EE6624 Questions for Weeks 1 and 2

Semester 2 2019

sound intensity level

1. A single violin produces a level of 60 dB SPL. What level is produced by 10 violins if all are played with equal intensity? Assume the signals are uncorrelated. (Ans: $I = 20 \times 10^{-6}$)

Solution:

The SPL of 1 violin at 60 dB is given by

$20\log_{10}(I/(20 \times 10^{-6}) = 60$

$I = 20 \times 10^{-3}$

10 violins is $10 \times I = 20 \times 10^{-2}$

SPL of 10 violins in dB is $20\log_{10}[(20 \times 10^{-2})/(20 \times 10^{-6})]= 80$ dB

2. "0 dB SPL means there is no sound intensity" - Indicate this statement as True or False?

Solution:

False. Because $20\log_{10}(I/(20 \times 10^{-6}) = 0$, $I = 20 \times 10^{-6}$.

3. The threshold for detecting a 1,000 Hz tone is 20 dB when no other sounds are present. When a 1,050-Hz tone is presented at the same time as the 1,000-Hz tone, the threshold for detecting the 1,000 Hz tone is 30 dB. Is there masking? If so, which tone is the signal and which is the masker? How much masking is there in decibels (Ans: 10 dB)?

Solution:

Yes, there is masking, as the threshold of detection has been raised by 10dB. From the question, it is clear that we are trying to detect the 1 kHz tone which makes it the SIGNAL and the 1050 Hz tone becomes the MASKER. Amount of Masking = 30dB-20dB=10dB

4. What is *inverse square law*? Calculate the intensity drop if the distance from a point source is doubled? State if there is any assumption. (Ans : 0.25)

Solution:
The law of decreasing power per unit area (intensity) of a wavefront with increasing distance from the source is known as the inverse square law, because intensity drops in proportion to the inverse square of the distance from the source.

$I \propto 1/r^2$ where I : sound intensity; r : distance

If distance is doubled, the intensity will drop by a factor of 4.

Also $I_2/I_1 = r_1^2/r_2^2$

If $r_2 = 2r_1$, $I_2/I_1 = ¼ = 0.25$. Hence $I_2 = 0.25I_1$

Answer in dB will be $10\log_{10}0.25 = -6$ dB

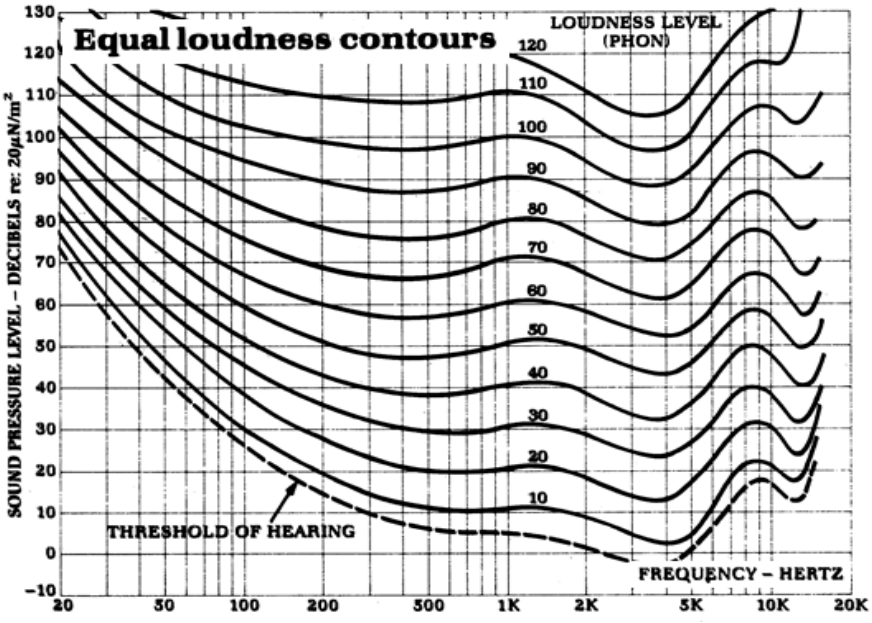
5. In designing a video game, a loud sound came on before an important soft sound and as such the soft sound is often masked. To solve this problem the designer simply made the soft sound come on first before the loud sound came on. Would this actually solve the masking problem?

Solution:
This may or may not solve the masking problem. The masking will change from post-masking (forward masking) to pre-masking (backward masking). However pre-masking is less effective than post-masking.

The exact answer will depends on the time gap in both cases and the signal strength. If the time gap is zero, it is safe to say that the signal will still be masked.

刻度

6. After proper calibration, a student tested in the classroom had thresholds of 23.9 dB SPL at 1,000 Hz, 30.6 dB SPL at 2,000 Hz, and 44.4 dB SPL at 4,000 Hz. These results suggest that the child has a hearing loss. Why? How many decibels in the hearing loss at each of the three frequencies given the Standardized Thresholds of Audibility?



Solution:

From the equal loudness curves, the threshold of hearing for a normal person at 1kHz, 2kHz and 4kHz is 4dB, 2dB and -3dB approximately. The child exhibits symptoms of deafness especially for the higher frequencies. The amount of hearing loss at

1kHz : (23.9 – 4) = 19.9 dB

2kHz : (30.6 – 2) = 28.6 dB

4kHz : (44.4 + 3) = 47.4 dB

7. In the laboratory, a listener's hearing is tested using headphones. The calibration for that headphone at 2,000 Hz shows that with 1 volt of sound level as the input to the headphone, the headphone produces 10 dB SPL as the output sound pressure level. The voltage into the headphone is 0.5 volts when the listener is at her threshold for detecting the 2,000-Hz tone. What is her threshold in dB SPL, given the calibration data and assuming the audio equipment is operating in a linear manner. Show your calculations? (Ans: ≈ 4 dB)

Solution:

1 volt results in 10 dB SPL. For 10 dB, the intensity will be

$10\log_{10}(I_1/10^{-12}) = 10 \rightarrow (I_1/10^{-12}) = 10 \rightarrow I_1 = 10^{-11}\text{W/m}^2$

Because the intensity (power) is proportional to the square of voltage, the halved voltage level implies intensity is divided by 4.

Hence threshold of hearing at 2 kHz is $10\log_{10}(I_1/(4X10^{-12})) \approx 4$ dB

8. People exposed to loud sounds (e.g. those who fire guns) often have more hearing loss in the region of 4,000 Hz. Why?

Solution:

The auditory canal or ear canal has the highest gain at approximately 4 kHz. Other frequencies are attenuated to different extent.

减弱

9 A conference room with a dimension of 30 m(L) x 20 m(W) x 15 m(H), and has an average absorption coefficient of surface of 0.6 for all surface, except the ceiling which has an absorption coefficient of 0.1. Calculate the reverberation time (in sec) using the Sabine's formula (Ans: 1.11 s).

