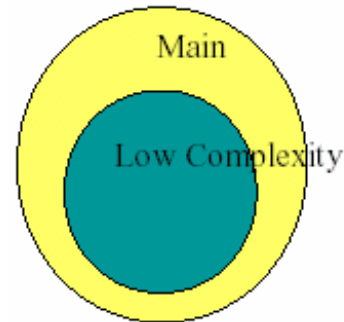


Encoder



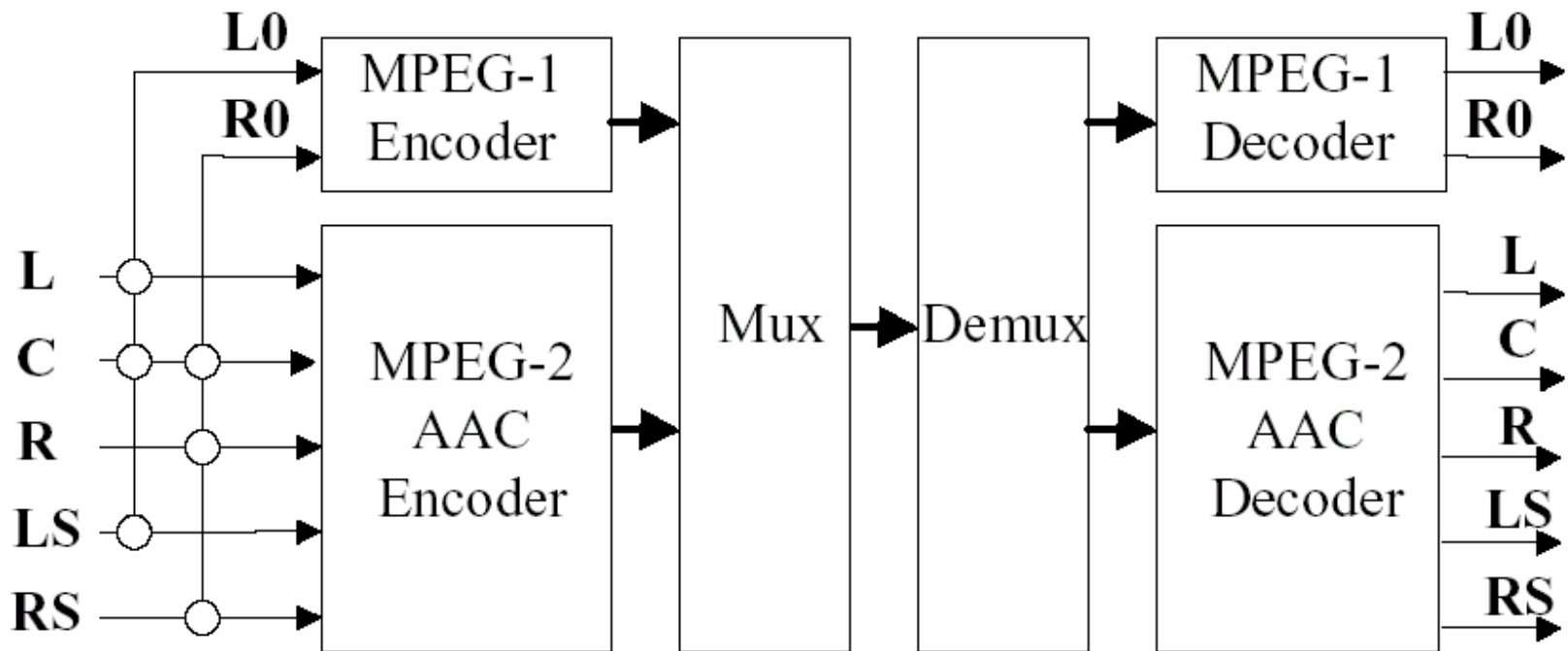
MPEG-2 AAC profiles

- Main profiles
 - Best quality, highest quality
 - 1024 (stationary) or
 - 128 MDCT (variant)
- Low complexity profile
 - No temporal noise shaping, no prediction
- Scalable sampling-rate profile
 - Scalable output sampling rates and complexity
 - Uses hybrid filterbanks (like MPEG-1 layer III)
 - No prediction, no coupling channel



