Chapter 6

Introduction to Psychoacoustics

1. INTRODUCTION

In the introduction to this book, we saw that the last stage in the coding chain is the human ear. A good understanding of how the human ear works can be a powerful tool in the design of audio codecs. The general idea is that quantization noise can be placed in areas of the signal spectrum where it least affects the fidelity of the signal, so that the data rate can be reduced without introducing audible distortion.

In this chapter, we examine the main aspects of psychoacoustics (the science that studies the statistical relationships between acoustical stimuli and hearing sensations) that are useful in the design of perceptual audio coders. The main goal of this chapter is to introduce the basic principles and data behind the masking models currently utilized in state-of-the-art audio coders. First, units for sound pressure level measurements and the range of human hearing are introduced. The hearing threshold and masking phenomena are discussed and their main empirical properties presented. We then examine the underlying mechanism of the hearing process and how the ear acts as a spectrum analyzer, analyzing sound in specific frequency units called critical bands. This will provide us with the foundation for developing psychoacoustic models, which link empirical masking data with the sound hearing sensation.

2. SOUND PRESSURE LEVELS

As we saw in Chapter 1, sound can be represented as a function of time. Sound reaches the human ear in the form of a pressure wave. It can be represented as the variation of the air pressure in time, p(t), where the pressure is defined as force per unit area. The unit of pressure in the MKS system is the Pascal (Pa) where 1 Pa = 1 N/m^2. Relevant values of sound pressure for audio applications vary between 10^{-5} Pa, which is close to the limits of human hearing at the most sensitive frequencies, and 10^2 Pa, which corresponds to the threshold of pain.

To describe such a wide range of relevant sound pressures, we usually choose to work in logarithmic units and define the sound pressure level, SPL, in units of dB as

$$SPL = 10 \log_{10} (p/p_o)^2$$

where $p_o = 20$ μ Pa is roughly equal to the sound pressure at the hearing threshold for tone frequencies around 2 kHz [Zwicker and Fastl 90].

We often also describe sounds in terms of the sound intensity. The sound intensity, I, is the power per unit area in the sound wave and it is proportional to p^2. The SPL (in units of dB) can also be calculated in terms of sound intensity as:

$$SPL = 10 \log_{10} (I/I_o)$$

The intensity I is measured in MKS units in terms of W/m^2 (1W = 1N m/s) and the reference sound intensity $I_o = 10^{-12}$ W/m^2 corresponds to a wave with the reference pressure p_o.

3. LOUDNESS

The hearing sensation that corresponds to sound levels is the loudness of the sound. The concept of loudness was first introduced by Barkhausen in the 1920s as a means to describe perceived sound intensities. The loudness level is defined as the level of a 1 kHz sound tone that is perceived as loud as the sound under examination for frontally incident plane fields. In general, the loudness of an audio signal depends on its duration and its temporal and spectral structure in addition to its intensity. The loudness unit is the phon, where the phon describes a curve of equal loudness as a function of frequency. It is interesting to note that the difference between values for the loudness measured in phones and values for the intensity measures in dB

decreases at high levels (see *Figure 1*). For example a 1 kHz tone at 100 dB is perceived almost as loud as a 100 Hz tone at 100 dB, while 1 kHz tone at 40 dB is perceived as about 20 dB louder than a 100 Hz tone at 40 dB [Fletcher and Munson 33]. It should be noted that, depending on how the equal loudness contours are measured, there might be differences in the data. Some of these differences can be accounted for by considering an attenuation factor necessary to produce equal loudness from frontally incident plane fields versus diffused sound fields [Zwicker and Fastl 90].

