Multimedia Signals and Systems

MP3 - Mpeg 1,2 layer 1,2,3 Audio Encoding

Kunio Takaya

Electrical and Computer Engineering

University of Saskatchewan

March 26, 2008

"A review of algorithms for perceptual coding of digital audio signals"
Painter, T. Spanias, A.
Dept. of Electr. Eng., Arizona State Univ., Tempe, AZ;
http://ieeexplore.ieee.org/iel3/4961/13644/00628010.pdf?arnumber=628010

MP3' Tech - Encoding engines source codes:
http://www.mp3-tech.org/programmer/encoding.html

"ECE-700 Filterbank Notes.", Why Filterbanks? Sub-band Processing:
Phil Schniter, Ohio State Univ. March 10, 2008. 1
http://www.ece.osu.edu/˜ schniter/ee700/handouts/filterbanks.pdf

** Go to full-screen mode now by hitting CTRL-L

Contents

1 Stages and Layers

Two points need to be distinguished. Firstly, MPEG works in stages. These stages are normally denoted in Arabic figures (MPEG-1, MPEG-2, MPEG-4). Each stage has target encoding bit rates and application areas as shown in Table-**??**. In terms of audio coding, MPEG-1 and MPEG-2 have both a three layers structure. Each layer represents a family of coding algorithms. These layers are denoted in Roman figures (Layer I, Layer II, Layer III).

MPEG Stages

Stage	Coding bit rate	Applications	Completion
MPEG-1	1 Mbps	Video CD	93 Mar.
MPEG-2	$4 \sim 10$ Mbps (SDTV) approx. 50 Mbps (HDTV)	DVD, Cable TV	95 Mar.
MPEG-4	$\sim$ 382 Kbps (QCIF) 128 Kbps $\sim$ 2 Mbps (CIF) 15 Mbps (SDTV) 38.4 Mbps (HDTV)	Cellular Phone Internet broadcasing	99 May

Different layers have been defined and they all have their own advantages. Moreover, the complexity increases going from the Layer I to the Layer III.

MPEG-1 Audio is designed for monaural and stereophonic audio signals sampled at 32 KHz, 44.1 KHz, or 48 KHz.

1. Layer I possesses the lowest complexity and is specifically targeted to applications where the complexity of the encoder plays an important role.

2. Layer II requires a more complex encoder as well as a slightly more complex decoder. Compared to Layer I, Layer II is able to suppress more redundacy in the signal and applies the psychoacoustic model in more efficient way. Video CD, 128 Kbps

3. Layer III is once again of an increased complexity and is targetted to applications needing the lowest data rates, by its suppression of the redundant signal and its improved extraction of feebly audible frequencys using its filter. MP3, 64 Kbps

2 How does MPEG audio work?

- MPEG audio compressor are based on a perceptual coding scheme. During a perceptual coding, the codec does not try absolutely to maintain a signal identical to original after encoding and decoding phases, but its goal is rather to insure that the output signal seems identical for a human ear.
- The first psychoacoustic effect that the perceptual coding uses is the masking effect. Some parts of the signal. due to the functioning of the human auditory system, are not audible. In the presence of strong sounds, you do not hear weak sounds. To suppress weak sounds, the encoder integrates a psychoacoustic model trying to mimic the human ear's behaviour. This psychoacoustic model analyzes the input signal to determine the masking level that affects the human auditory system, then estimates the minimal audible level.

- During its quantification and coding phase, the encoder tries to allocate the number of bits based on the masking properties as well as the size of the authorized data rate by adjusting the quantization levels.

3 MPEG-1 Multiplexing and Synchronization

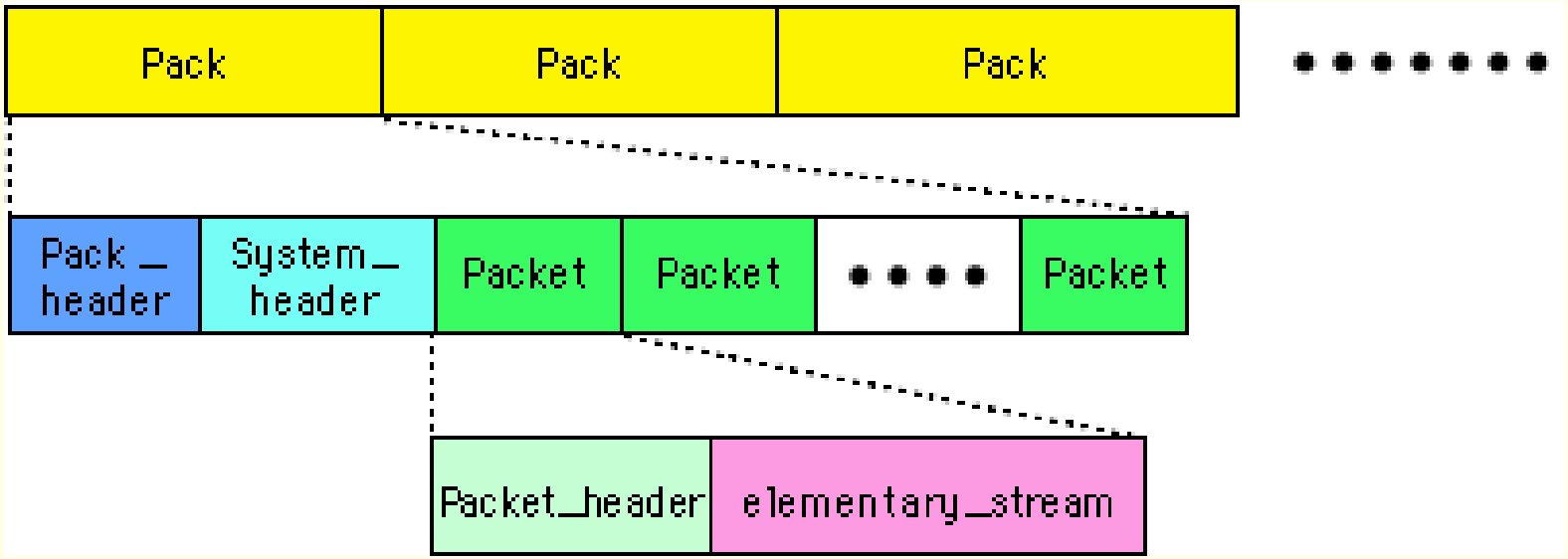


Figure 1: MPEG-1 Packet Structure

Pack a pack header [pack starting code, SCR (System Clock Reference), multiplexing bit rate (mux rate), a system header [bit rate of individual streams, and buffer size...]]

Packet a packet header [stream identifier, packet size, PTS (Presentation Time Stamp), DTS (Decoding Time Stamp) and other control information for synchronized play back]

Multiplexing A video packet and an audio packet are alternated to realize the multiplexing between video and audio.

Synchronization MPEG-1 system decoder has STC running at 90 KHz. Synchronized when

Pack: STC = SCR

Packet: STC = PTS or STC = DTS (if decoded time is different from presentation time)

4 Overview of the MP3 Techniques

To achieve a great deal of compression on audio data, the MP3 uses a few techniques and tricks. The MP3's principle consists of several conceptual components. Perceptual coding is one of the unique feature of the MP3. Such key components are briefly introduced to provide a common background.

MP3' Tech web site: Overrview and Source Codes

4.1 The minimal audition threshold

The minimal audition threshold of the ear is not linear. It is represented, according to the law of Fletcher and Munson, by a curve dug between 2Khz and 5Khz. It is not therefore necessary to code sounds situated under this threshold, because they will not be perceived.

4.2 The masking effect

The MP3 is based on masking properties of the human ear. When you look at the sun and if a bird passes ahead, you do not see it because of the too predominant light of the sun. In audio, it is similar. During strong sounds, you do not hear the weakest sounds. Take as an example a piece of organ: when the organist does not play, you hear the breath in the piping, and when he plays, you no longer hear it because it is masked. Masking effect is frequency dependent, and masks perception before and after the masking sound.

It is therefore not necessary to code all the sounds. This is the first property used by the MP3 format to earn some space. For this the MP3 encoder uses a psychoacoustic model modeling the behavior of the human ear.

4.3 The bytes reservoir

Often, some passages of a musical piece can not be coded to a given rate without altering the musical quality. The MP3 then uses then a short reservoir of bytes that acts as a buffer by using capacity from passages that can be coded to an inferior rate in the given flow.

4.4 The Joint Stereo coding

In the case of a stereophonic signal, the MP3 format can then use a few more tools, reffered as Joint Stereo (JS) coding, to further shrink the compressed file size.

In many mid-range Hi-fi sets , there is a unique subwoofer.
However you usually do not have the feeling that the sound comes

from this boomer, but rather from satellite speakers. Indeed for very low and very high frequencies, the human ear is no longer able to locate the spacial origin of sounds with full accuracy. The mp3 format can therefore (optionally) revert to such a trick by using what is called Intensity Stereo (IS). Some frequencies are then recorded as a monophonic signal followed by a few additional information in order to restore a minimum of spatialisation.

The second joint stereo tool is called Mid/Side (M/S) stereo. When the left and the right channels are quite similar, then a middle (L+R) and a side (L-R) channels are encoded instead of left and right. This allows to reduce the final file size by using less bits for the side channel. During playback, the MP3 decoder will reconstruct the left and right channels.

4.5 The Huffman coding

The MP3 also uses the classic technique of the Huffman algorithm. It acts at the end of the compression to code information, and this is not therefore itself a compression algorithm but rather a coding method.

This coding creates variable length codes on a whole number of bits. Higher probability symbols have shorter codes. Huffman codes have the property to have a unique prefix, they can therefore be decoded correctly in spite of their variable length. The decoding step is very fast (via a correspondence table). This kind of coding allows to save on the average a bit less than 20% of space.

It is an ideal complement of the perceptual coding: During big polyphonies, the perceptual coding is very efficient because many sounds are masked or lessened, but little information is identical,

so the Huffmann algorithm is very seldom efficient. During "pure" sounds there are few masking effects, but Huffman is then very efficient because digitalized sound contains many repetitive bytes, that will then be replaced by shorter codes.

5 MPEG-1 Audio Layer-I and Layer-II

MPEG-1 Audio Layer-I and Layer-II are basically the same encoder. Layer-II is more efficient due to some improvements made in scale factor encoding, bit assignment algorithm, and quantization. The block diagram of Layer-II audio encoder is shown in Fig. 2.