Solution

Based on the equation given in lecture notes $RT_{60} = \frac{0.163V}{\sum a_i s_i}$

V=30x20x15=9000m^3.

a_1s_1=30 x 20 x 0.6 = 360 m^2 (floor)

a_2s_2=30 x 20 x 0.1 = 60 m^2 (ceiling)

$2a_3s_3$=2x30 x 15 x 0.6 = 540 m^2 (long side)

$2a_4s_4$=2x15x 20 x 0.6 = 360 m^2 (short side)

RT60 = 0.163 x 9000/(360+60+540+360) = 1.11 second

10. Compute the data rates for the following audio signals:

a) A stereo signal sampled at 44.1 kHz using 16 bits/sample
b) Five channels surround sound sampled at 44.1 kHz using 16 bits/sample.
c) In addition to (b), also include 1 (low frequency enhanced) channel sampled at 500Hz using 16 bit/sample.

Solution:

a) Data rate = 44100 samples x 16 bits x 2 = 1.4112 Mb/s
b) Data rate = 44100 samples x 16 bits x 5 = 3.528 Mb/s
c) Data rate = 3.528 Mb + 500 sample x 16 bits x 1 = 3.536 Mb/s

11. It is required to store a two-hour 5-channel sound track using a format of 24-bit/sample and sampling at 96 kHz.

a) How much storage (in bits) and data rate are needed for storing this sound track?
b) A CD has a throughput of 1.411 Mbps, what is the compression ratio needed to store this track into a CD?
c) A DVD has a throughput of 6.144 Mbps, what is the compression ratio needed to store this track into a DVD?

Solution:

(a) The data rate = 96 k sample x 24 bits x 5 = **11.52 Mbps**

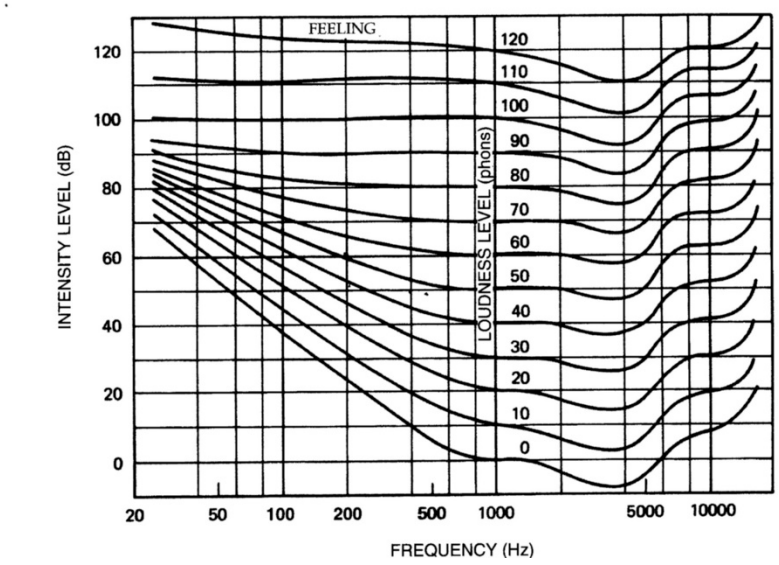
At a sampling frequency of 96 kHz, a two-hour sound track requires:
2 hr x 3600 sec/hr x 11.52 Mbps = **82.944 Gbits** (or 10.368 Gbytes)

(b) A CD has a throughput of 1.411 Mbps requires a compression ratio of 11.52/1.411 = **8.16 : 1**.

(c) A DVD has a throughput of 13.824 Mbps requires a compression ratio of 11.52/6.144 = **1.875:1**

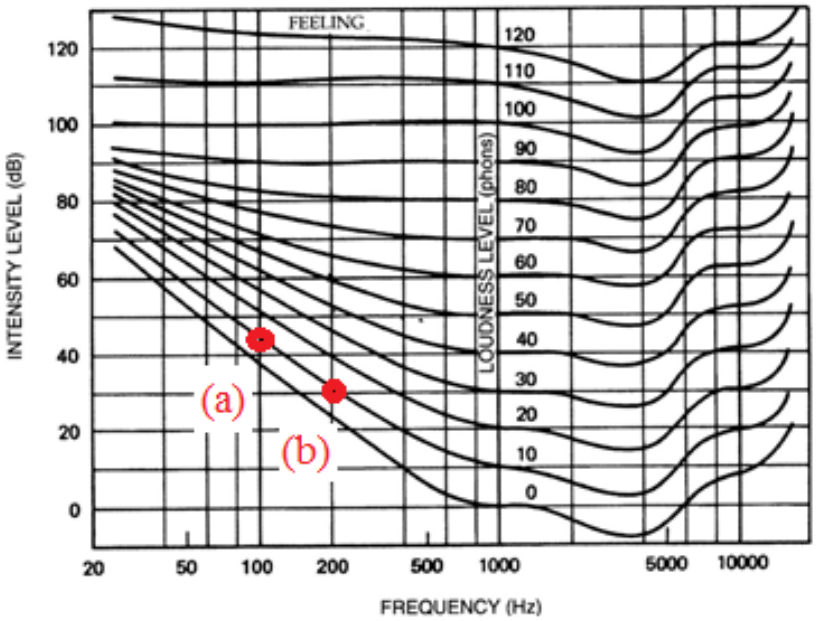
12. Exercises to test the understanding of the loudness level contours.

a) What is the intensity level of a pure tone at 100 Hz that has a loudness level of 10 phons?
b) What must the intensity level of 200Hz pure tone in order for it to sound just as loud as a 1000 Hz tone played at 10 dB?
c) What is perceived as louder, a 100 Hz tone played at 60 dB or a 10,000 Hz tone played at 60 dB?
d) What must be the intensity level of a 500Hz tone in order for it to be perceived at the same loudness as a 10,000 Hz tone played at 90 dB?
e) What is the intensity level of a 2000 Hz tone at 0 phon? What is the implication of this intensity level?
f) What is the total dB level of a 100 Hz tone and a 1000Hz tone each having an intensity of 20 dB, if they played together?
g) What is the total intensity level (in dB) of a 100 Hz tone and a 1000 Hz tone, each having a loudness level of 70 phons, if they are played together?