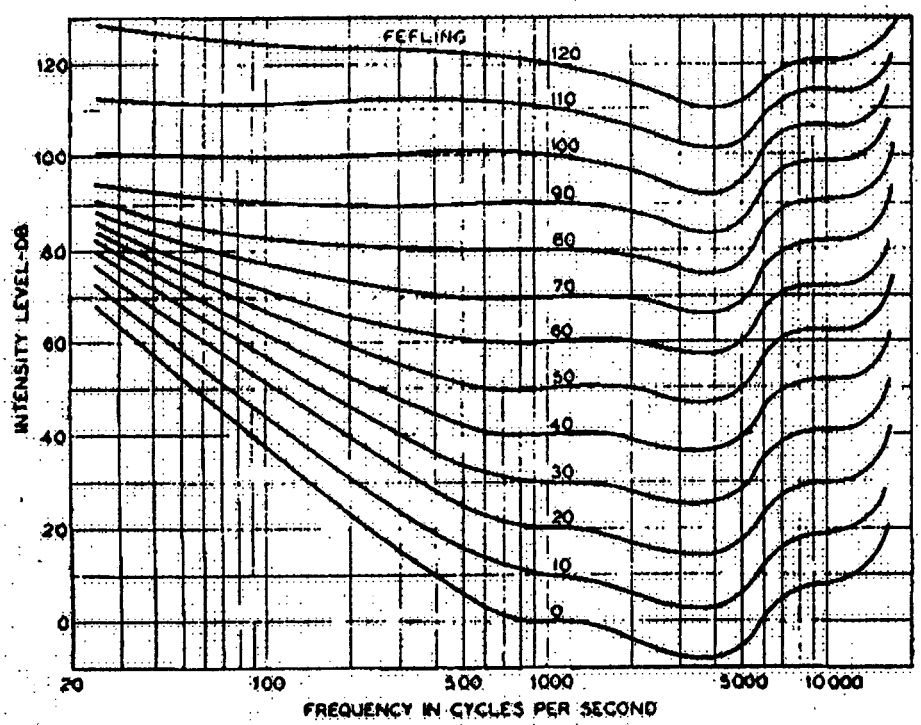

Figure 1. Loudness contours from [Fletcher and Munson 33].

4. HEARING RANGE

The human ear can cover a wide range of SPLs. *Figure 2* shows the hearing area of a typical human ear [Zwicker and Fastl 90]. The graph illustrates different SPL curves as function of frequency. The frequencies shown in the abscissa vary between 20 Hz and 20 kHz, which is generally considered the frequency range of audible sounds. It should be noted, however, that recent findings imply that particularly sensitive subjects can hear sounds at frequencies above 20 kHz.

The curve in the lower part of the graph represents the threshold in quiet, which is the level of audibility for pure tones in steady state conditions. The dotted line extending upwards from the threshold in quiet between 2 and 20 kHz represents the hearing loss commonly seen in subjects exposed to loud sounds in the mid-range frequency region. The threshold of pain is the dashed line at the top of the diagram. The area between the threshold in quiet and the threshold of pain represents the human hearing range.