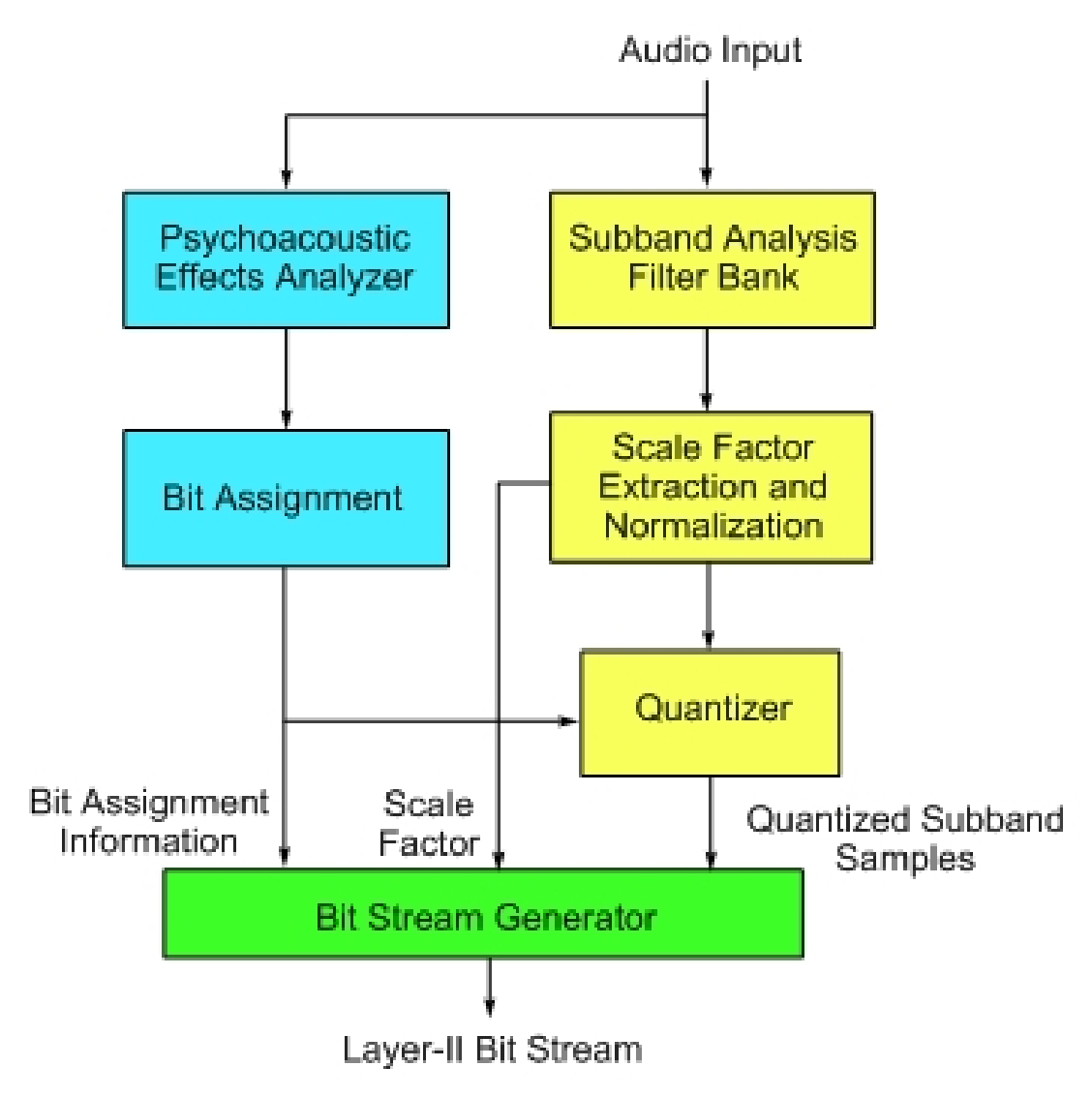


Figure 2: MPEG-1 Layer-II Block Diagram

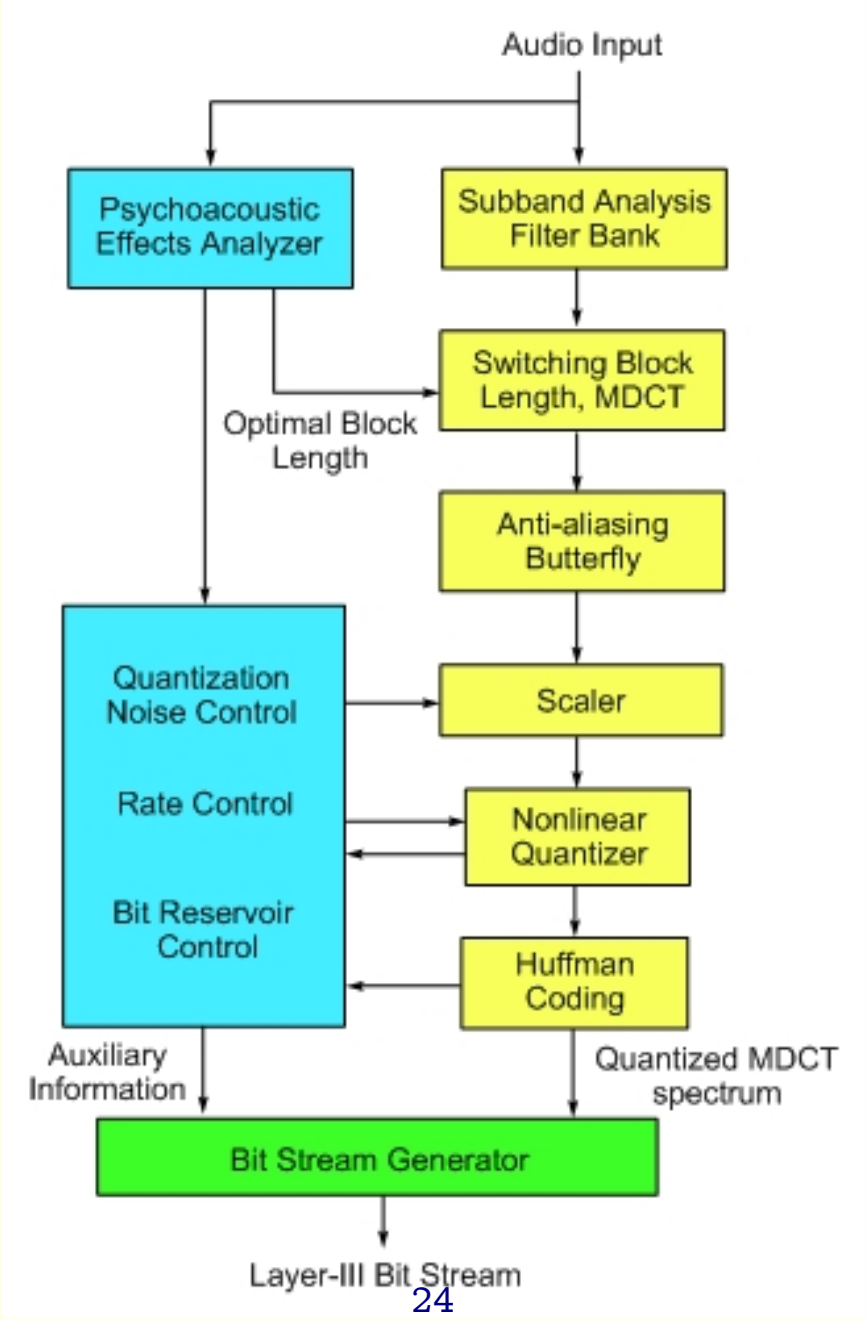
Audio input signal is sampled at one of the sampling rate, 32 KHz, 44.1 KHz, or 48 KHz, then put in a block (frame) of a certain number of samples. The block of data is fed into two major paths of processing. One path (yellow) is for subband analysis using a subband filter bank. The subband filter bank subdivide the input signals into 32 equally divided frequency bands, producing 32 subband signals down sampled at 1/32 of the sampling rate. Then, the maximum absolute amplitude is searched for each subband block. The subband signal is divided by the maximum value to normalize the data to fit in the scale of [-1, 1]. The logarithmic value of the maximum absolute amplitude is recorded as the scale factor of a suband.

In another path (blue), psychoacoustic effects are analyzed. Masking threshold of the psychoacoustic masking effect is calculated first. Based on the masking level, a number of bits to be

used to qauntize a subband signal is assigned to each subband. FFT is first applied to a frame of audio input. From the calculated spectrum, masking thresholds (permissible quantization noise power for each subband) are calculated. The number of quantization bits for each subband is iteratively determined in a loop by considering the masking thresholds and the number of total bits allowed for a frame under the constraint of bit rate.

The quantizer linearly quantizes each and every subband with the assigned number of quantization bits. The bit stream generator produces the output bit stream by multiplexing quantized sbuband signals, assigned number of quantization bits per subband, scale factors, and adding the header.

6 MPEG-1 Audio Layer-III

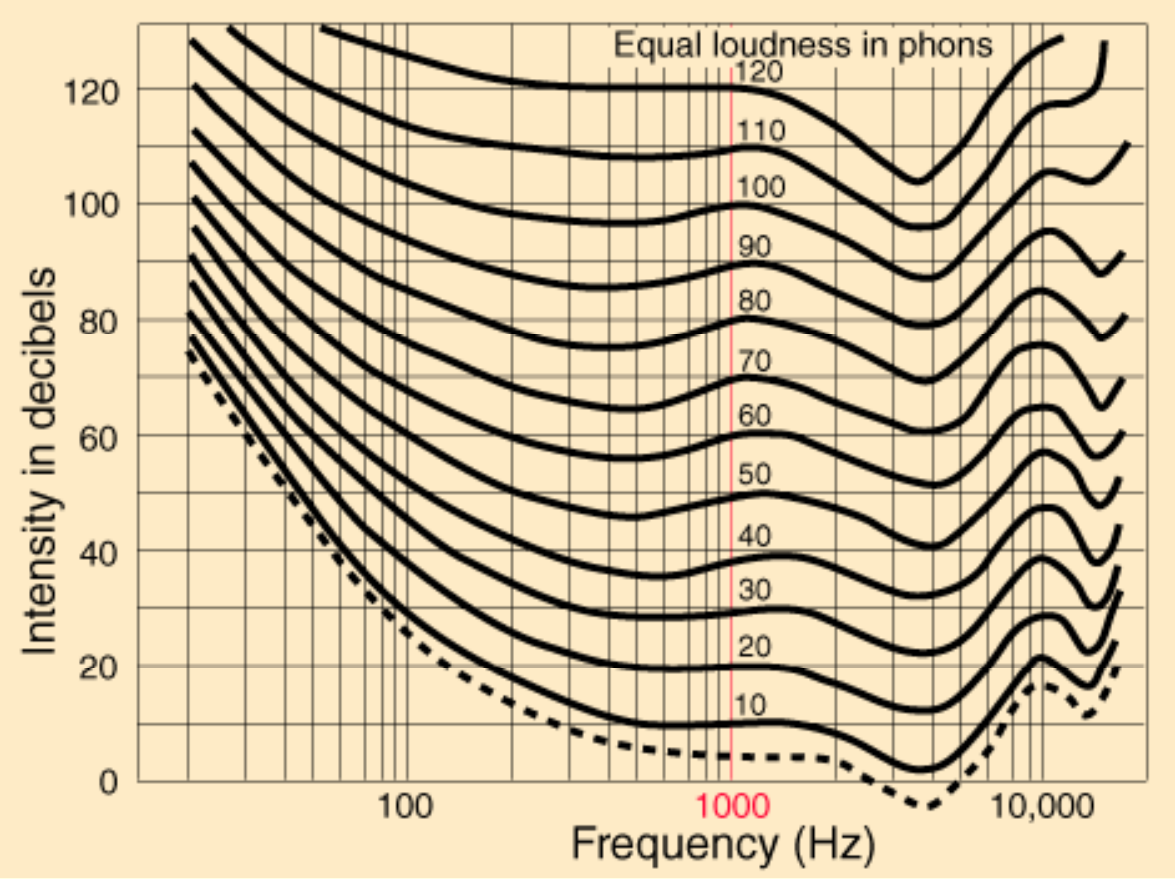

Audio Input
Psychoacoustic Effects Analyzer
Subband Analysis Filter Bank
Switching Block Length, MDCT
Optimal Block Length
Anti-aliasing Butterfly
Quantization Noise Control
Rate Control
Bit Reservoir Control
Scaler
Nonlinear Quantizer
Huffman Coding
Auxiliary Information
Quantized MDCT spectrum
Bit Stream Generator
Layer-III Bit Stream

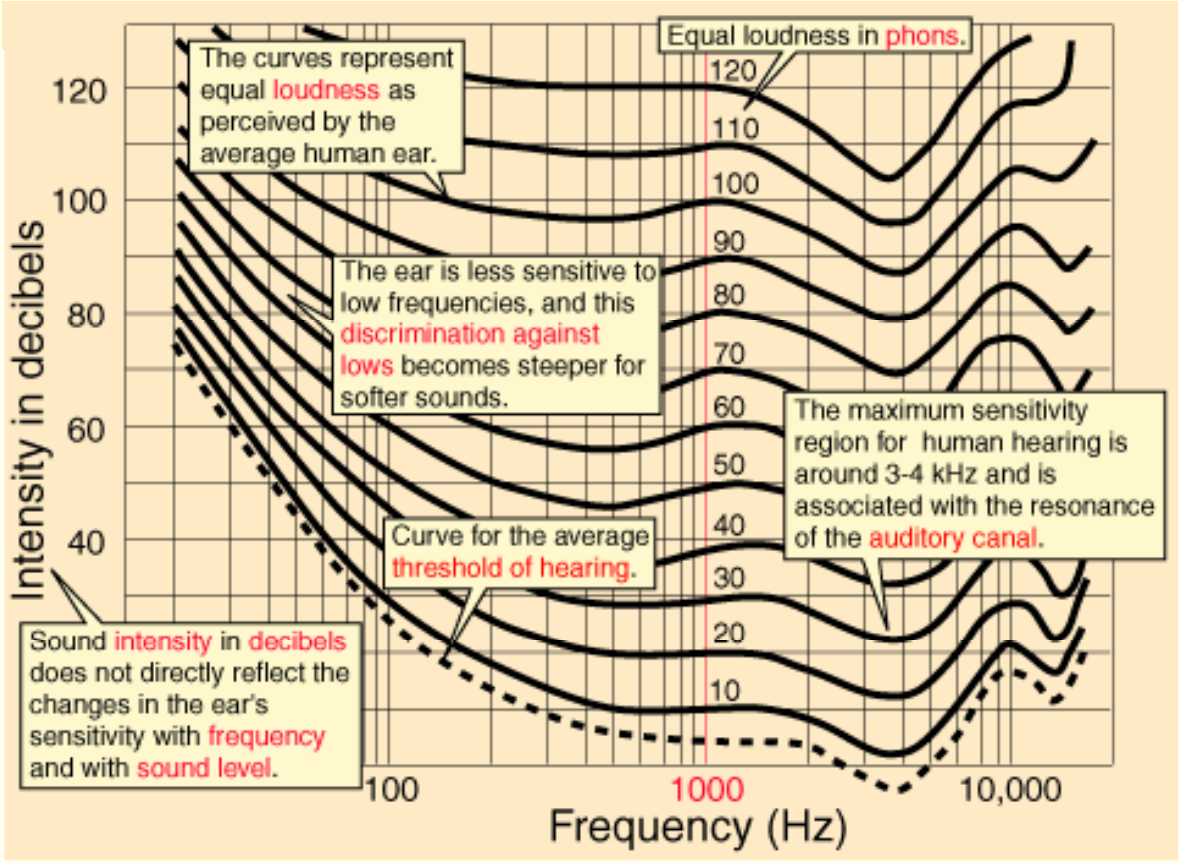
The MPEG-1 Audio Layer-III algorithm shown in Fig. 3 is substantially different from Layer-I and Layer-II. The 32 subband filter bank remains the same as the Layer-II. The subband signals output from the filter band are then transformed into the frequency domain by MDCT (Modified Discrete Cosine Transform). MDCT specrum is processed by the anti-aliasing butterfly to remove the aliased spectral components, caused by the down sampling involved in the filter bank. The MDCT spectrum is processed by the scaler, then quantized and finally encoded by the Huffman coding. The stage of scaler and non-linear quantizer interacts with the spychoacoustic analyzer. The permissible quantization noise power, bit rate, and the size of buffered data in the bit reservoir are considered to determine the quantization step size and the scale factors for individual subband. These are determined by an iterative loop operation. Two block sizes, short and long, are switched depending on the characteristics of input

signal, when MDCT is applied in order to suppress the pre-echo noise bothersome for hearing.