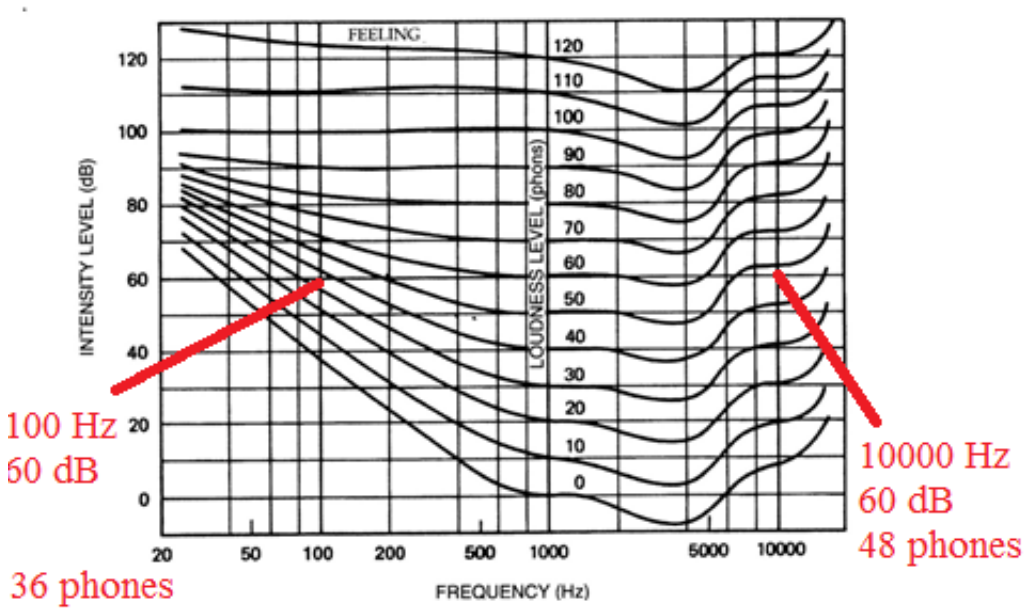


Solution:

a). 45 dB, b). 30 dB (see the figure below)



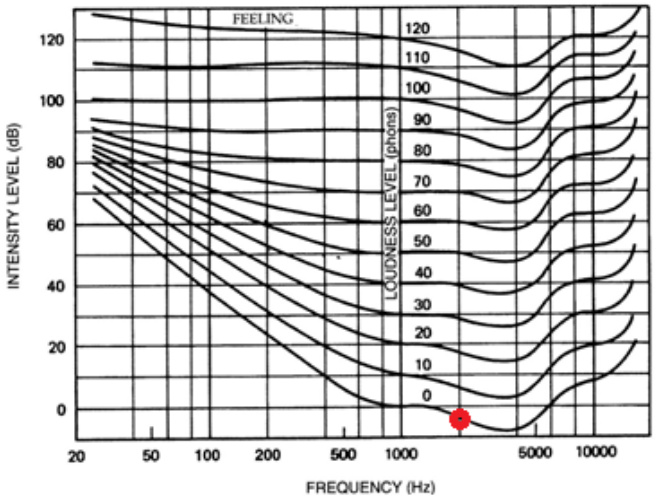
c). 100 Hz at 60 dB is perceived as 35 phons vs. 10,000 Hz at 60 dB is perceived as 48 phons. Therefore, the latter is louder.



d). What must be the intensity level of a 500Hz tone in order for it to be perceived at the same loudness as a 10,000 Hz tone played at 90 dB?
80 dB

e). What is the intensity level of a 2000 Hz tone at 0 phon? What is the implication of this intensity level?

–5 dB. The negative dB simply means that the intensity of the sound is below the min intensity of $I_{\min} = 10^{-12}\ W/m$



f). What is the total dB level of a 100 Hz tone and a 1000Hz tone each having an intensity of 20 dB, if they played together?

$$10\log\frac{I}{I_{\min}} = 20 \Rightarrow \frac{I}{I_{\min}} = 10^{(number\ of\ bels)} = 10^2$$

$$I_{combined} = 2x10^2 xI_{\min}$$

$$\therefore \frac{I_{combined}}{I_{\min}} = 2x10^2 = 10^{number\ of\ bels}$$

$$Number\ of\ bels = 2.301(or 23\,dB)$$

Increased by only 3 dB

(g) What is the total intensity level (in dB) of a 100 Hz tone and a 1000 Hz tone, each having a loudness level of 70 phons, if they are played together?

Need to convert phon level to dB.

A 1000 Hz tone at loudness level of 70 phons

→ 70 dB intensity.

A 100 Hz tone at 70 phon loudness level → 78 dB.

Or expressed by

$$\frac{I_{1000}}{I_{\min}} = 10^7 \qquad \frac{I_{100}}{I_{\min}} = 10^{7.8}$$

$$I_{tot} = (10^7 + 10^{7.8})xI_{\min} = 7.3x10^7 I_{\min}$$

$$Number\ of\ bels = \log(7.3x10^7) = 7.86(or\ 78.6dB)$$

Questions of EE6424 for Weeks 4 and 5

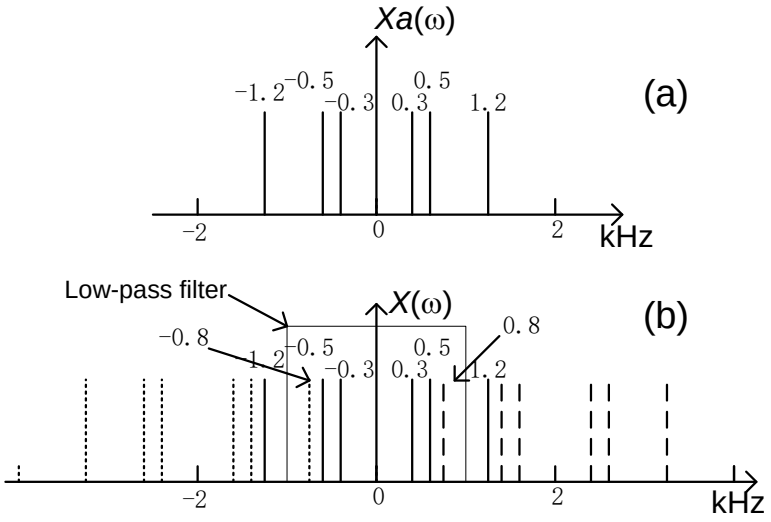
Semester 2 2018

Question 1

A continuous-time signal $x_a(t)$ is composed of a linear combination of sinusoidal signals of frequencies 300 Hz, 500 Hz, and 1.2 kHz. The signal $x_a(t)$ is sampled at 2.0 kHz rate, and the sampled sequence is passed through an ideal lowpass filter with a cutoff frequency of 850 Hz, generating a continuous-time signal $y_a(t)$. What are the frequency components present in the reconstructed signal $y_a(t)$?

Solution

It is known that $x_a(t)$ is being sampled at 2 kHz rate, and the Fourier transform Xa(w) of $x_a(t)$ has the frequencies of ±300 Hz, ±500 Hz, and ±1.2 kHz, as seen in figure (a) below.



Using aliasing formula, Figure (b) shows the frequency components in $x_a(f)$ + $x_a(f+F_s)$ + xa(f-F_s). For reconstruction the original signal from the samples, we use low pass filter (LPF) with a cut-off frequency of $F_s/2$=1kHz (see in figure (b)). All these frequency components smaller than 1 kHz will be used for reconstruction. Therefore, these frequency components are ±300 Hz, ±500 Hz, and ±0.8 kHz. Therefore, 0.8 kHz is the aliasing frequency.

Question 2

Find the 6-bit binary representation 0.725 by using the midtread and midrise quantizers. Also find the errors of these representations, respectively.