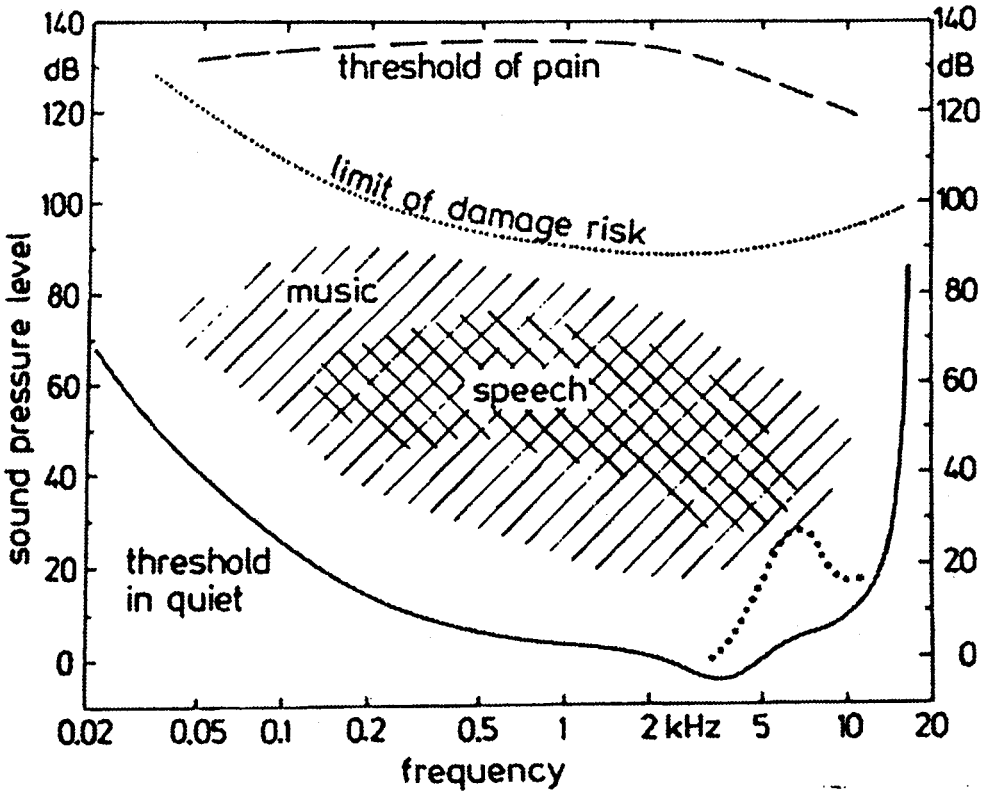

Figure 2. Hearing area from [Zwicker and Fastl 90]

Human speech typically falls into the frequency range comprised between 100 Hz and 8 kHz, and has SPLs ranging from about 30 dB up to around 70 dB, with typical conversation levels at values at about 50-60 dB. Music typically has a wider range in both frequency and SPLs than speech. For example, the A_0 note in the piano is tuned at 27.5 Hz (C_0 is tuned at about 16 Hz) while the highest note of the piccolo is at a frequency of about 8.4 kHz. Moreover, harmonics of musical instruments such as the violin and cymbals can reach frequencies above 15 kHz. The dynamic range for music typically varies between 20 dB and 95 dB. Around 100 dB is the onset of risk for hearing damage. At about 120 dB is the threshold of pain.

5. HEARING THRESHOLD

The hearing threshold, or threshold in quiet, represents the lowest sound level that can be heard at a given frequency. Even in extremely quiet conditions, the human ear cannot detect sounds at SPLs below the threshold in quiet. This curve is extremely important for audio coding since frequency components in a signal that fall below this level are irrelevant to our perception of sound and therefore they do not need to be transmitted. In addition, as long as the quantization noise in frequency components that are transmitted falls below this level, it will not be detectable by the human hearing process.

The threshold in quiet is also important in describing how loud we perceive sounds to be. In particular, the equal loudness contours display a shape that is nearly parallel to the threshold in quiet for low loudness levels (20 phones or below) suggesting that it is the difference between a sound and the threshold in quiet that determines the loudness for soft sounds. For loud sounds, the SPL itself plays a more important role in the determination of loudness. According to [Zwicker and Fast 90] the threshold in quiet corresponds to the equal loudness contour described by phon = 3.

The threshold in quiet can be measured by recording the sound pressure level of the lowest sound level that elicits a listener's response that the sound is audible. The frequency dependence can be tracked by giving the test subject a switch which changes between continuously incrementing and continuously decrementing the sound pressure level of a test tone whose frequency is slowly sweeping from low to high values and vice-versa. The test subject is instructed to switch to decrementing the sound pressure level when the sound is definitely audible and to switch to incrementing the pressure level when the sound is definitely inaudible. Typically, the results produce zigzag curves such as that in *Figure 3* from [Zwicker and Fastl 90] with a range of roughly 6 dB between the point where the sound is definitely audible and where it is definitely inaudible. The average of the two curves marking the top and bottom of the zigzags is used as the assessment of the threshold in quiet. According to [Zwicker and Fastl 90], the reproducibility of the threshold in quiet for a single subject is within ±3 dB. In addition, the frequency dependence of this curve has been recorded in a similar manner for many subjects with normal hearing.