The major difference between the Layer-II and Layer-III is that the output from the subband filter bank in the time domain is endcoded in the Layer-II, while the Layer-III encodes the MDCT spectrum in the frequency domain. The bit stream generator multiplexes the quantized MDCT spectra, and add auxiliary information such as the MDCT block size, quantization step size, scale factors, information regarding the data and table area for the Huffman coding.

7 Threshold of Hearing





Equal Loudness Curves and Annotation

Sound level measurements in decibels are generally referenced to a standard threshold of hearing at 1000 Hz for the human ear which

can be stated in terms of sound intensity:

$$I_0 = 10^{-12} watts/m^2 = 10^{-16} watts/cm^2$$

or in terms of sound pressure:

$$P_0 = 2 \times 10^{-5} Newtons/m^2 = 2 \times 10^{-4} dyne/cm^2$$

This value has wide acceptance as a nominal standard threshold and corresponds to 0 decibels. It represents a pressure change of less than one billionth of standard atmospheric pressure. This is indicative of the incredible sensitivity of human hearing. The actual average threshold of hearing at 1000 Hz is more like $2.5x10^{-12} watts/m^2$ or about 4 decibels, but zero decibels is a convenient reference. The threshold of hearing varies with frequency, as illustrated by the measured hearing curves. The threshold of pain is:

$$10^{13} I_0 \text{ or } 120 dB$$

Sound Intensity:

Sound intensity is defined as the sound power per unit area. The usual context is the measurement of sound intensity in the air at a listener's location. The basic units are watts/m2 or watts/cm2. Many sound intensity measurements are made relative to a standard threshold of hearing intensity I0: The most common approach to sound intensity measurement is to use the decibel scale:

$$I \quad (dB) = 10 \log_{10} \frac{I}{I_0}$$

Sound Pressure:

Since audible sound consists of pressure waves, one of the ways to quantify the sound is to state the amount of pressure variation relative to atmospheric pressure caused by the sound. Because of the great sensitivity of human hearing, the threshold of hearing corresponds to a pressure variation less than a billionth of atmospheric pressure.

The standard threshold of hearing can be stated in terms of pressure and the sound intensity in decibels can be expressed in terms of the sound pressure:

$$I \ \ (dB) = 10 \log_{10} \frac{I}{I_0} = 20 \log_{10} \frac{P}{P_0}$$

Phons:

Two different 60 decibel sounds will not in general have the same loudness. Saying that two sounds have equal intensity is not the same thing as saying that they have equal loudness. Since the human hearing sensitivity varies with frequency, it is useful to plot equal loudness curves which show that variation for the average human ear. If 1000 Hz is chosen as a standard frequency, then each equal loudness curve can be referenced to the decibel level at 1000 Hz. This is the basis for the measurement of loudness in phons. If a given sound is perceived to be as loud as a 60 dB sound at 1000

Hz, then it is said to have a loudness of 60 phons.
60 phons means "as loud as a 60 dB, 1000 Hz tone"
The loudness of complex sounds can be measured by comparison to 1000Hz test tones, and this type of measurement is useful for research, but for practical sound level measurement, the use of filter contours has been commonly adopted to approximate the variations of the human ear.
Reference:
HyperPhysics Georgia State University

8 Psychoacoustic Principles

Most current audio coders achieve compression by exploiting the fact that irrelevant signal information is not detectable by even a well trained or sensitive listener. Irrelevant information is identified during signal analysis by incorporating into the coder several psychoacoustic principles:

1. absolute hearing thresholds
2. critical band frequency analysis
3. simultaneous masking
4. the spread of masking along the basilar membrane
5. temporal masking

Combining these psychoacoustic notions with basic properties of signal quantization has led to the development of perceptual entropy, a quantitative estimate of the fundamental limit of audio signal compression.

Reference:

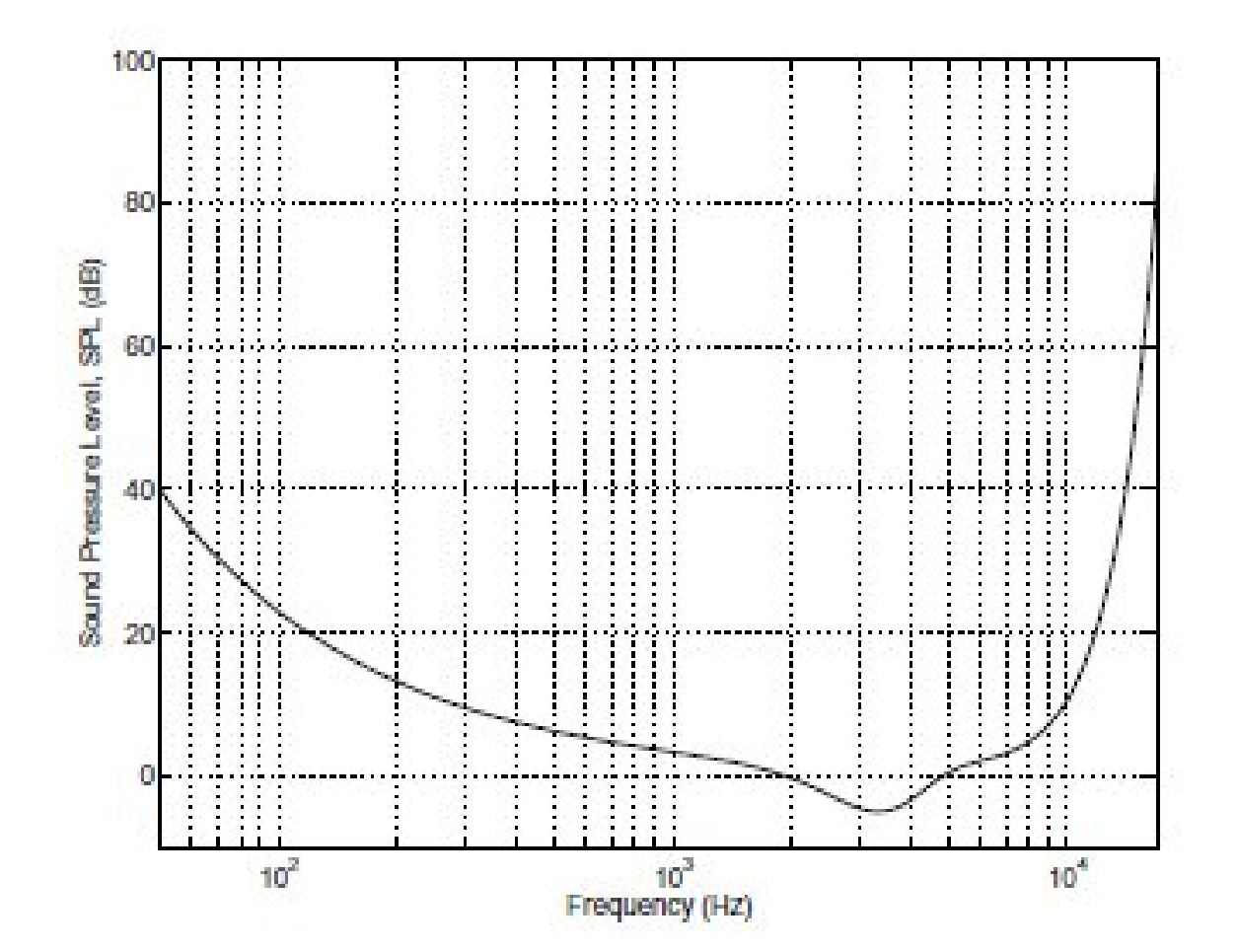
1. Ted Painter, Perceptual Coding of Digital Audio

2. Ted Painter, A Review of Algorithms for Perceptual Coding of Digital AudioSignals

8.1 Absolute Threshold of Hearing

The absolute threshold of hearing is characterized by the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. (Fletcher 1940) The quiet threshold is well approximated by the nonlinear function,

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4$$



The Absolute Threshold of Hearing

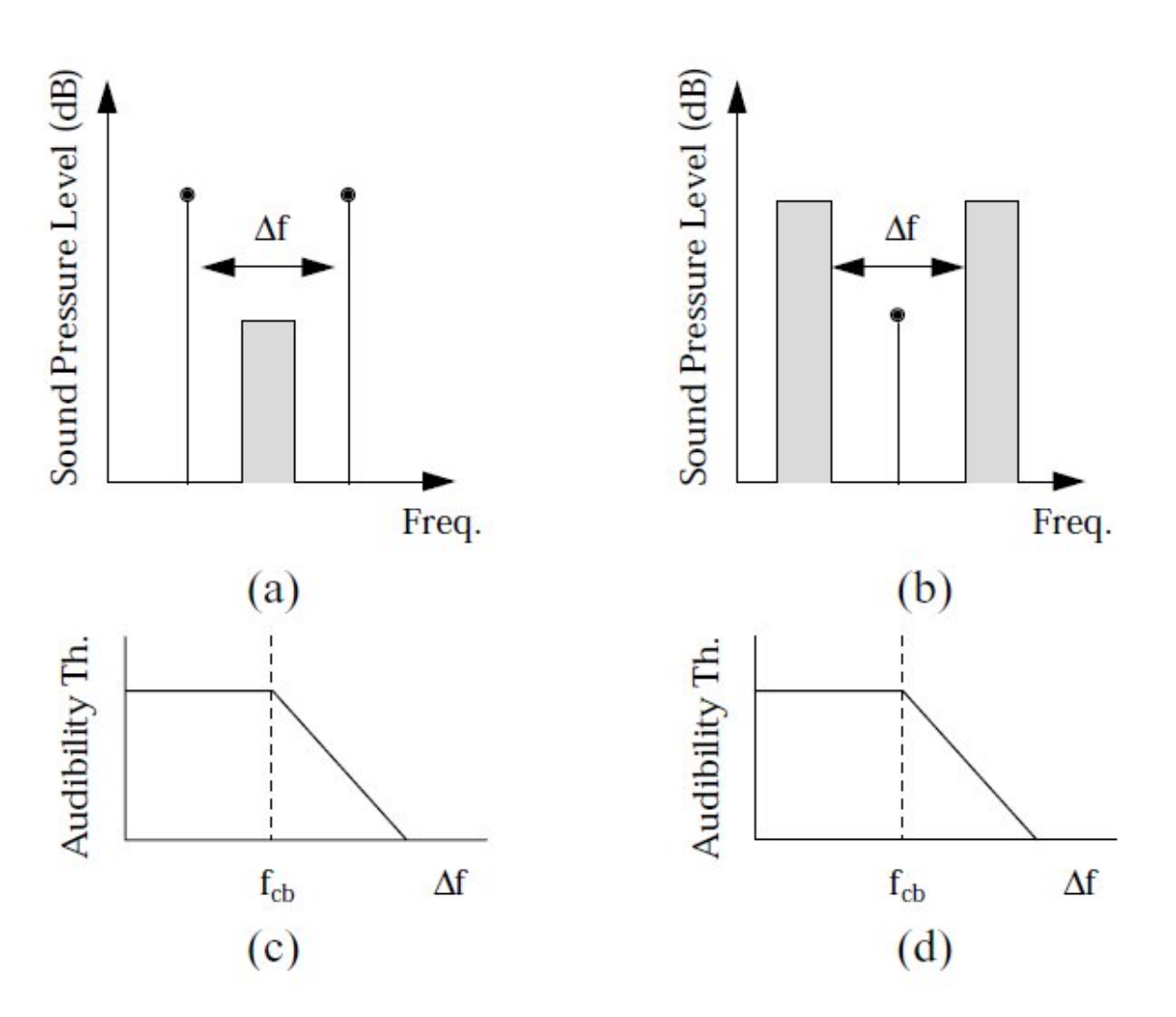
When applied to signal compression, $T_q(f)$ can be interpreted as a maximum allowable energy level for coding distortions in the frequency domain. Algorithm designers have no a priori knowledge regarding actual playback levels, therefore the sound pressure level (SPL) curve is often referenced to the coding system by equating the lowest point on the curve (i.e., 4 kHz) to the energy in +/- 1 bit of signal amplitude.

MATLAB code

```
ff=[];
for pw=1:4
    for v=1:0.5:9.5
        ff=[ff, v*10^pw];
    end
end
f=ff(9:56); n=size(f,2);
TqValue=[];
for i=1:n
    fv=f(i);
    TqValue=[TqValue,Tq(fv)];
end
semilogx(f,TqValue); axis tight;

function valTq = Tq(f)
valTq=3.64*(f/1000)^(-0.8)-6.5*exp(-0.6*(f/1000-3.3)^2)+10^(-3)*(f/1000)^4;
```

8.2 Critical Bands - critical bandwidth



- A frequency-to-place transformation takes place in the inner ear, along the basilar membrane. Distinct regions in the cochlea, each with a set of neural receptors, are tuned to different frequency bands.
- In the experimental sense, critical bandwidth can be loosely defined as the bandwidth at which subjective responses change abruptly. For example, the perceived loudness of a narrowband noise source at constant sound pressure level remains constant even as the bandwidth is increased up to the critical bandwidth. The loudness then begins to increase.