Solution:

Refer to the lecture notes. Also note that the number is positive and smaller than 1, the number of bits is 6.

(a) Midrise

The sign bit is 0 since the number is positive0. Based on

$$|\text{code}| = \begin{cases} 2^{B-1} - 1, & \text{when} \, |\text{number}| \geq 1 \\ \text{int}[2^{B-1}.|\text{number}|], & \text{elsewhere} \end{cases}$$

We have |code|=int[2^{6-1}*0.725]=int(23.2)=23 → 10111. Therefore, the code word is 010111. The number point is assumed to be after the sign bit.

Let us consider the de-quantization. The equation is

$$|\text{number}| = (|\text{code}| + 0.5) / 2^{B-1} = (23 + 0.5) / 2^5 = 0.734375$$

The error is |0.725 – 0.734375| = 0.009375.

(b) Midread

The encoding equation is

$$|\text{code}| = \begin{cases} 2^{B-1} - 1, & |\text{number}| \geq 1 \\ \text{int}\{[(2^B - 1).|\text{number}| + 1] / 2\}, & \text{elsewhere} \end{cases}$$

Then |code|=int[(2^6-1)*0.725+1]/2]=int[23.3375]=23 → 010111.

In this case, the midread representation is dequantized by

$$|\text{number}| = 2.|\text{code}| / (2^B - 1)$$

the error is |0.725-0.730 $= 2 \times 23 / (2^6 - 1) = 0.730158|$

Question 3

Quantization is the mapping of continuous amplitude values into codes that can be represented with a finite number of bits. Two commonly used uniform quantizers are the midtread and midrise quantizers.

i). Draw their input-output diagrams and state the differences between these quantizers.

ii). Select a suitable quantizer for audio input and explain your selection.

iii). If the overall input signal ranges from -3 to +3 v, determine the input range per code using 16-bit midread and midrise quantizers, and comment on their differences.

iv). What would happen when the amplitude is more than 3 v?

Solution

i) Plot the input output relation with the following explanations

- There are two commonly used uniform quantizes: midtreat and midrise.
- For a B bit number system, the former has 2^B-1 levels and the latter has 2^B levels.
- Also the former has a zero-valued level and the latter does not have any zero level.

ii) A suitable quantizer for audio input is the midtreat quantizer. The reason is that audio signal has higher dynamic range and it is sometimes required to represent a complete silence of a musical track, which is not achievable b midrise quantizer.

iii) If the overall inlut/output signal range is from -3 v to 3v, the input step size is

$$\Delta = 2 \times 3 / 2^{16} = 0.0915527 \text{ mV} \quad \text{(midtrise)}$$

$$\Delta = 2 \times 3 / (2^{16} - 1) = 0.09155413 \text{mV} \quad \text{(midtrise)}$$

The difference is not significant for a 16 bit quantization, which may not have noticeable performance difference.

iv) When the continuous amplitude is larger/smaller than 3 V, it is being mapped into the highest value of the output sample. In another words, the quantizer dips all output to highest or lowest possible level. In this case, the range of the output is still in the range of |3| v.

Question 4

Find the signal to noise ratio of a R bit fixed point system. It is assumed that the probability of rounding error is evenly distributed in a step size Δ. Verify the rule of thumb to be 6 dB SNR increase for each bit increase of the binary representation.

Solution:

Assuming the rounding error q is evenly distributed between the range of $-\Delta/2$ and $\Delta/2$, and its probability of error is $1/\Delta$. The energy of errors is

$$Q = \int_{-\Delta/2}^{\Delta/2} q^2 \frac{1}{\Delta} dq = \frac{\Delta^2}{12}$$

The uniform quantizer of R bits has the step size $\Delta = 2x_{max}/2^R$, where x_{max} is the maximum (or minimum) value of the range of the quantizer. Then

$$Q = \int_{-\Delta/2}^{\Delta/2} q^2 \frac{1}{\Delta} dq = \frac{x_{\max}^2}{3*2^{2R}}$$

If the input signal has energy E_{in}, the signal to noise ratio of the obtained by the quantizer is

$$\begin{aligned} SNR &= 10\log_{10}(\frac{E_{in}}{Q}) = 10\log_{10}(E_{in}\frac{3*2^{2R}}{x_{\max}^2}) \\ &= 10\log_{10}(\frac{E_{in}}{x_{\max}^2}) + 20*R*\log_{10}2 + 10*\log_{10}3 \\ &\approx 10\log_{10}(\frac{E_{in}}{x_{\max}^2}) + 6.021*R + 4.771 \end{aligned}$$

It is seen that about 6 dB increase in SNR is achieve when R is increased by 1. For example, a 16 bit representation generally has more than 90dB SNR for most of the time.

The actual SNR also depends on the energy of input signal E_{in}. Therefore, it is always desired to make E_{in} to be maximum.

Question 5

Table 1 lists five input symbols and their corresponding occurring probabilities and Huffman code words.

Table 1 Huffman code table

Input Symbol	Probability	Huffman code word
1	1/11	0011
2	1/11	0010
4	5/11	1
5	2/11	01
6	2/11	000

(i) Given the input data stream X = [4 5 6 6 2 5 4 4 1 4 4], compute the entropy of this data stream and the coding efficiency achieved by the code words listed in Table 1.

(ii) Describe the main steps of Huffman coding process.

Solution:

(i).

$$E = \sum p_n \log_2(1/p_n) = \sum p_n \log_{10}(1/p_n)/\log_{10} 2$$

$$= [\frac{1}{11}\log_{10}(11) + \frac{1}{11}\log_{10}(11) + \frac{5}{11}\log_{10}(11/5) + \frac{2}{11}\log_{10}(11/2) + \frac{2}{11}\log_{10}(11/2)]/\log_{10} 2$$

$$= [\frac{1\times 1.0413}{11} + \frac{1\times 1.0413}{11} + \frac{5\times 0.3424}{11} + \frac{2\times 0.7403}{11} + \frac{2\times 0.7403}{11}]/0.301 = 2.0404$$

The average length

$$L = \sum L_n p_n = 4*1/11 + 4*1/11 + 1*5/11 + 2*2/11 + 3*2/11$$

$$= (4+4+5+4+6)/11 = 23/11$$

Coding efficiency

$$E/L = 2.0404*11/23 = 0.9758$$

(ii) . The coding steps are as follows:

- Arranging all the symbols according to the descending order of their occurring probability;
- Combining the two symbols with the lowest probabilities by the calculating the total probability;
- Assigning 0 and 1 to the branches of these two symbols;
- Repeating all these above steps for other symbols;
- The code words are obtained by reading the 0 and 1 from the branches of the coding three.

Question 6

A signal is sampled at the frequency of 1024 Hz. It is processed by DFTs of length N=512 and 2048, respectively. The component at 256 Hz in the input signal is coincident on the valid index value of k=128, and the component of 270.5 Hz is located between k = 135 and 136.