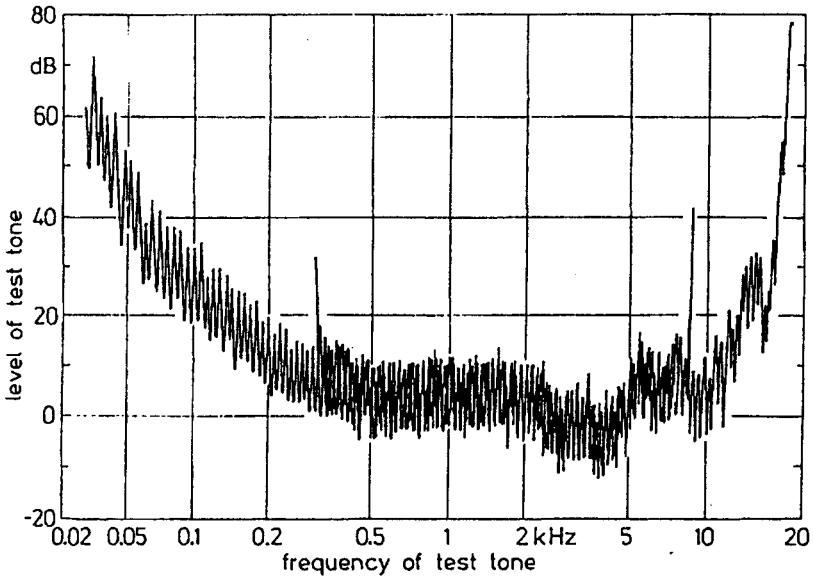

Figure 3. Sample results from an experimental assessment of the threshold in quiet from [Zwicker and Fastl 90]

The frequency dependence of the hearing threshold has been fairly well established. The threshold is relatively high at low frequencies. It is at an SPL of around 40 dB at 50 Hz and almost drops to 0 dB by 500 Hz. It remains almost constant near 0 dB between 500 and 2 kHz. It can then drop below zero between 2 kHz and 5 kHz for listeners with good hearing. For frequency above 5 kHz, there are peaks and valleys that vary from subject to subject but the threshold is generally rising. Typically, the threshold then increases quite rapidly above 16 kHz. While for frequencies below 2 kHz the threshold seems to be largely independent of age, above 2 kHz it is shifted to a value almost 30 dB higher at 10 kHz for 60-year old subjects than for 20 year old subjects. *Figure 4* shows a comparison plot of the threshold in quiet for test subjects of various ages [Zwicker and Fastl 90].

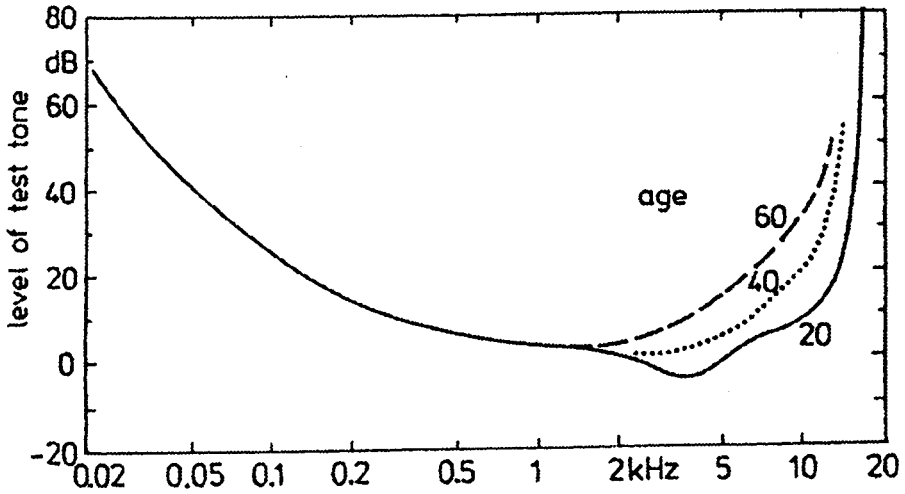

Figure 4. Threshold in quiet for listeners of different ages from [Zwicker and Fastl 90]

As shown in [Terhardt 79], one can obtain a good approximation of the threshold in quiet by utilizing the following frequency dependent function:

$$A(f)/\mathrm{dB} = 3.64(f/\mathrm{kHz})^{-0.8} - 6.5e^{-0.6(f/\mathrm{kHz}-3.3)^2} + 10^{-3}(f/\mathrm{kHz})^4$$

where the threshold in quiet is modeled by taking into consideration the transfer function of the outer and middle ear and the effect of the neural suppression of internal noise in the inner ear (see also Section 9 later in this chapter). A graph of the frequency dependence of this function can be seen in *Figure 5.* Notice how it reasonably mimics the behavior of the experimentally derived curves shown in the prior figures.

One should be aware that, in order to be able to compare a signal with the threshold in quiet, it is important to know the exact playback level of the audio signal. In general, the playback level is not known a priori in the design of a perceptual audio coder. A common assumption is to consider the playback level as such that the smallest possible signal represented in the audio coding system under design will be presented close to 0 dB. This is equivalent to aligning the fairly flat bottom of the threshold in quiet, corresponding to frequencies of roughly 500 Hz to 2 kHz, with the energy level represented by the least significant bit of the spectral signal amplitude in the system under design.