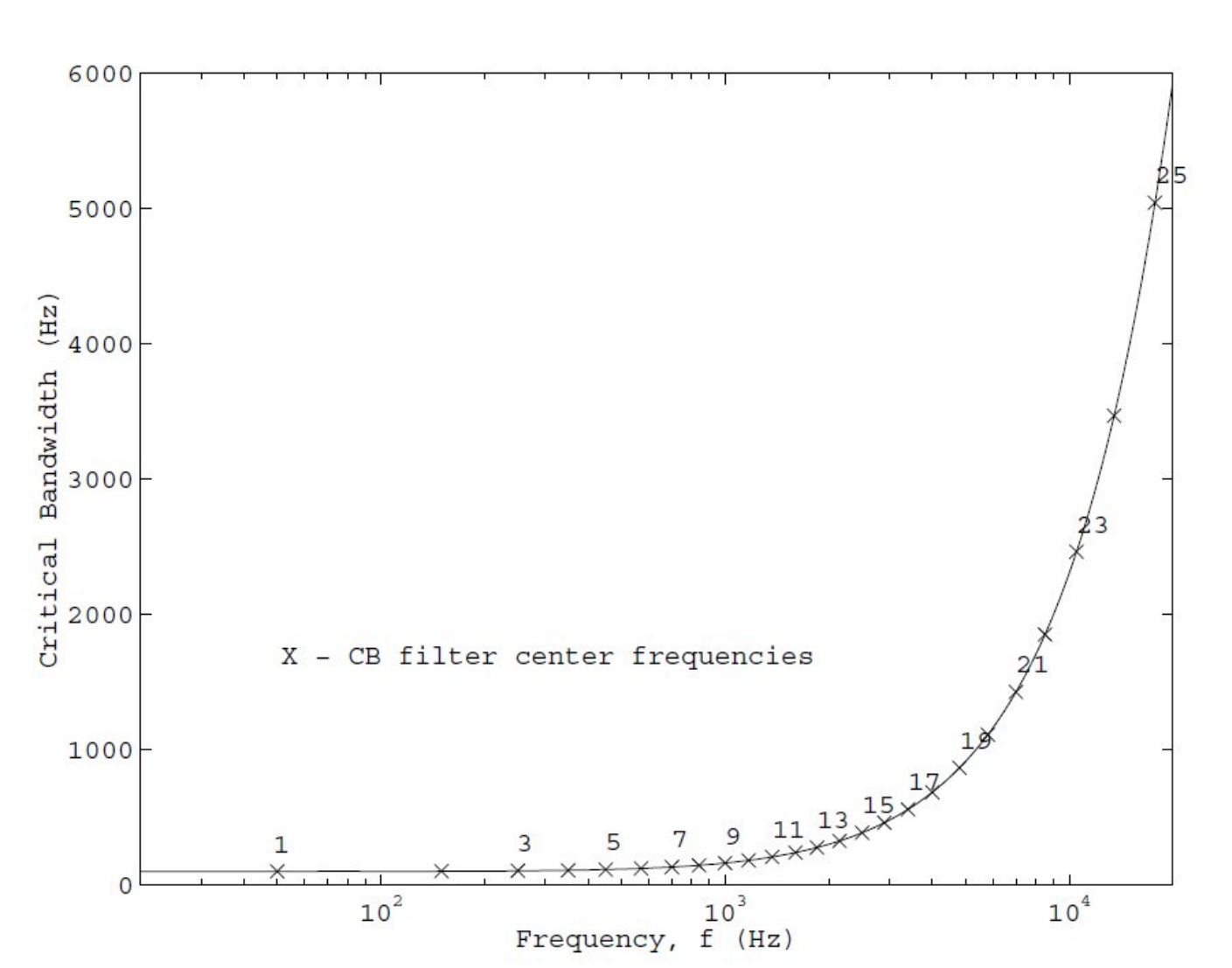
- In a different experiment Fig. (a,c), the detection threshold for a narrowband noise source between two masking tones remains constant as long as the frequency separation between the tones remains within a critical bandwidth. Beyond this bandwidth, the threshold rapidly decreases.
- A similar notched-noise experiment can be constructed with masker and maskee roles reversed Fig. (b,d). Critical bandwidth tends to remain constant (about 100 Hz) up to 500 Hz, and increases to approximately 20% of the center frequency above 500 Hz.

Experiment

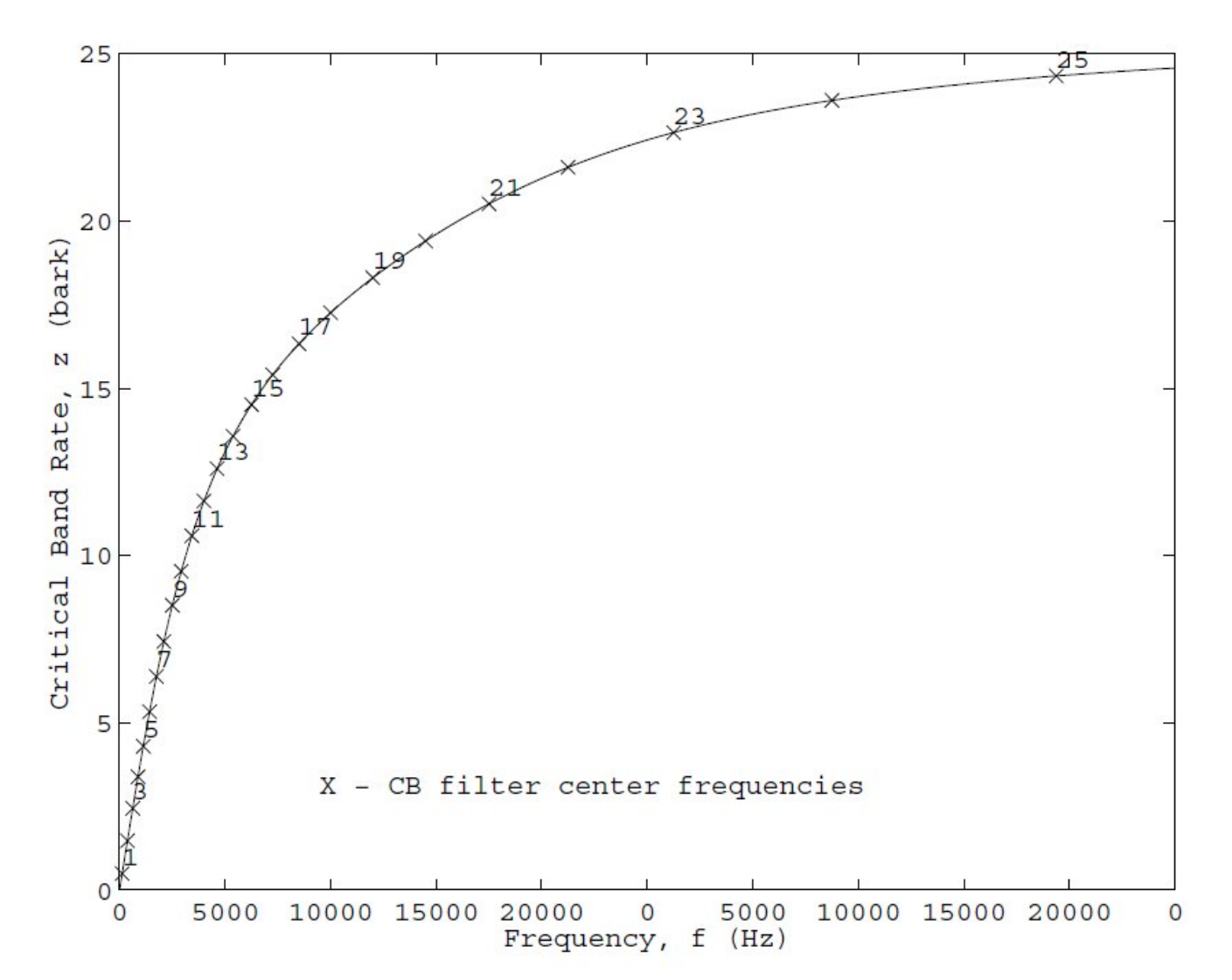
The following MATLAB script runs a simple experiment to test masking effect in a critical band centered at 4000 Hz. Find your threshold of hearing by varying the maskee noise level relative to the masker's (2 sine waves) power at this frequency.

```
% Psycoacoustic Masking Experiment at Most Sensitive frequency f=4000Hz
% Two sinusoidal Maskers at fc1 and fc2
% Noise signal (maskee) to be masked is BP signal between fn1 and fn2
% K. Takaya Feb. 5, 2007
amp = input('Enter maskee noise level (masker f1,f2 level is 0.25) =');
fs = 44100;  fsh = 22050; % CD sampling frequency
fc=4000;
fc1=3700;  fc2=4400;  % Critical Bandwidth of fc=4000
fn1=3850;  fn2=4200;  % Noise frequency range
w1=fn1/fsh;
w2=fn2/fsh;
[B,A]=butter(5,[w1,w2]);
t = 0:1/fs:1;
x = randn(size(t));
y = filter(B, A, x);
```

```
y = y / max(abs(y));
wavplay(amp*y, fs);
disp('Noise Done');
pause
v1 = 0.25 * sin(2 * pi * fc1 * t);
v2 = 0.25 * sin(2 * pi * fc2 * t);
v= v1+v2;
wavplay(v, fs);
disp('Sine Done');
pause
w = v + amp*y;
wavplay(w, fs);
disp('Combo Done');
pause
yf=abs(fft(w));
figure(1); plot(yf); axis tight;
```



Critical Bandwidth BW_c



Critical Bandwidth $z(f)$ in Bark

For an average listener, critical bandwidth is conveniently approximated by

$$BW_c(f) = 25 + 75[1 + 1.4(f/1000)^2]^{0.69}(\text{Hz}) \tag{1}$$

Although the function BW_c is continuous, it is useful when building practical systems to treat the ear as a discrete set of bandpass filters which obeys Eq. 1. A distance of 1 critical band is commonly referred to as one bark in the literature.

$$z(f) = 10\arctan(0.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right](\text{Bark}) \tag{2}$$

This function Eq. 2 is often used to convert from frequency in Hertz to the bark scale. Thus, one critical bandwidth comprises one bark.

Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100
2	150	100-200
3	250	200-300
4	350	300-400
5	450	400-510
6	570	510-630
7	700	630-770
8	840	770-920
9	1000	920-1080
11	1370	1270-1480
12	1600	1480-1720
13	1850	1720-2000

Band No.	Center Freq. (Hz)	Bandwidth (Hz)
14	2150	2000-2320
15	2500	2320-2700
16	2900	2700-3150
17	3400	3150-3700
18	4000	3700-4400
19	4800	4400-5300
20	5800	5300-6400
21	7000	6400-7700
22	8500	7700-9500
23	10,500	9500-12000
24	13,500	12000-15500
25	19,500	15500-

problem

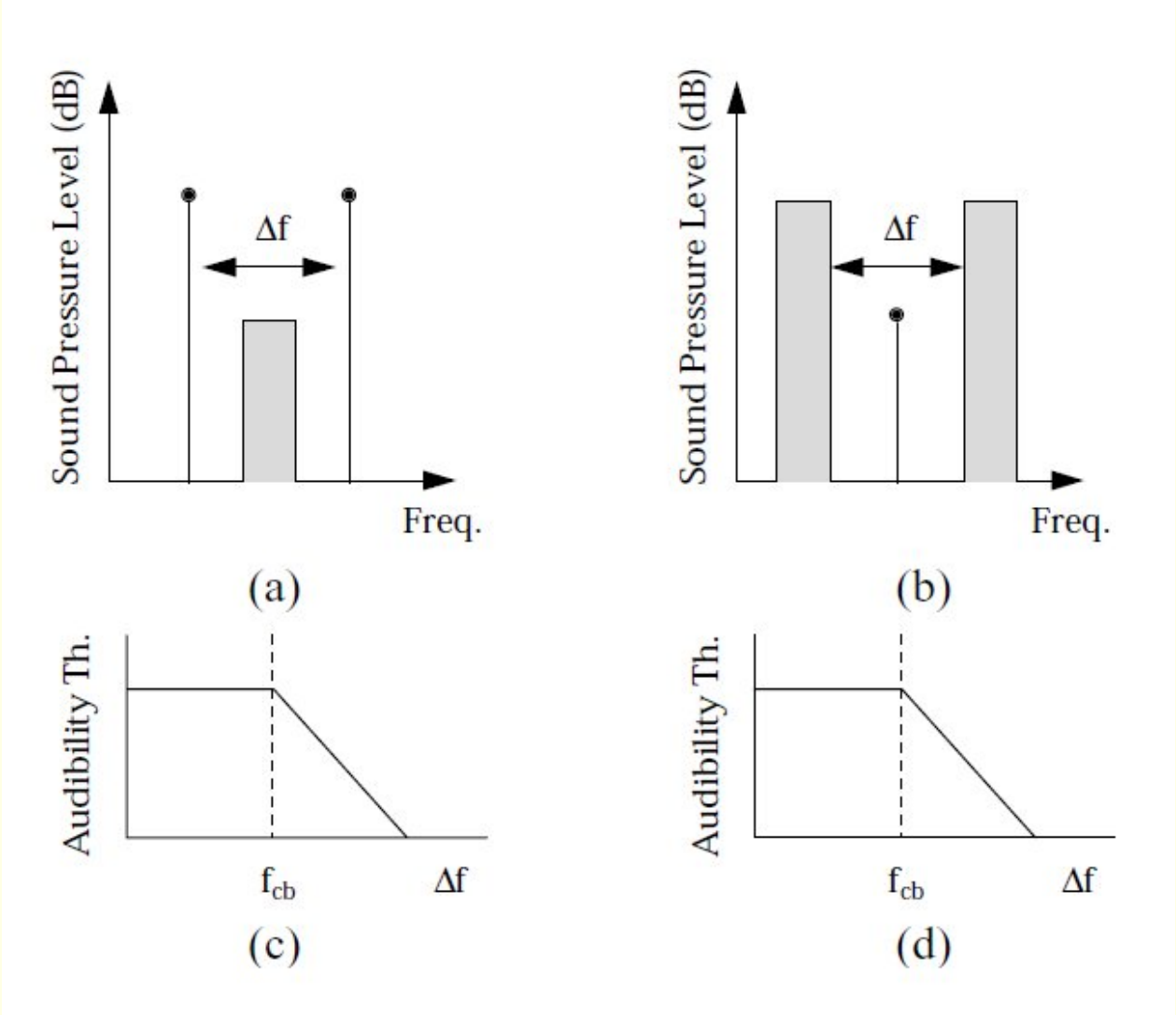
Plot the curves of BW_c and $z(f)$ by using the formula of BW_c given by Eq. 1 and $z(f)$ given by Eq. 2. Then, confirm that the critical bands are equally spaced when measured in terms of Bark. List the values of bandwidth in Bark for all of the critical bands from 1 to 25.