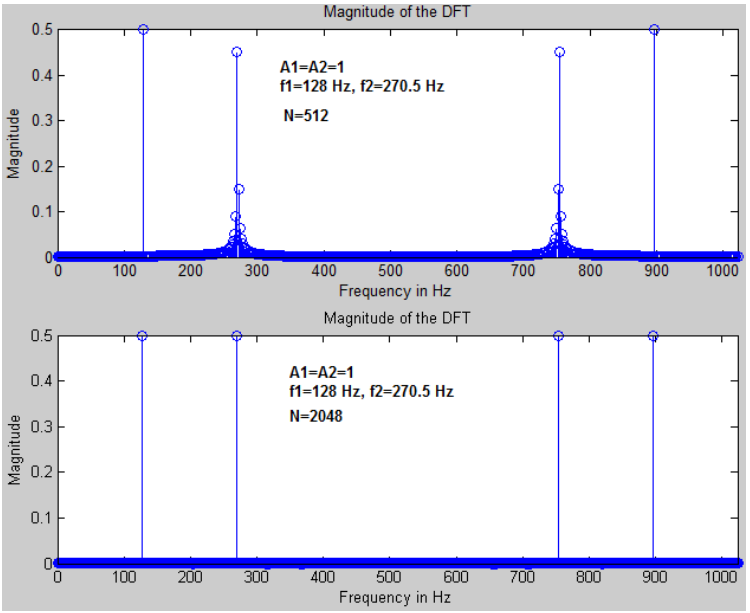
i). When N = 512, can we see these two components from the DFT outputs? Why?

ii). When N = 2048, can we have any different observation from that in i).

iii). Are there any general comments made by comparing the observation from i) and ii)?

Solution:

i). When N= 512, the frequency resolution is F_s/N=1024/512 = 2 Hz. The frequency component of 256 Hz ca be clearly observed as $X(128)$. However, the location of the component at 270.5 Hz is not on the any value of k. It is between k= 135 and 136. Therefore there is no valid point represented by the DFT.



However, this component can still be observed due to the windowing effects of the input segment. The energy of this component is distributed into a number of components. See the figure.

ii). When N= 2048, the frequency resolution is F_s/N=1024/2048 = 0.5 Hz is sufficient to represent the component at 270.5 Hz. Therefore, it can be seen as a single peak in the DFT.

iii). The general comments are:

- The window effect is desirable to observe the frequency components when the frequency resolution is not sufficient.
- Large value of N leads to high frequency resolution.
- The signal energy is always reserved in the DFT, but not concentrated at the signal frequency in the DFT when the frequency resolution is not sufficient.

Question 7

In practice, the lossy coding algorithm calculates the sound pressure level (SPL) in the frequency domain. The definition of SPL in dB for a sinusoidal wave is

$$\text{SPL} = 96\ \text{dB} + 10\log_{10}(A^2) \tag{1}$$

where it is assume that when the amplitude A=1, SPL =96 dB. It is also known that

$$A^2 = \frac{4}{N^2 < w >^2} \sum_{peaks} |X(k)|^2 \tag{2}$$

where $X(k)$ is the DFT value with a length of N and $<w^2>$ is the window power factor, which is 1 for rectangular window (putting (2) into (1), we get equation in page 13).

$$SPL_{DFT} = 96dB + 10\log_{10}\left[\frac{4}{N^2 < w >^2} \sum_{peaks} |X(k)|^2\right]$$

i). Assume a signal contains two frequencies, and the amplitude of one frequency is smaller than that of the other. By measurement, it is find the SPL of the smaller frequency is 7 dB less than the SPL of the other one. Find out the magnitude difference of the two frequencies.

ii). Using the FFT outputs obtained in Q1) to calculate the SPL values assuming a_1=0 and a_2=1.0 and 0.6 in the given program for N=128, 256 and 512, respectively.

iii). Comments on your results in terms of SPLs

Solution:

Assuming the magnitude of one frequency is A and the other one is kA, where k is smaller than 1.

1). From (1), we have

$$[96\ \text{dB} + 10\log_{10}(A^2)] - [96 + 10\log_{10}(k^2A^2)] > 7\ \text{dB}$$

$$10\log_{10}(A^2)] - 10\log_{10}(k^2A^2) = 10\log_{10}(1/k^2) > 7\ \text{dB}$$

$$\text{or } k^2 = 10^{-0.7} \quad \rightarrow \quad k = 0.4467$$

Therefore, one frequency is smaller by 0.4467 times of the other.

ii). The equation

$$SPL_{DFT} = 96dB + 10\log_{10}\left[\frac{4}{N^2 < w >^2}\sum_{peaks} |X(k)|^2\right]$$

The sum in the second term is to compute all the signal energy in the frequency domain. The table shows the SPLs for different N values

N=	128	256	512
A=1	96	96	96
A=0.6	91.58	91.58	91.58

The same results are also obtained from the equation of

$$SPL_{DFT} = 96dB + 10\log_{10}(A^2)$$

iii). From these results, the following observations are made

- The SPL values are independent of the values of *N*. This is the most important property of this formula.
- Although the frequency components may not be accurately located when N is small, the SPL values do not change since the signal energy is not lost, but may be distributed to other frequencies.

Question 8

Consider an input bitstream, Y = [2 5 6 6 2 5 5 4 1 4 4], {N = 11} chosen over a data set of V = [0 1 2 3 4 5 6 7]. The probabilities, p_i = [0, 1/11, 2/11, 0, 3/11, 3/11, 2/11, 0].

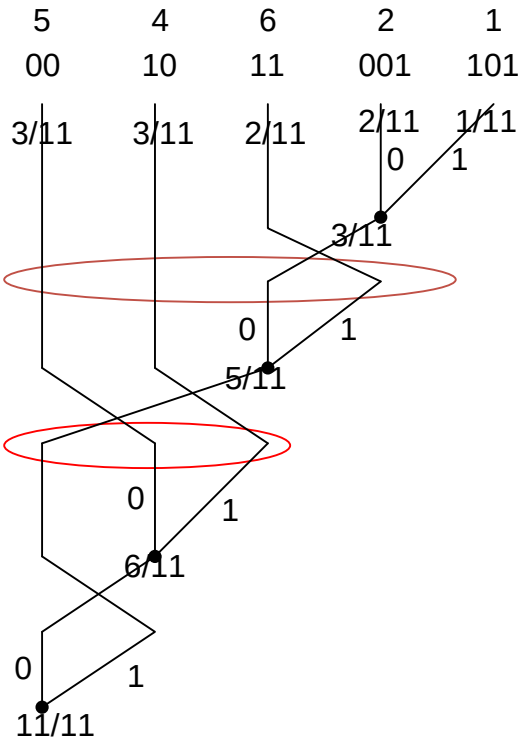
(a) Draw the Huffman code tree and determine the Huffman codeword.

(b) Compute the total length of the Huffman encoded bitstream.