Simcast

- To achieve backward compatibility at the cost of higher bitrate

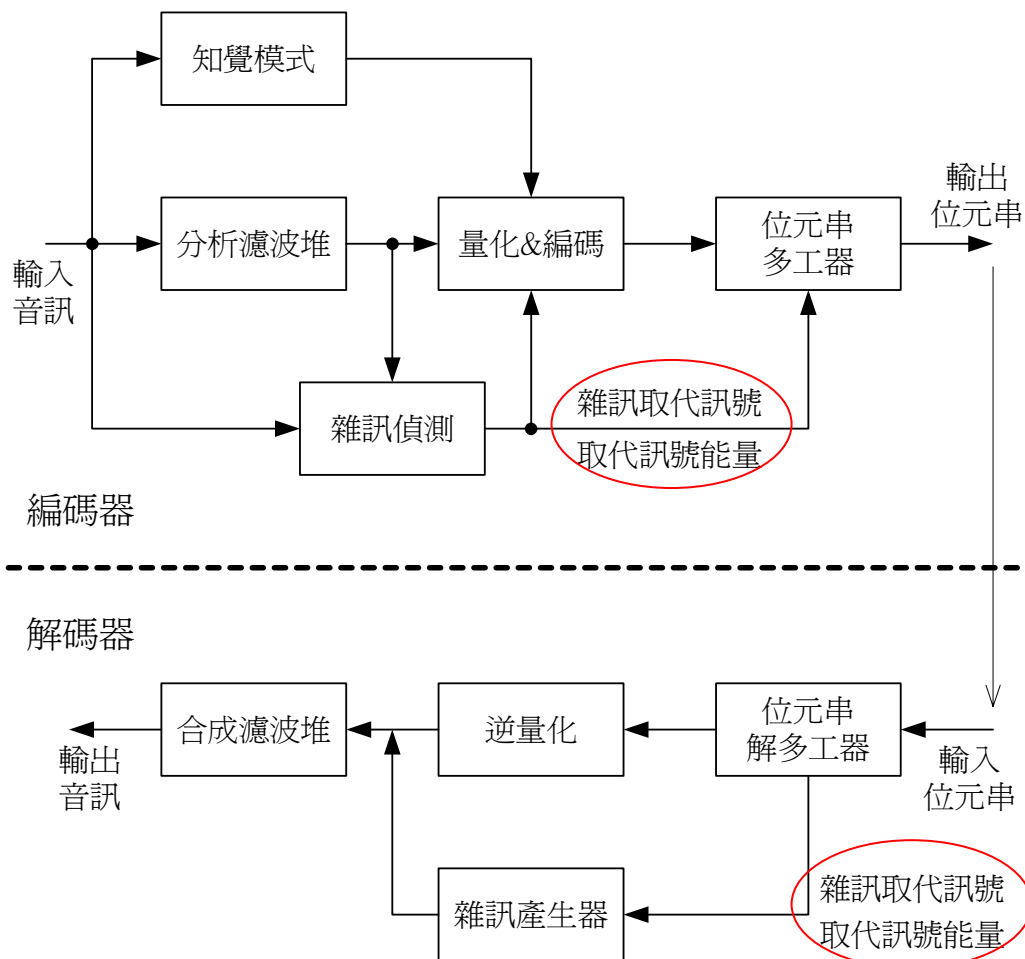




MPEG-4 GA audio

- ISO/IEC 14496-3
- GA : general audio
- MPEG-2 AAC + PNS, LTP, and TwinVQ tools
- PNS
 - Perceptual noise shaping
 - Used to improve the coding performance of noise-like signal