8.3 Simultaneous Masking and the Spread of Masking

Simultaneous (In-band) Masking

- Masking refers to a process where one sound is rendered inaudible because of the presence of another sound.
- Simultaneous masking refers to a frequency domain phenomenon which has been observed within critical bands (in-band). For the purposes of shaping coding distortions it is convenient to distinguish between two types of simultaneous masking, namely tone-masking-noise, and noise-masking-tone.
- In the first case, a tone occurring at the center of a critical band masks noise of any subcritical bandwidth or shape, provided the noise spectrum is below a predictable threshold directly related to the strength of the masking tone.

- The second masking type follows the same pattern with the roles of masker and maskee reversed. A simplified explanation of the mechanism underlying both masking phenomena is as follows.
- The presence of a strong noise or tone masker creates an excitation of sufficient strength on the basilar membrane at the critical band location to effectively block transmission of a weaker signal.



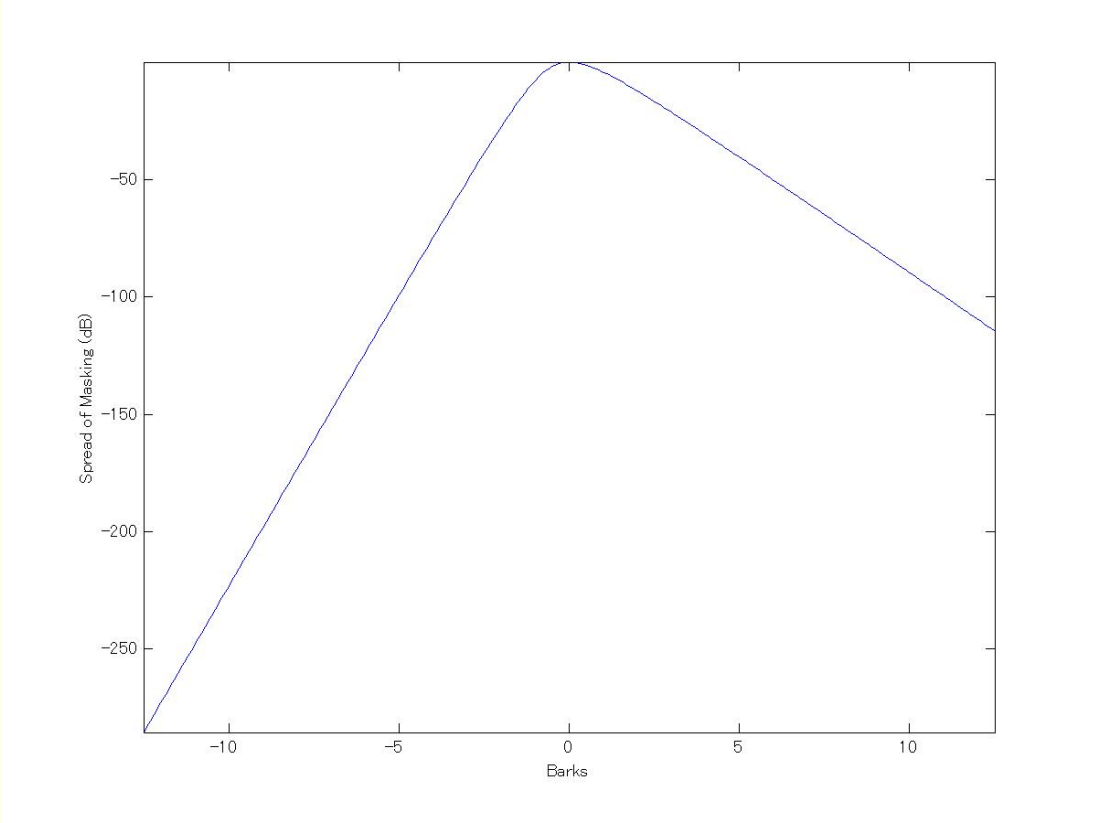
Marker-Maskee Experiment for Critical Bands

Spread of Masking - Inter-band masking

Inter-band masking has also been observed, i.e., a masker centered within one critical band has some predictable effect on detection thresholds in other critical bands. This effect, also known as the spread of masking, is often modeled in coding applications by an approximately triangular spreading function which has slopes of +25 and -10 dB per bark. A convenient analytical expression is given by:

$$SF_{dB}(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2}$$

where x has units of barks and $SF_{dB}(x)$ is expressed in dB.



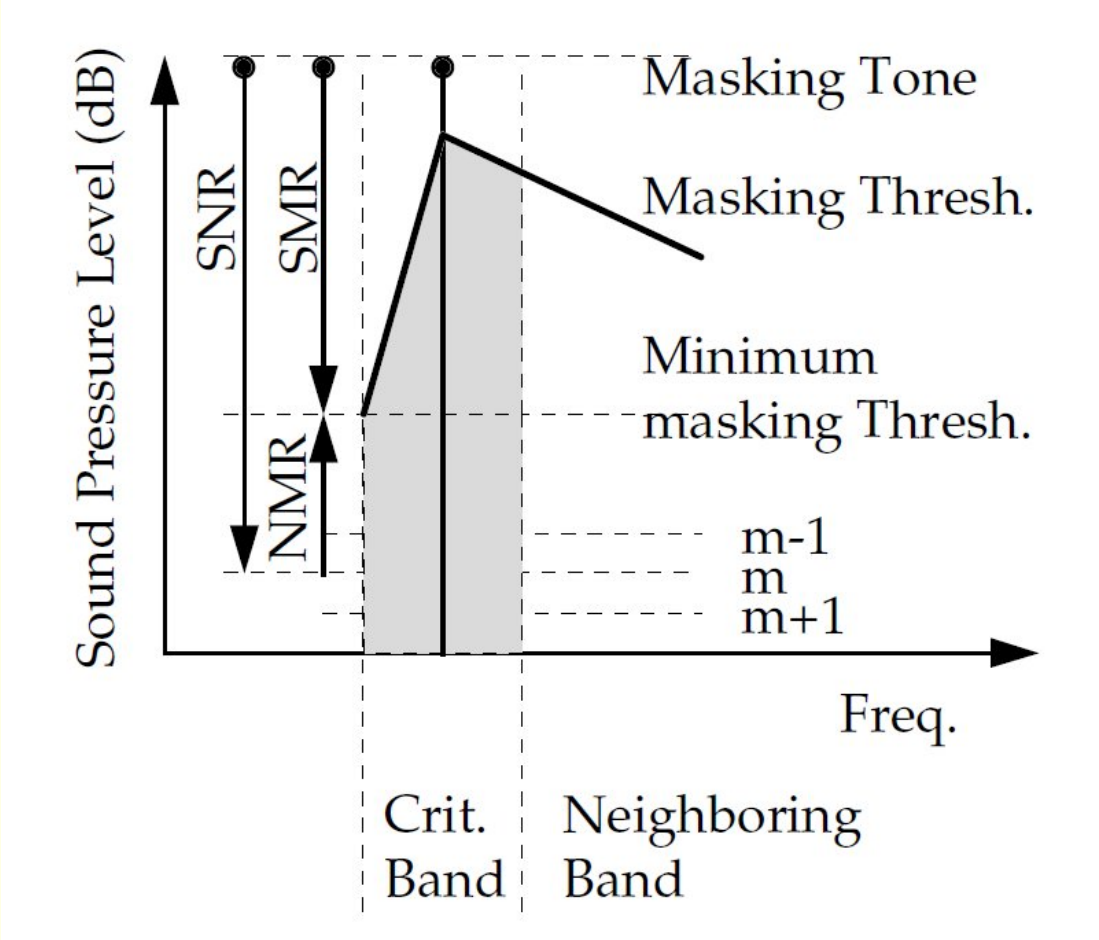
Spreading of Masking - Approximated Function

problem

Plot the above function of spread of masking for the bark range from -12.5 to +12.5 and confirm that the slope of the curve is +25 dB before the masking tone at 0 bark, and -10dB after the masking tone.

```
x0=[-12.5:0.1:12.5];
sfv=[];
for x=x0
sfv=[sfv,SF(x)];
end
plot(x0,sfv); axis tight;
xlabel('Barks'); ylabel('Spread of Masking (dB)');

function valSF = SF(x)
valSF=15.81+7.5*(x+0.474)-17.5*sqrt(1+(x+0.474)^2);
```



Conceptual Diagram of Masking Effects

Notions of critical bandwidth and simultaneous masking in the audio coding context give rise to some convenient terminology

illustrated in Fig., where we consider the case of a single masking tone occurring at the center of a critical band. All levels in the figure are given in terms of dB. A hypothetical masking tone occurs at some masking level. This generates an excitation along the basilar membrane which is modeled by a spreading function and a corresponding masking threshold. For the band under consideration, the minimum masking threshold denotes the spreading function in-band minimum. Assuming the masker is quantized using an m-bit uniform scalar quantizer, noise might be introduced at the level m. Signal-to-mask ratio (SMR) and noise-to-mask ratio (NMR) denote the log distances from the minimum masking threshold to the masker and noise levels, respectively.

9 Application of Psychoacoustic Principles: ISO 11172-3 (MPEG-1) PSYCHOACOUSTIC MODEL 1

- It is useful to consider an example of how the psychoacoustic principles described thus far are applied in actual coding algorithms. The ISO/IEC 11172-3 (MPEG-1, layer 1) psychoacoustic model 1 determines the maximum allowable quantization noise energy in each critical band such that quantization noise remains inaudible.
- In one of its modes, the model uses a 512-point DFT for high resolution spectral analysis (86.13 Hz), then estimates for each input frame individual simultaneous masking thresholds due to the presence of tone-like and noise-like maskers in the signal spectrum. A global masking threshold is then estimated for a

subset of the original 256 frequency bins by (power) additive combination of the tonal and non-tonal individual masking thresholds.

- This section describes the step-by-step model operations. The five steps leading to computation of global masking thresholds are as follows:

1. Spectral Analysis and SPL (Sound Pressure Level) Normalization
2. Identification of Tonal and Noise Maskers
3. Decimation and Reorganization of Maskers
4. Calculation of Individual Masking Thresholds
5. Calculation of Global Masking Thresholds

9.1 Spectral Analysis and SPL Normalization

First, incoming audio samples of b bit integer, $s(n)$, are normalized according to the FFT length, N , and the number of bits per sample (signed integer), b, using the relation

$$x(n) = \frac{s(n)}{N\left(2^{b-1}\right)}$$

Normalization references the power spectrum to a 0-dB maximum.

The normalized input, $x(n)$, is then segmented into 12 ms frames (512 samples) using a 1/16th overlapped Hann window such that each frame contains 10.9 ms of new data. A power spectral density (PSD) estimate, $P(k)$, is then obtained using a 512-point FFT.

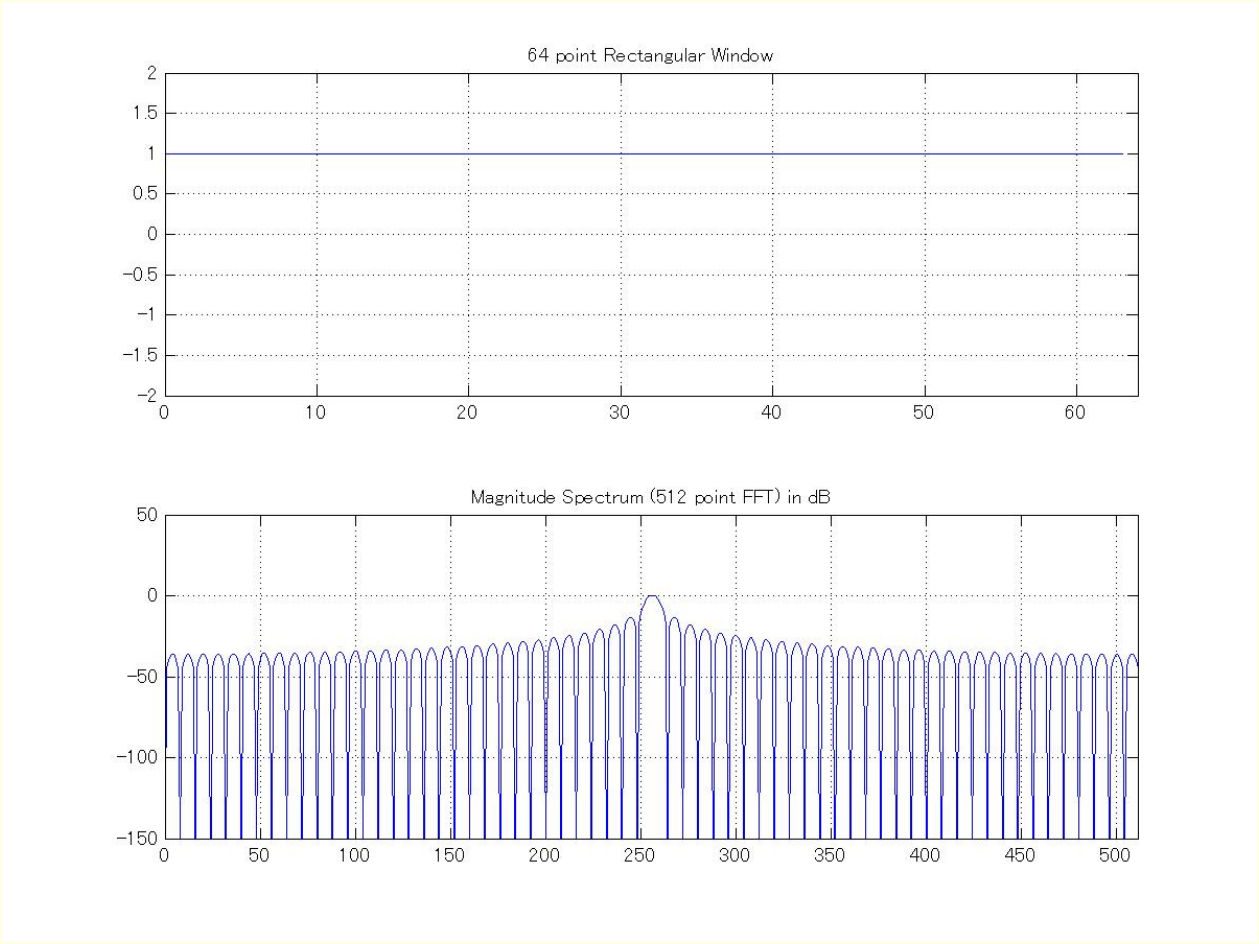
$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi nk}{N}}$$