(c) Compute the entropy of the input bitstream

Solution

(a) It is noted that the probability for V = [0 3 7] is 0. Therefore, we consider only Y'=[1 2 4 5 6]



Steps:

Reorder according to the probability after each level of merging

(a). The symbols are ordered according the probability decrement. If more than one symbol have the same the probability, they can be arranged in any order. Such an order should be maintained in the entire coding process.

(b). Always assign 1 (or 0) in the same side of the input branch of a node

(c). Obtain the code bit always from the top to the root. For example for symbol 1, we obtain the code word (101) from the branches at the top level to the bottom level. (In fact, it is also possible to get the code word by reading 0 or 1 from the root to the top. The coding efficiency will not be changed).

The code words for these symbols are listed below

Input symbol	Prob.	Huffman code
4	3/11	10
5	3/11	00
6	2/11	11

2	2/11	001
1	1/11	101

Based on the table above, the given input sequence requires 3+2+2+2+3+2+2+2+3+2+2 = **25 bits.**

For the set of above code words, the entropy is defined as

$$Entropy = \sum_{n}^{codes} p_n \log_2(1/\ p_n)$$

Therefore, we have E =-{2 x 3/11*$\log_2$(3/11) + 2 x 2/11* $\log_2$(2/11)+1 x 1/11* $\log_2$(1/11)} = **2.231 bit/symbol.**
Note that $\log_2(x) = lg_{10}(x)/\log_{10}(2)$

The average code word length is defined as $L = \sum_{n} p_n l_n$

For this case, we have
L = 2x2x 3/11+1x2x2/11 + 1x3x2/11 + 1x3x1/11= 2.2727 bits

The coding efficiency E/L = 2.231/2.27 =0.98 or 98%

Question 9
Let us use the filter $h_0(n) = [1/2^{1/2}, 1/2^{1/2}, 0,0]$ and the corresponding z transform of this filter is $H_0(z) = (1+ z^{-1})/2^{1/2}$.

i). Design a two channel filter bank that meet the perfect reconstruction condition.

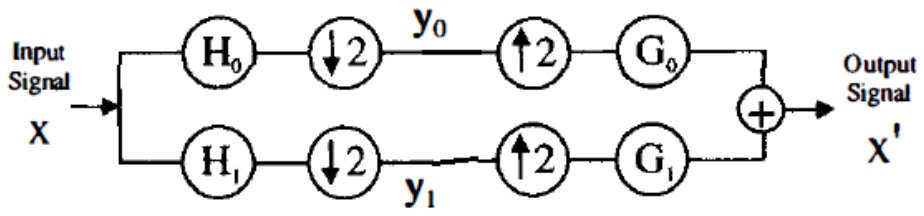
ii) Verify the perfect reconstruction based on the operations performed in the time domain by the two channel filter analysis and synthesis banks.

Solution:

i). Based on the relationships among the filter responses for analysis and synthesis filter banks, we have

$H_0(z) = (1+ z^{-1})/2^{1/2}$	$h_0(n) = [1/2^{1/2}, 1/2^{1/2}, 0,0]$
$H_1(z) = -(1- z^{-1})/2^{1/2}$	$h_1(n) = [-1/2^{1/2}, 1/2^{1/2}, 0,0]$
$G_0(z) = (1+ z^{-1})/2^{1/2}$	$g_0(n) = [1/2^{1/2}, 1/2^{1/2}, 0,0]$

$G_1(z) = (1- z^{-1})/2^{1/2}$ $g_1(n) = [1/2^{1/2}, -1/2^{1/2}, 0,0]$



Two-channel perfect reconstruction filter bank

ii). Assuming the input signal x(n) = { . . . , 0, 0, x(0), x(1), x(2), . . . }, we can easily verify

$y_0(n)$={ . . . ,0, $[x(0)+x(1)]/2^{1/2}$, ~~$[x(1)+x(2)]/2^{1/2}$~~ , $[x(2)+x(3)]/2^{1/2}$, ~~$[x(3)+x(4)]/2^{1/2}$~~ ...}

$y_1(n)$={ . . . ,0, $[x(0)-x(1)]/2^{1/2}$, ~~$[x(1)-x(2)]/2^{1/2}$~~, $[x(2)-x(3)]/2^{1/2}$, ~~$[x(3)-x(4)]/2^{1/2}$~~, ...}

The crossed terms are those being ignored by the down-sampling operation.

After up-sampling and filtering with the synthesis filters, these two series become

{... 0, [x(0)+x(1)]/2, [x(0)+x(1)]/2, [x(2)+x(3)]/2, [x(2)+x(3)]/2, ...}

{... 0, [x(0)-x(1)]/2, -[x(0)-x(1)]/2, [x(2)-x(3)]/2, -[x(2)-x(3)]/2, ...}

Finally, when added together to get the output signal, we find that

x'(n) = { 0, …, 0, x(0), x(1), x(2), x(3), …}

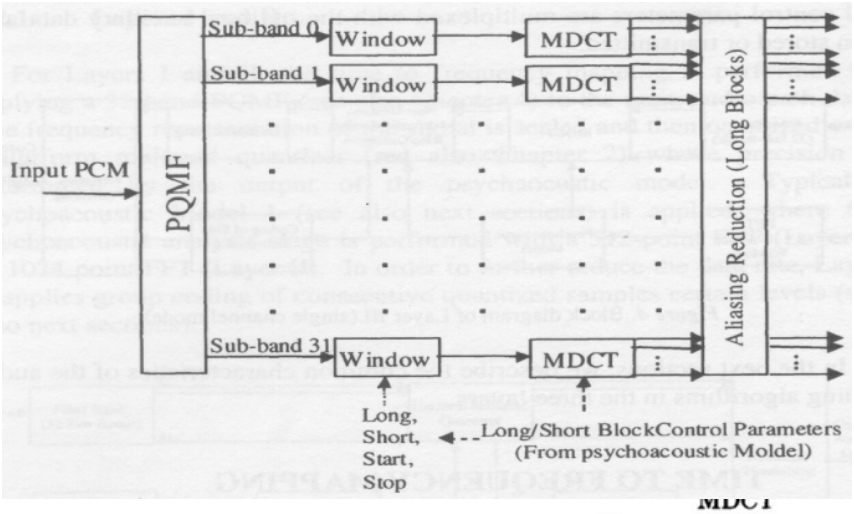
Questions of EE6424 for Weeks 5 and 6

Semester 2 2018

1) In MPEG layer III, the filter bank is a hybrid consisting of the 32-channel PQMF and followed by a time-variant MDCT. The MDCT filter bank consists of 18-frequency lines MDCT for steady state signal or 6-frequency lines MDCT for transient-like signals. If the sampling frequency of 48 kHz is used,
   (a) What are the frequency and time resolution?
   (b) How these results are compared to Mpeg layer I and II?

Solution

See the figure below for the analysis in the frequency domain.