PNS

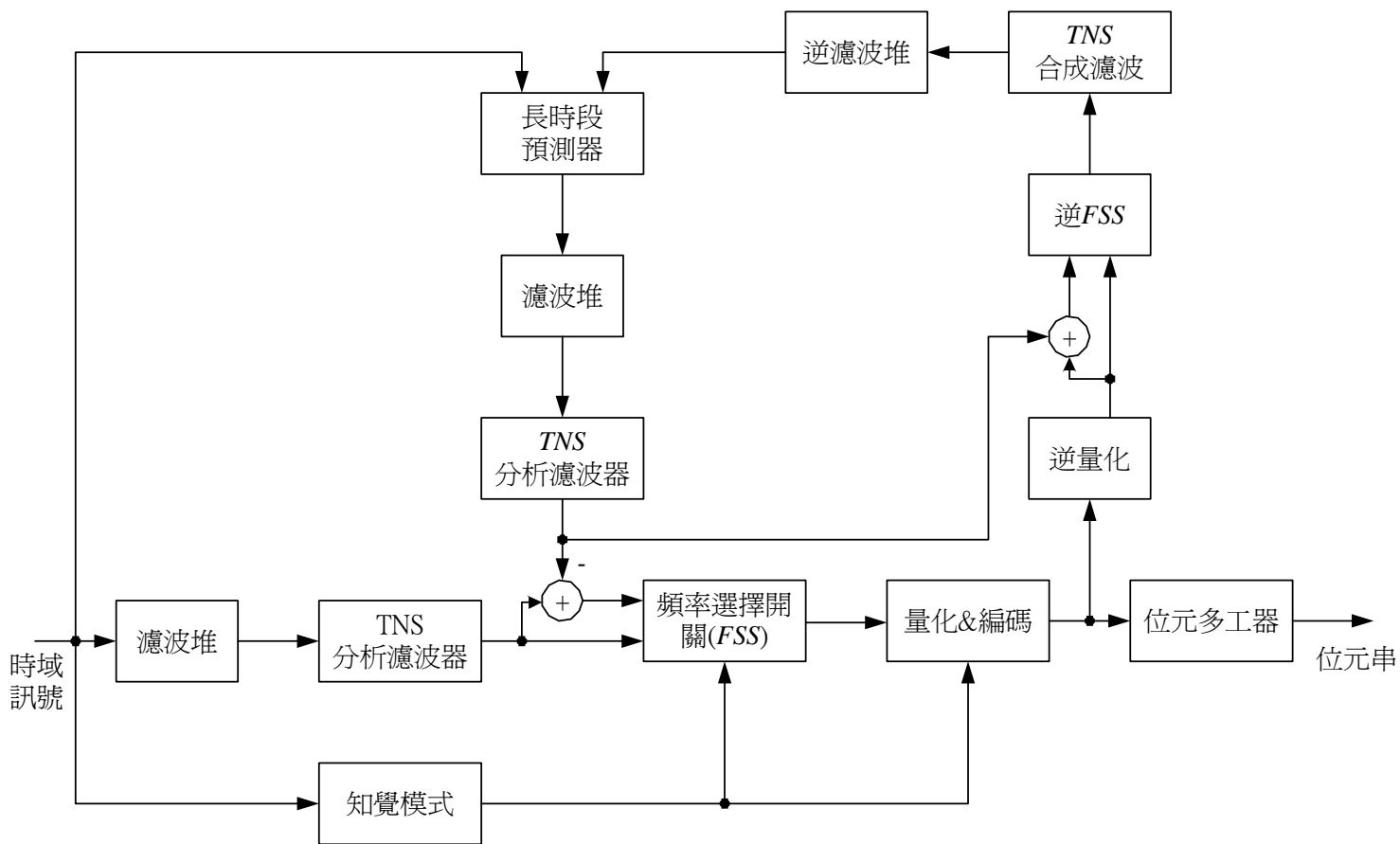




LTP

- Long term prediction
 - Used to estimate the redundancy due to signal periodicity
 - Integrated with perceptual model (LTP for speech signals is usually accomplished in time domain)
 - Pitch lag and gain
 - Significant performance gain in stationary, resonant segments

LTP





TwinVQ

- Transform-domain weighted interleave vector quantization
- **Low bit rate coding**
 - ≤ 16 Kb/s
 - ≥ 6 Kb/s
- Two steps :
 - Coefficients normalization
 - Parameters in normalization should be quantized and send out as side information
 - Quantization by weighted VQ
 - Interleave
 - VQ

Quantization of TwinVQ

■ NTT DoCoMo