$$X(k) = \sum_{n=0}^{N-1} x(n)w(n)e^{-j\frac{2\pi nk}{N}}.$$

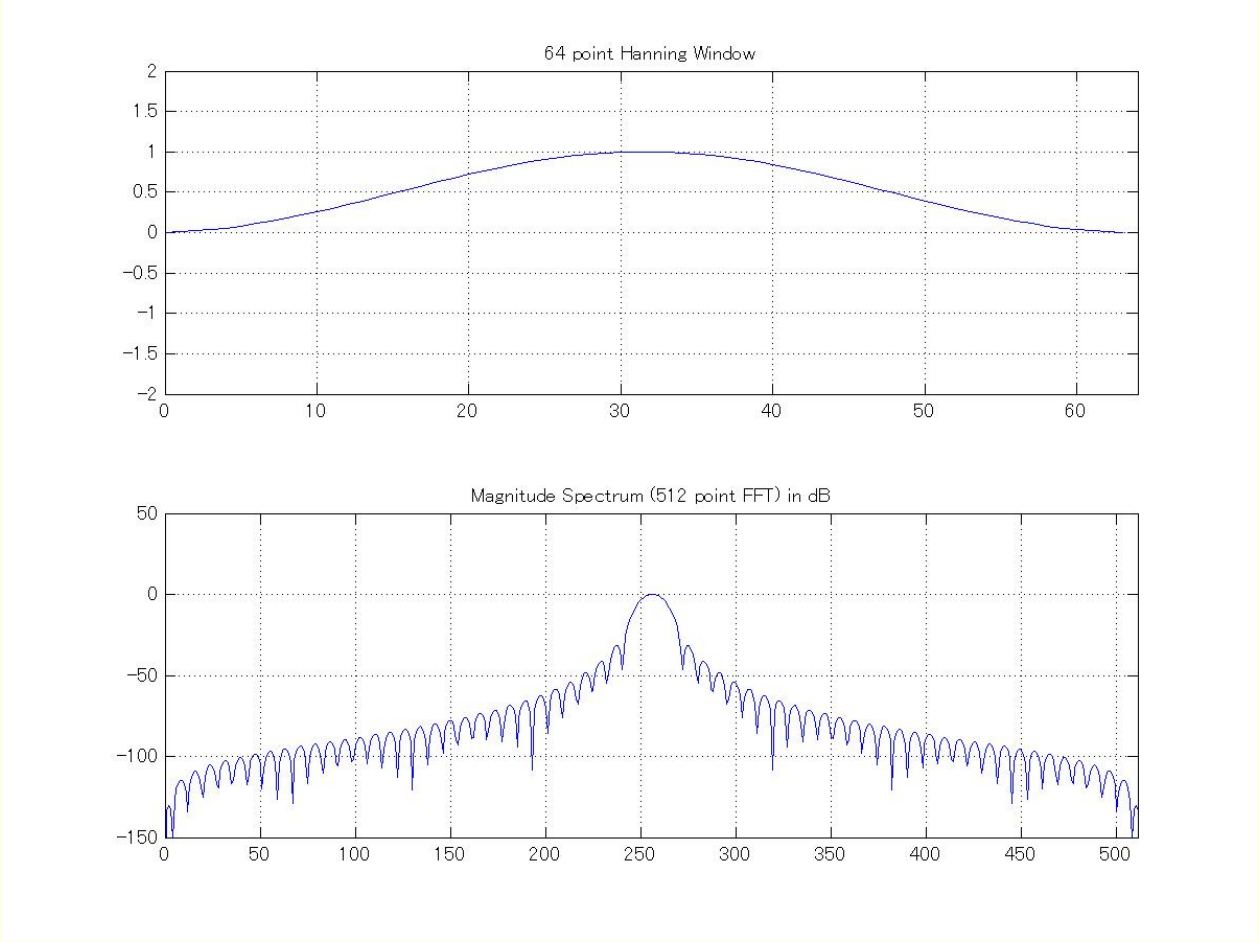
The Hanning window (Hann window) defined by

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

is used to reduce the spectrum leakage from other frequencies to the analysing frequency.



Spectrum of Rectangular (time) Window



Spectrum of the Hanning Window

A power spectral density (PSD) estimate, $P(k)$, is then obtained from $X(k)$ computed by a 512-point FFT (Fast Fourier Transform), a fast algorithm to compute DFT (Discrete Fourier Transform). PSD resulting from 512 FFT has 256 spectral components (harmonics).

$$P(k) = PN + 10 \log_{10} |X(k)|^2 \quad \text{for} \quad 0 \leq k \leq \frac{N}{2}$$

where the power normalization term, PN, is the reference sound pressure level of 96 dB.

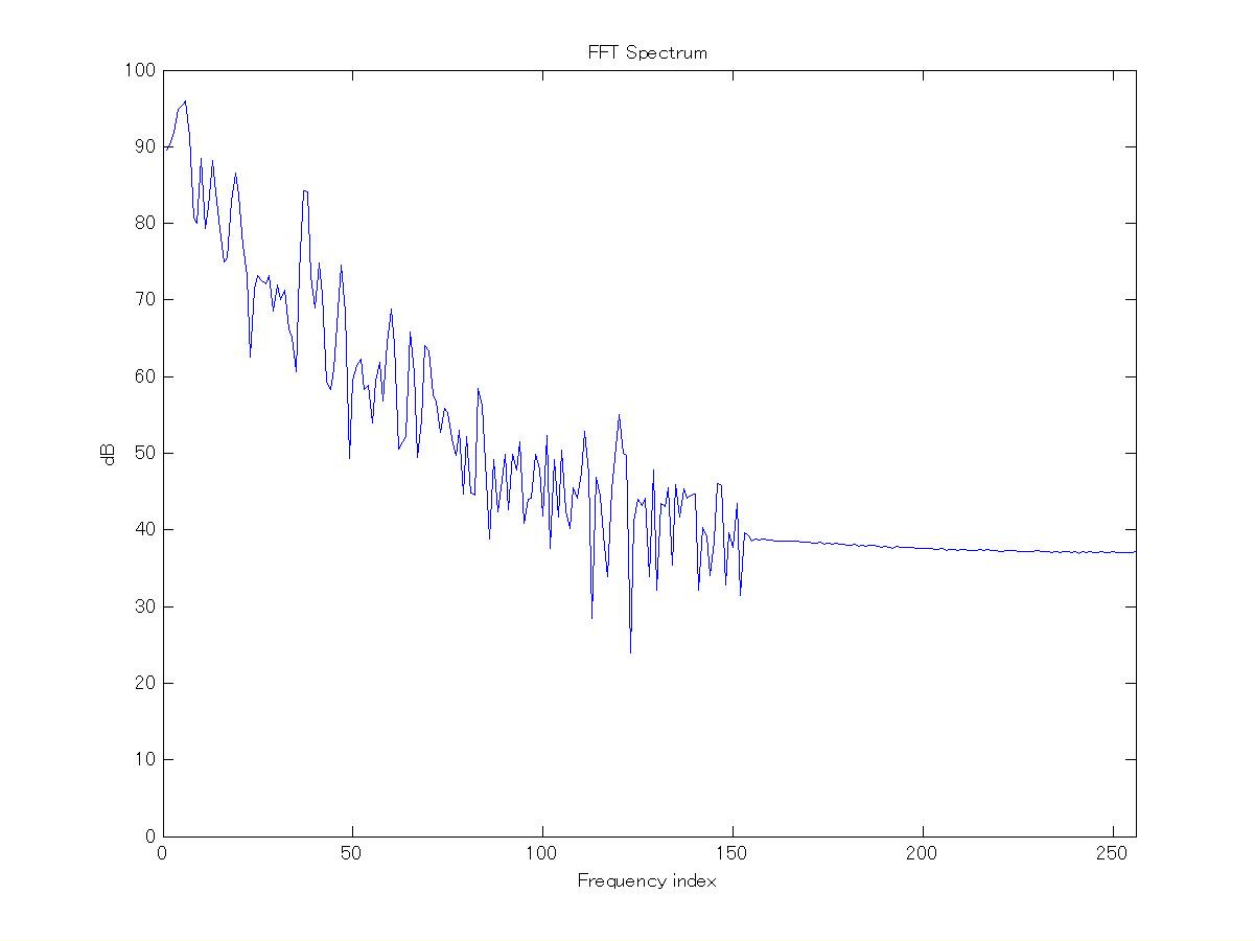
Problem

Matlab_MPEG_1_2_4.zip contains a MATLAB program that simulates all of MP3 spychoacoustic masking threshold calculations. A subroutine FFT_Analysis.m calculates Power Spectral Density (PSD). Main program is Test_MPEG.m. Apply this program to a music piece in *.wav of your choice to see its PSD. Slide the time window of 512 samples to find the first block so that no zero padding is applied to the analysis. The PSD of "Eine Kleine Nachtmusik" by Mozart is shown below. The key part of processing in FFT_Analysis.m is shown below.

```
%   Compute the auditory spectrum using the Fast Fourier Transform.
%   The spectrum X is expressed in dB. The size of the transform si 512 and
%   is centered on the 384 samples (12 samples per subband) used for the
%   subband analysis. The first of the 384 samples is indexed by n:
%   ..............................................
%      |       |  384 samples    |        |
%      n-64    n                 n+383    n+447
%   A Hanning window applied before computing the FFT.
%
% Prepare the Hanning window
h = sqrt(8/3) * hanning(FFT_SIZE);

% Power density spectrum
X = max(20 * log10(abs(fft(s .* h)) / FFT_SIZE), MIN_POWER);

% Normalization to the reference sound pressure level of 96 dB
Delta = 96 - max(X);
X = X +  Delta;
```



PSD of “Eine Kleine Nachtmusik” by Mozart

9.2 Identification of Tonal and Noise Maskers

After PSD estimation and SPL normalization, tonal and non-tonal masking components are identified.

Tonal maskers

Local maxima in the sample PSD which exceed neighboring components within a certain bark distance by at least 7 dB are classified as tonal. Specifically, the tonal set, S_T, is defined as

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

where,

$$\Delta_k \in \begin{cases} 2 & 2 < k < 63 & \text{0.17-5.5 KHz} \\ (2,3) & 63 \leq k < 127 & \text{5.5-11 KHz} \\ (2,\cdots,6) & 127 \leq k < 256 & \text{11-20 KHz} \end{cases}$$

Tonal maskers, $P_{TM}(k)$, are computed from the spectral peaks listed in S_T as follows