			MDCT
Frequency resolution at 48 kHz	750 Hz	750 Hz	41.66 Hz
Time resolution at 48 kHz	0.66 ms	0.66 ms	4 ms
Impulse response (LW)	512	512	1664
Impulse response (SW)	—	—	896
Frame length at 48 kHz	8 ms	24 ms	24 ms

The block diagram of MPEG Audio Layer –III

- In the figure, a filter bank divides the input signal into 32 subbands. Good understanding on the process used for this spectrum analysis is needed. The information can be found in the lecture notes. Below are the points relevant to this question.

- The input signal is segmented by a window of 512 points (see lecture note). The input signal is firstly divided into 32 subband by the filter bank.
- For each subband, 36 samples (long window) or 12 samples (short window) are transformed by the MDCT. Due to 50% of window overlapping for eliminating the windowing effects, the MDCT produces 18 or 6 output samples in the frequency domain.

(a) The freq. resolution of the MPEG layer III is computed when the long window is used. The first stage of the freq. mapping generates 32 subbands and the MDCT for each subband has 18 freq. lines/subband. Therefore, the number of freq. components over the range of Fs/2 is 18 x 32 = 576. Then the best freq. resolution = (Fs/2)/(no. of freq. components)), i.e., = (48 k/2)/576= **41.66 Hz**.

The time resolution is based on the shortest window used in layer III. Therefore, best time resolution = (6*32)/48k= **4 msec**.

By using the window switching, these resolutions are used for time-to-frequency mapping.

(b) In the case of Layers I and II, only filter bank (or PQMF) is used to divide the input into 32 subbands (see the table in Page 1) and the frequency resolution based on the 32 subbands. The signal bandwidth is between 0 and 24 kHz. Therefore the frequency resolution = 24k/32 = 750Hz.

In the case of layers I and II, the minimum number of inputs for PQMF is 32 so that PQMF produces at least one output for each subband. Therefore,

Time resolution = 32/48kHz = 0.66msec.

The impulse response of the PQMF filter used in layer III is 512/48kHz = 10.67 msec.

It should be noted that switching between long and short window are not used in MPEG layer I and II.

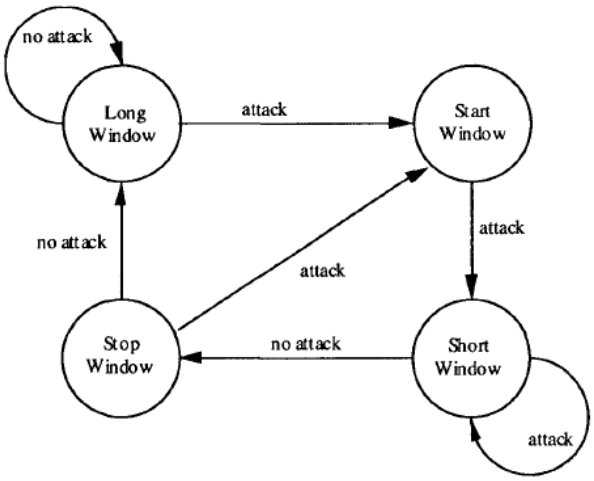
2) Audio coder usually uses a 50% overlapping transform, namely the modified discrete cosine transform (MDCT) to blend one frame into the next frame without boundary artefacts. Two sine windows of lengths 64 and 16 are used for long window and short window, respectively.

   a) Draw the graphs to show (describe) the switching from long to short windows and switching from short to long windows, and state the duration of overlap in the diagram.

   b) What are the frequency and time resolution of the long and short window when a sampling frequency of 48 kHz is used?

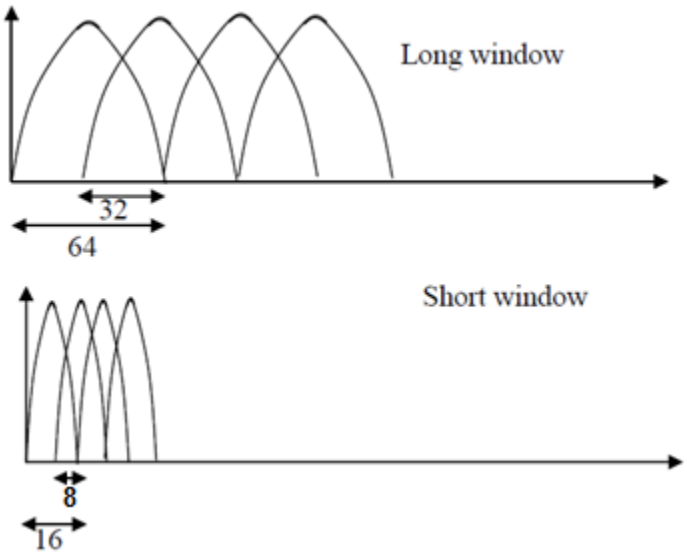
Solutions:

a). A long window is 64 samples and a short window is 16 samples. The state transitional diagram is

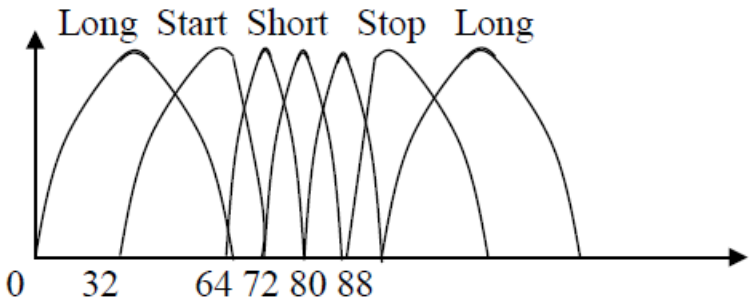


Block switching state diagram for Layer III from [ISO/IEC 11172-3]

For 50% window overlapping:



When switching from long-to-short window or from short-to-long window, transition windows are required as shown in the figure below.
The transition windows have different slopes that overlap with long and short windows.



When the system receives an 'attack' indication requiring short window, the start window (or transition window) is used and short window follows.

Note that the left half of the start window is the same as the left half of long window and the right half of start window is the same as the right half of short window. This is necessary for meeting the requirements of perfect reconstruction during switching.

When the system does not receive the 'attack' indication, it means that long window is needed, the switching from short-to-long window is performed similarly by using short window, transition window (stop window) and long window.

b). Frequency resolution of long window is:

$$(Fs/2)/(64/2) = 48k/64 = \mathbf{750\ Hz}$$

where Fs/2 is the frequency range of the MDCT and 64/2 is the number of output points of the MDCT with 50% of overlapping.

Frequency resolution of short window is:

$$(Fs/2)/(32/6) = 48k/16 = \mathbf{3{,}000\ Hz}$$

Time resolution of long window (because of 50% overlap, we take only 32 new samples) is:

$$32/48k = \mathbf{0.66\ msec}$$

Similarly, the time resolution of short window (because of 50% overlap, we take only 16 new samples)

$$8/48k = \mathbf{0.167\ msec}$$

3). The tonal maskers are found by the following definitions

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

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

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

Discuss the number of frequency components to be used for maskers at different frequency. Explain why?