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

Noise maskers

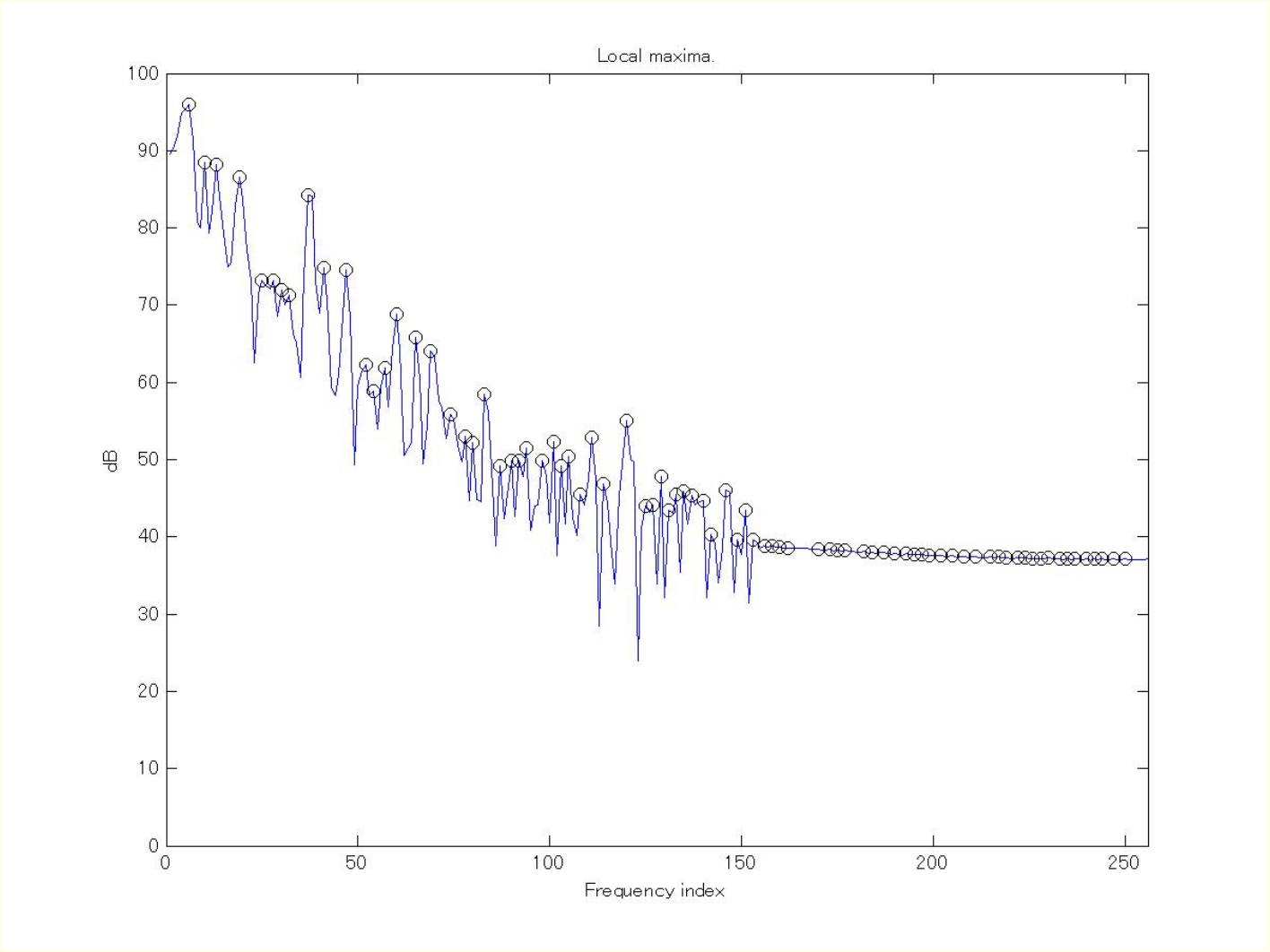
A single noise masker for each critical band, $P_{NM}(\bar{k})$, is then computed from (remaining) spectral lines not within the $\pm\Delta_k$

neighborhood of a tonal masker using the sum,

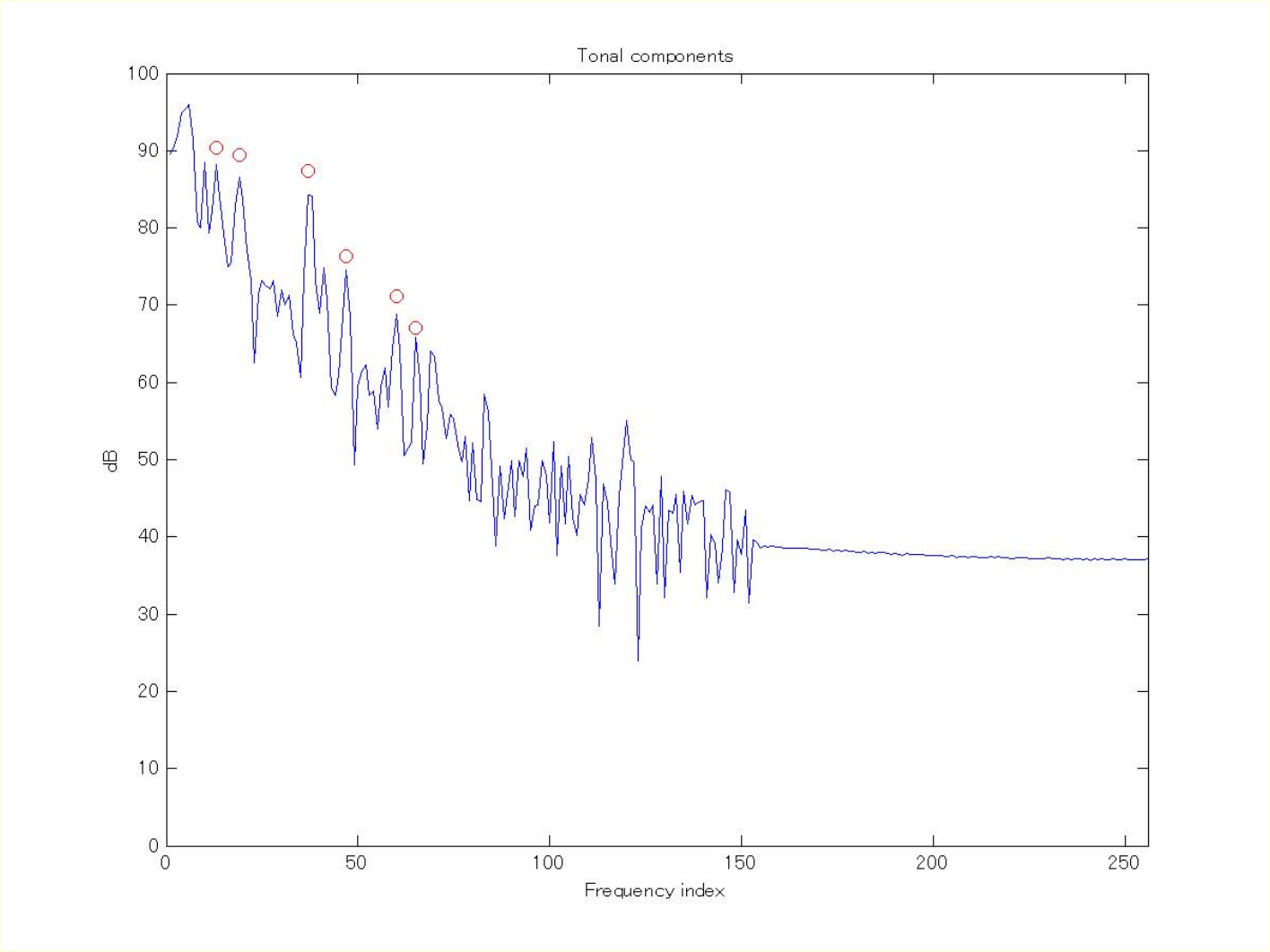
$$P_{NM}(\bar{k}) = 10\log_{10}\sum_{j} 10^{0.1P(j)} \quad \text{dB}$$

for all $P(j)$ not the member of $P_{TM}(k, k \pm 1, k \pm \Delta_k)$

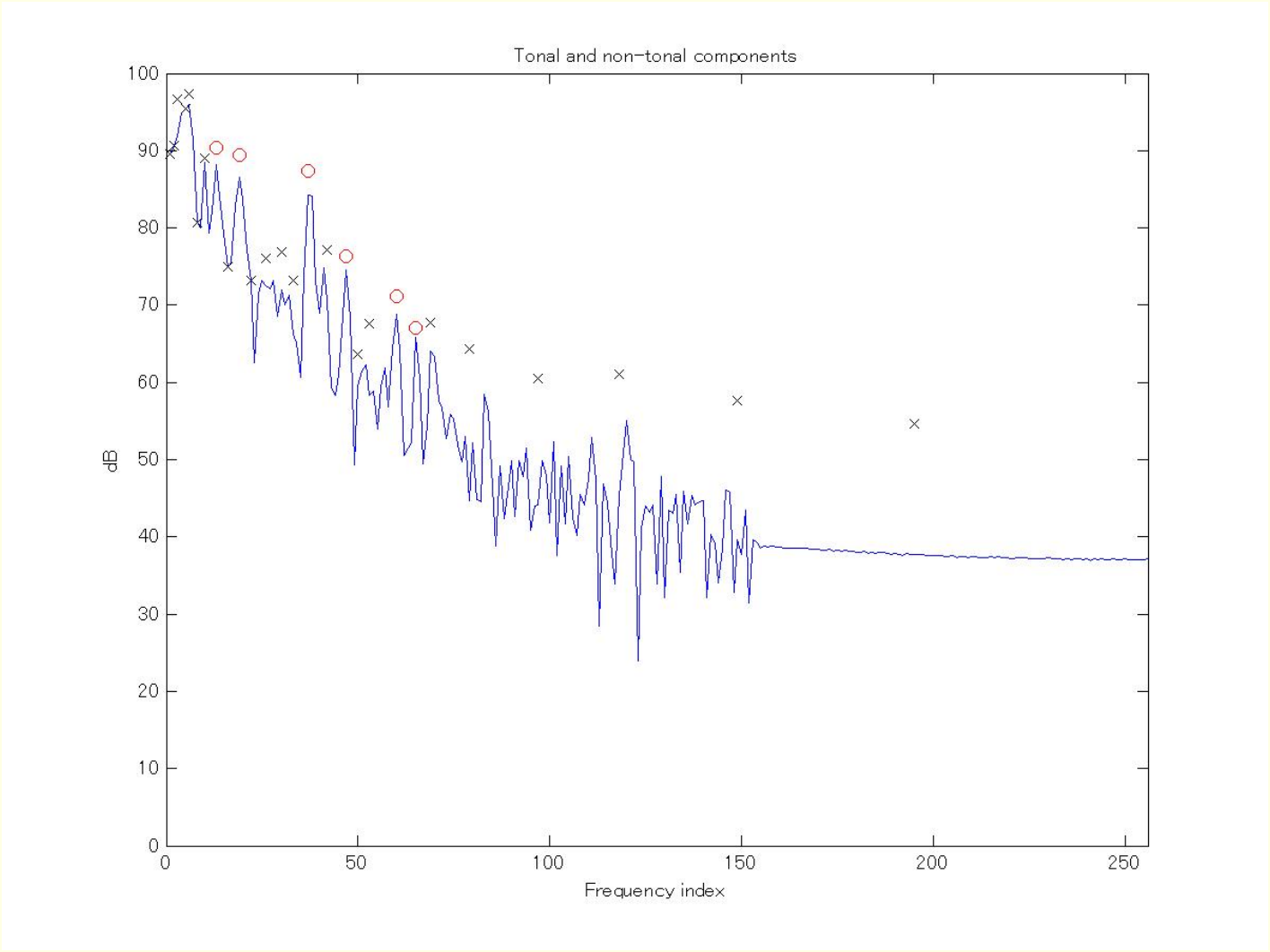
where, $\bar{k} = \left(\prod_{j=l}^{u} j\right)^{\frac{1}{u-l+1}}$ and l and u are the lower and upper spectral line boundaries of the critical band, respectively.



(1) local maxima



(2) tonal components



(3) tonal and non-tonal components of Eine Kleine Nachtmusik

Problem

A subroutine Find_tonal_components.m contained in the MP3 spychoacoustic masking simulation program Matlab_MPEG_1_2_4.zip first calculates the local maxima of Power Spectral Density (PSD). From the obtained local maxima of PSD, tonal components are calculated based on Equations described above. Then, non-tonal components and the frequencies of the critical band are calculated. Main program is Test_MPEG.m. Apply this program to a music piece in *.wav chosen in the previous **Problem** to show the 3 figures generated by Find_tonal_components.m, (1) local maxima, (2) tonal components, and (3) tonal and non-tonal components.

9.3 Decimation and Reorganization of Maskers

In this step, the number of maskers is reduced using two criteria. First, any tonal or noise maskers below the absolute threshold are discarded, i.e., only maskers which satisfy

$$P_{TM,NM}(k) \geq T_q(k)$$

are retained, where $T_q(k)$ is the SPL of the threshold in quiet at spectral line k. Next, a sliding 0.5 Bark-wide window is used to replace any pair of maskers occurring within a distance of 0.5 Bark by the stronger of the two.

After the sliding window procedure, masker frequency bins are

reorganized according to the subsampling scheme,

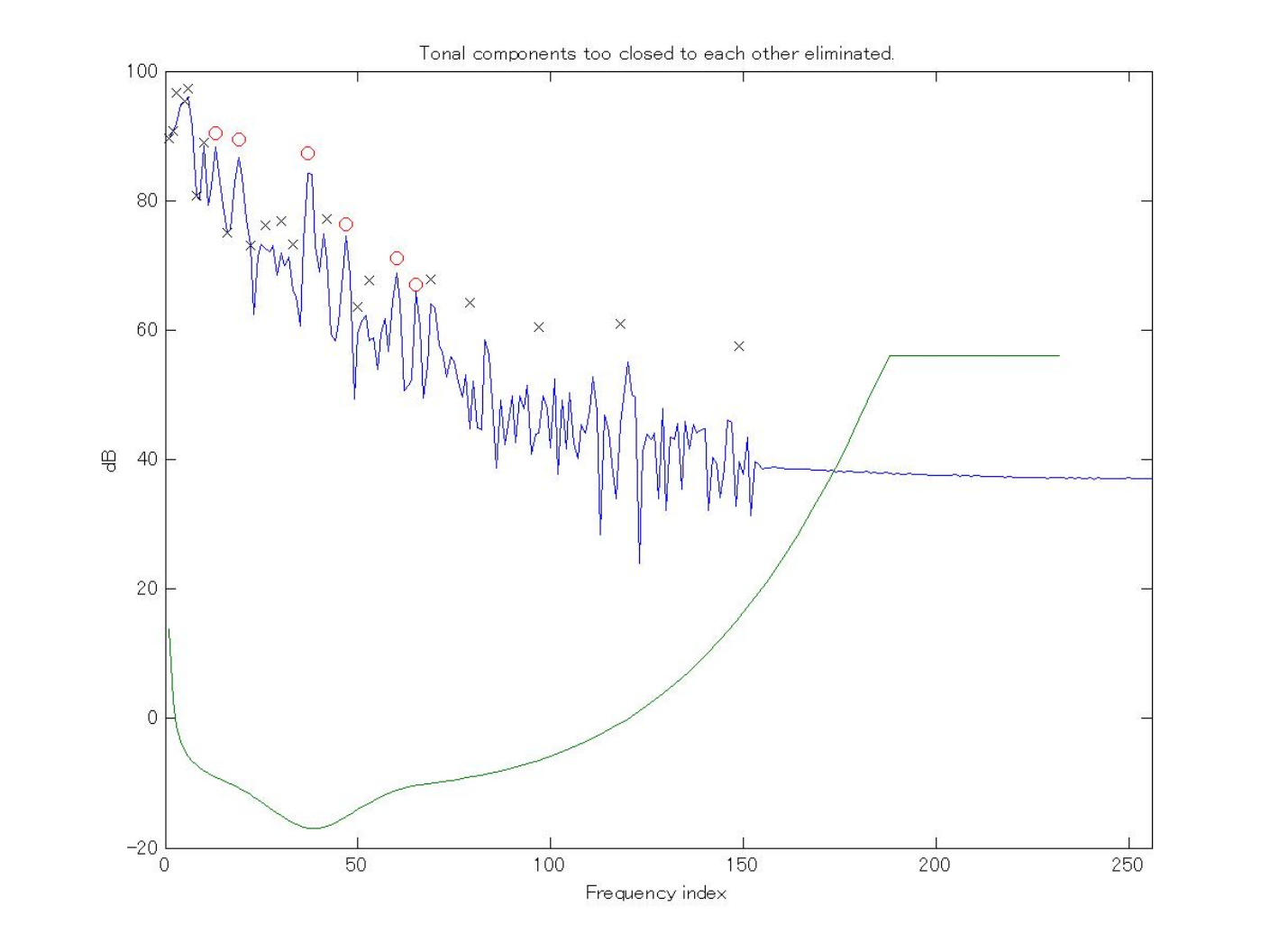
$$P_{TM,NM}(i) = \begin{cases} P_{TM,NM}(k) & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases}$$