Solution:

This question is to enhance the understanding that how the spectrum of the signal (from FFT) is used for calculating the masking effects. The equation

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

defines the process how to find the master from the signal spectrum. The equation

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

defines the number of frequency components to be considered for identifying a masker. In particular, the number of neighbouring frequencies to be considered is

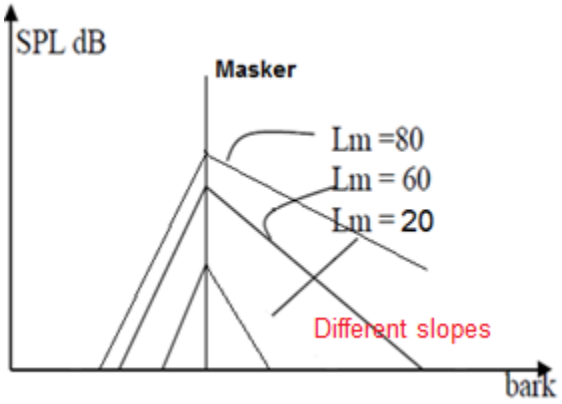
- 2 on each size when the masker frequency is in the range of 0.17- 5.5 k Hz;
- 3 on each side when the masker frequency is in the range of 5.5 – 11.0 k Hz, and
- 6 on each side when the masker frequency is larger than 11.0 k Hz.

In general, the frequencies > 11.0 k Hz have small amplitudes with less importance.

4. Explain why a low-frequency tone is able to better mask a tone of higher frequency than vice versa? Your explanation must also include the 2-slope spreading function to illustrate the effect of bark scale difference between the masker and maskee frequency on the masking curve.

Solution

This question is to demonstrate that the masking effects of the spreading function on different frequencies.



Consider a simple masking curve characterized by a 2-slope spreading function. The slope is steeper at the low frequency compared to the high frequency side as shown in the figure.

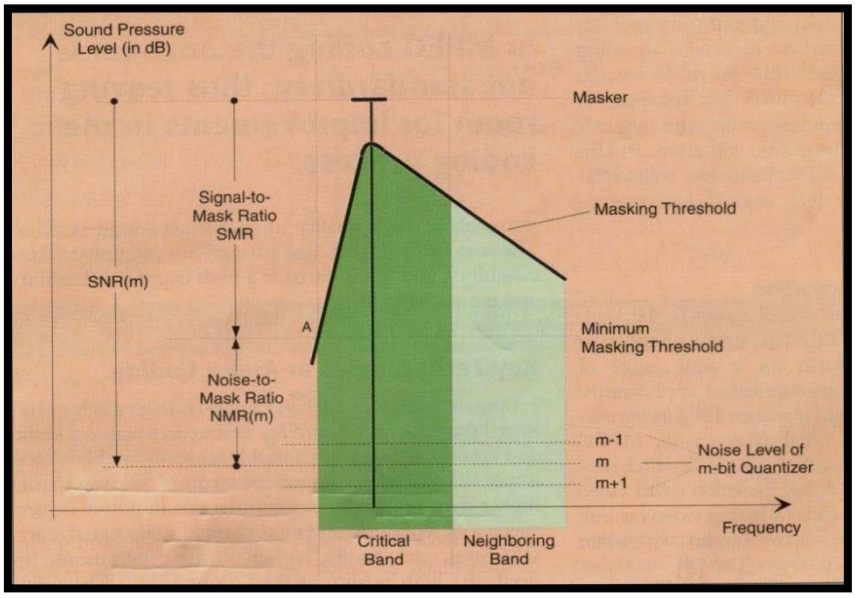
- These masking curves implies that low-frequency tone (as a masker) can better mask a higher frequency tone (as a maskee).
- The masking curves also tend to vary with different SPL level of the masker. The shape of the spreading function becomes more obviously asymmetric because the slopes on the high frequency side become smaller as SPL increases (see the figure).
- In general, the spreading function can be approximated by a triangular function (see lecture note). A constant slope of -27 dB/bark is used for the low-frequency slope. The high-frequency slope is decreased as the increase of SPL values.

This observation tells us as SPL increases, the masking for the frequencies higher than the masker frequency is increased compared to that for those frequencies lower than the masker frequency.

4) A minimum Signal-to-Mask Ratio (SMR) of 2 dB is observed for a noise masker with a centre freq at 250 Hz, 3 dB for 1kHz masker, and 5 dB for 4 kHz masker.
   i). Which is a better noise masker? And why?
   ii). The minimum SMR for a tone-masking-tone experiment is found to be 15 dB. How is this value compared to that of the noise-masking-tone experiment?
   iii). Can a tone mask narrowband noise?
   iv). What happens if the masking signal increases its SPL from 60 to 80 or 100 dB? Is there any change in the minimum masking threshold?

Solution:

To answer the above questions, it is better to understand the Figure.



SNR: signal-to-noise ratio means the comparison between the energies (or power) from the signal and the noise, which is generally the quality measure of the useful signal for applications.

SMR: signal to mask ratio define the difference between signal energy and the maximum level over which the noise is audible. This value is used to determine the additional number of bits that are used to determine the number of bits for the masker.

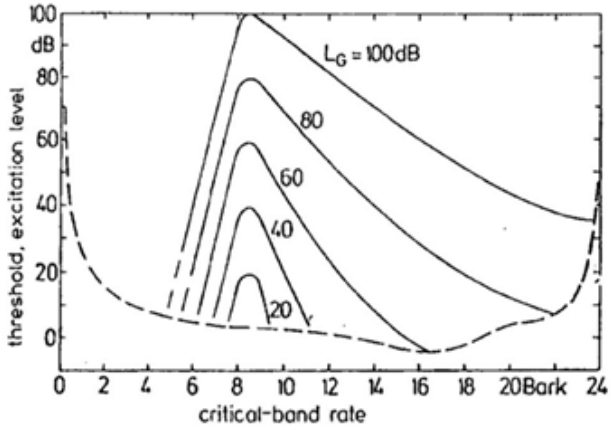
MNR=SNR-SMR: the noise-to-masker ratio measures the difference between the minimum masking threshold and the noise level. It cannot be negative because the noise is to be audible. Also see the minimum SMR in the figure.

CB: critical band defines the bandwidth in which our hearing system is not able to tell signals at different frequencies if they have the same magnitude.

(i). In freq. masking, the higher the SMR, the lesser the masker will mask the maskee. Therefore, the best masker among the three is centred at 250 Hz, which allows more tonal signal to be masked.

(ii). The min. SMR for a tone-masking-tone experiment is around 15 dB. For a noise-masking-tone experiment, the min. SMR is lesser than 10 dB which implies that noise is a better masker than tone.

(iii) A tone can mask a narrowband noise if the tone is strong enough.



(iv). If the masking signal increases its SPL from 60 to 80 to 100 dB, the slope towards the higher freq. becomes more gradual as masking increases. This implies that the masking curves cover a larger freq. region and mask more higher freq. maskees.

The minimum masking threshold will change as the SPL of the masker increases. This can be seen from the figure in the previous page that the masking threshold will be changed upward as the SPL is increased.