The net effect is 2:1 decimation of masker bins in critical bands 18-22 and 4:1 decimation of masker bins in critical bands 22-25 , with no loss of masking components. This procedure reduces the total number of tone and noise masker frequency bins under consideration from 256 to 106. An example of decimation for the equal SPL is shown in the table below.

k	i	decimate
50	50	keep
51	52	zero
52	52	keep
100	100	keep
101	104	zero
102	104	zero
103	104	zero
104	104	keep

Problem

A subroutine Decimation.m— contained in the MP3 spychoacoustic masking simulation program Matlab_MPEG_1_2_4.zip does all processes of decimination described in this sub-section. Apply this program to a music piece in *.wav chosen in the previous **Problem** to see if any of SPL's are elimnated due to (1) any tonal or noise maskers are below the absolute threshold, (2) any pair of maskers occurring within a distance of 0.5 Bark is replaced by the stronger of the two. (3) 2:1 decimation of masker bins in critical bands 18-22 and 4:1 decimation of masker bins in critical bands 22-25.



Tonal and non-tonal maskers after decimation. Only one non-tonal masker SPL under the absolute threshold was eliminated.

9.4 Calculation of Individual Masking Thresholds

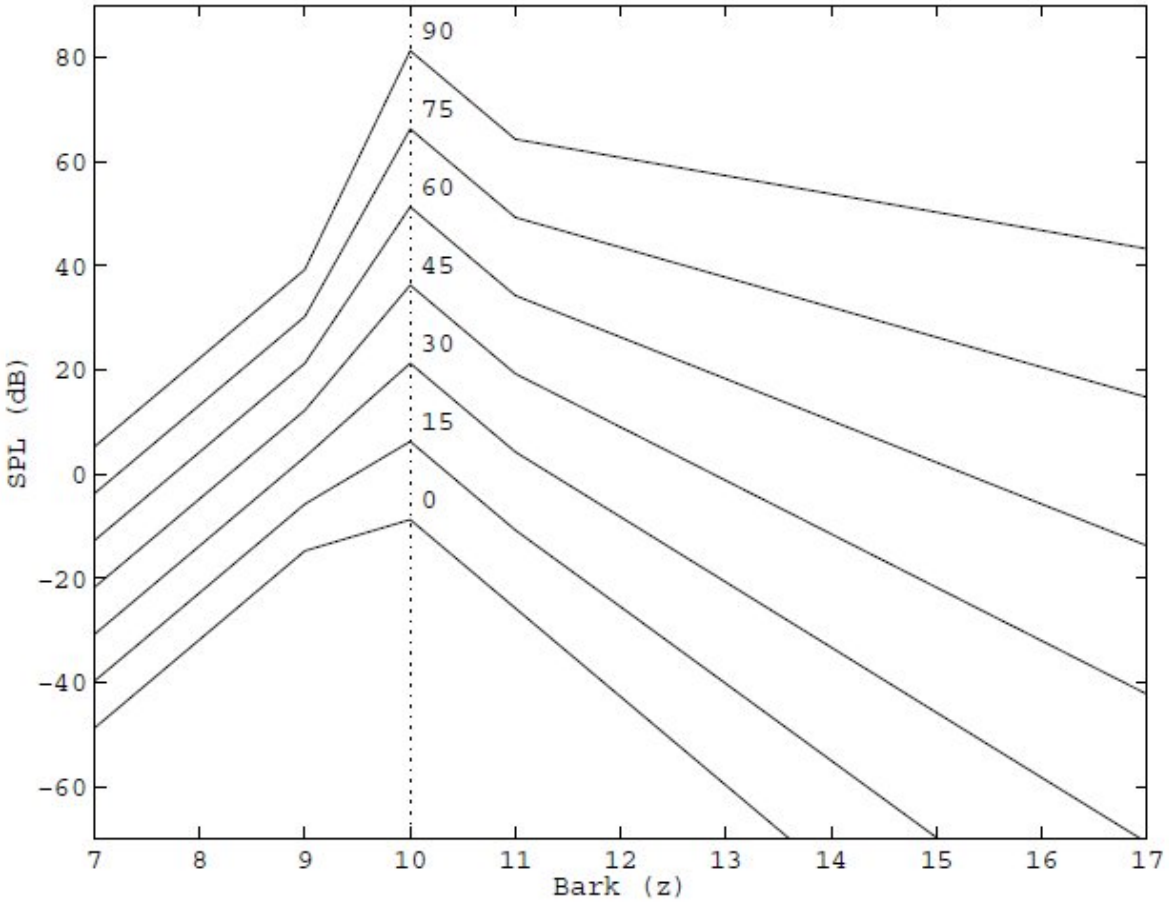
Having obtained a decimated set of tonal and noise maskers, individual tone and noise masking thresholds are computed next. Each individual threshold represents a masking contribution at frequency bin i due to the tone or noise masker located at bin j (reorganized during step 3). Tonal masker thresholds, $T_{TM}(i,j)$, are given by

$$T_{TM}(i,j) = P_{TM}(j) - 0.275z(j) + SF(i,j) - 6.025 \quad \text{dB}$$

where $P_{TM}(j)$ denotes the SPL of the tonal masker in frequency bin j, $z(j)$ denotes the Bark frequency of bin j,

and the spread of masking from masker bin j to maskee bin i, $SF(i,j)$, is modeled by the expression,

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



Prototype spreading functions at z=10 as a function of masker level

$SF(i,j)$ is a piecewise linear function of masker level, $P_{TM}(j)$, and Bark maskee-masker separation, $\Delta_z = z(i) - z(j)$. $SF(i,j)$ approximates the basilar spreading (excitation pattern) given. As shown in the figure, the slope of $T_{TM}(i,j)$, decreases with increasing masker level. This is a reflection of psychophysical test results, which have demonstrated that the ear's frequency selectivity decreases as stimulus levels increase. It is also noted here that the spread of masking in this particular model is constrained to a 10-Bark neighborhood for computational efficiency. This simplifying assumption is reasonable given the very low masking levels which occur in the tails of the basilar excitation patterns modeled by $SF(i,j)$.

Individual noise masker thresholds, $T_{NM}(i,j)$, are given by

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

where $T_{NM}(i,j)$ denotes the SPL of the noise masker in frequency bin j, $z(j)$ denotes the Bark frequency of bin j, and $SF(i,j)$ is obtained by replacing $P_{TM}(j)$ with $P_{NM}(j)$.

Problem

A subroutine Individual_masking_thresholds.m contained in the MP3 spychoacoustic masking simulation program Matlab_MPEG_1_2_4.zip calculates individaul masking thresholds of tonal maskers $T_{TM}(i,j)$, and non-tonal maskers $T_{NM}(i,j)$ using the spreading function $SF(i,j)$. Apply this program to a music piece in *.wav chosen in the previous **Problem** to plot the individaul masking thresholds of a frame.

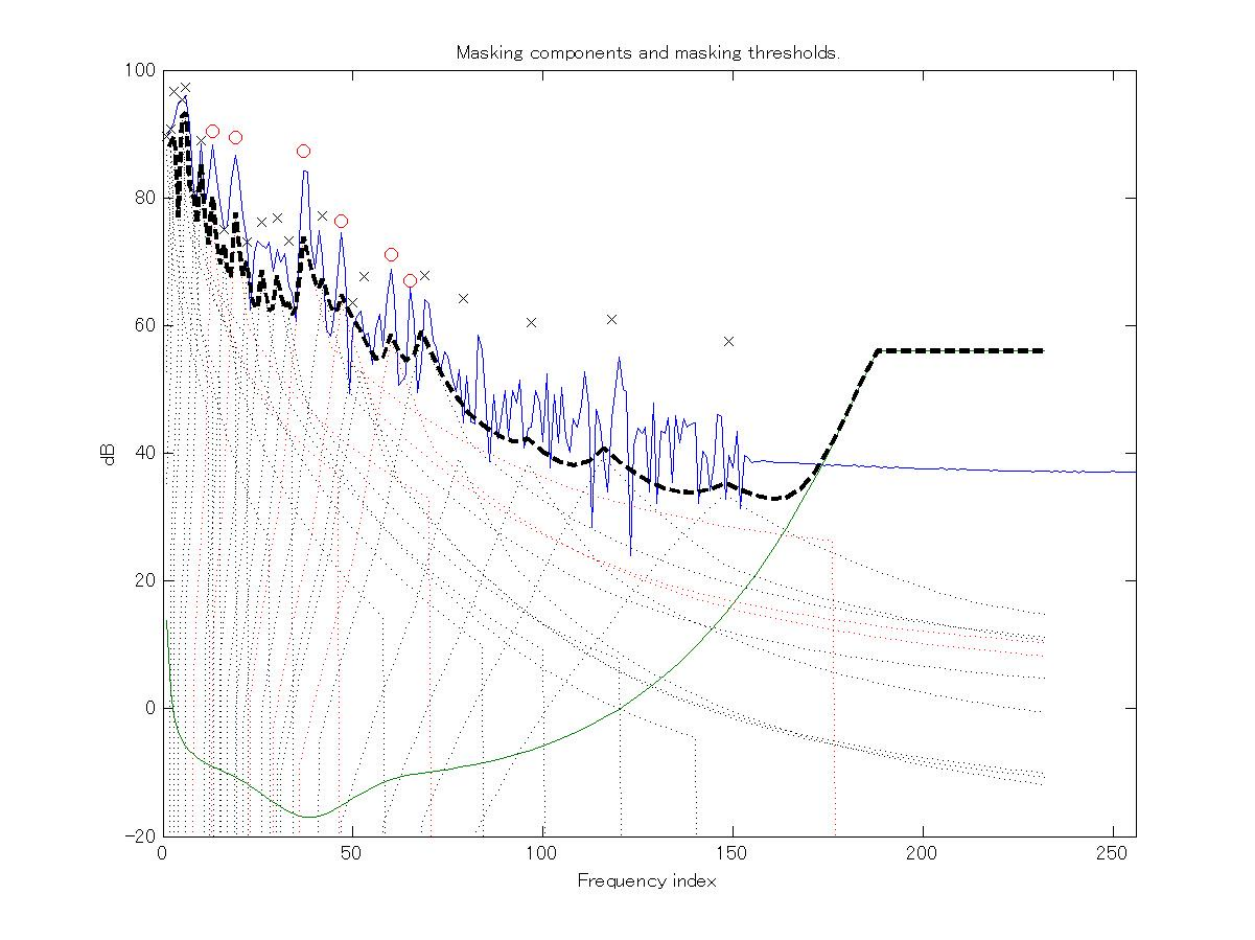
9.5 Calculation of Global Masking Thresholds

In this step, individual masking thresholds are combined to estimate a global masking threshold for each frequency bin in the subset given by Eq. 9.4. The model assumes that masking effects are additive. The global masking threshold, $T_g(i)$, is therefore obtained by computing the sum,

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

where $T_q(i)$ is the absolute hearing threshold for frequency bin i, $T_{TM}(i,l)$ and $T_{NM}(i,m)$ are the individual masking thresholds, and L and M are the number of tonal and noise maskers, respectively, identified previously.

In other words, the global threshold for each frequency bin represents a signal dependent, power additive modification of the absolute threshold due to the basilar spread of all tonal and noise maskers in the signal power spectrum. The next Fig. shows global masking threshold obtained by adding the power of the individual tonal and noise maskers to the absolute threshold in quiet.



Individaul masking thresholds for both tonal and non-tonal maskers. The global masking threshold is the sum of all individual masking thresholds.

10 End

$$R^{\mu}_{\ \nu} - \frac{1}{2} R\, \delta^{\mu}_{\ \nu} = \frac{8\pi G}{c^4} T^{\mu}_{\ \nu}$$

Here $T^{\mu}_{\ \nu}$ is tensor of energy momentum.

black	blue
red	magenta
green	cyan
yellow	