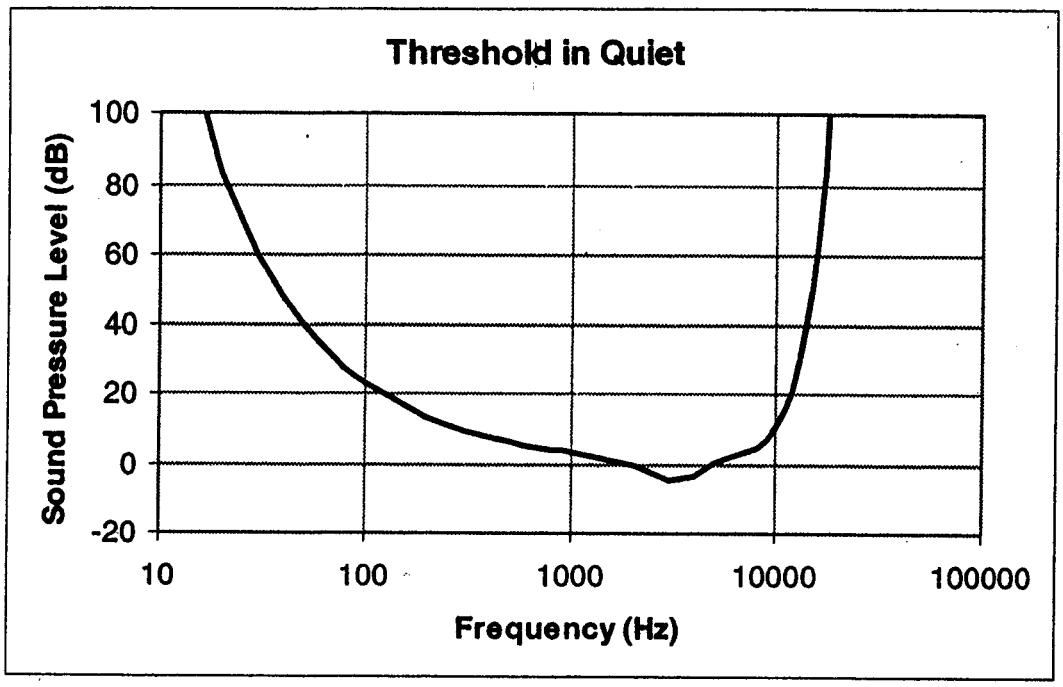

Figure 5. Approximate formula for the threshold in quiet

6. THE MASKING PHENOMENON

Masking of soft sounds by louder ones is part of our everyday experience. For example, if we are engaged in a conversation while walking on the street, we typically cease conversation while a loud truck passes since we are not be able to hear speech over the truck noise. This can be seen as an example of masking: when the louder masking sound (the truck) occurs at the same time as the maskee sound (the conversation), it is no longer possible to hear the normally audible maskee. This phenomenon is called simultaneous or frequency masking. Another example of frequency masking occurs when in a performance one loud instrument (masker) masks a softer one (maskee) that is producing sounds close in frequency. In general simultaneous masking phenomena can be explained by the fact that a masker creates an excitation in the cochlea's basilar membrane (see also next sections) that prevents the detection of a weaker sound exciting the basilar membrane in the same area.

Masking can also take place when the masker and the maskee sounds are not presented simultaneously. It this case we refer to this phenomenon as temporal masking. For example, in speech a loud vowel preceding a plosive consonant tends to mask the consonant. Temporal masking is the dominant effect for sounds that present transients, while frequency masking is dominant in steady state conditions. For example, in coding sharp

instrument attacks like those of castanets, glockenspiel, temporal masking plays a more important role than frequency masking.

6.1 Frequency Masking

Figure 6 illustrates frequency masking. In this figure, we see a loud signal masking two other signals at nearby frequencies. In addition to the curve showing the threshold in quiet, the figure shows a curve marked "masking threshold"[2] that represents the audibility threshold for signals in the presence of the masking signal. Other signals or frequency components that are below this curve will not be heard when the masker is present. In the example shown in *Figure 6*, the two other signals fall below the masking threshold, so they are not heard even though they are both well above the threshold in quiet. Just like with the threshold in quiet, we can exploit the masking thresholds in coding to identify signal components that do not need to be transmitted and to determine how much inaudible quantization noise is allowed for signal components that are transmitted.

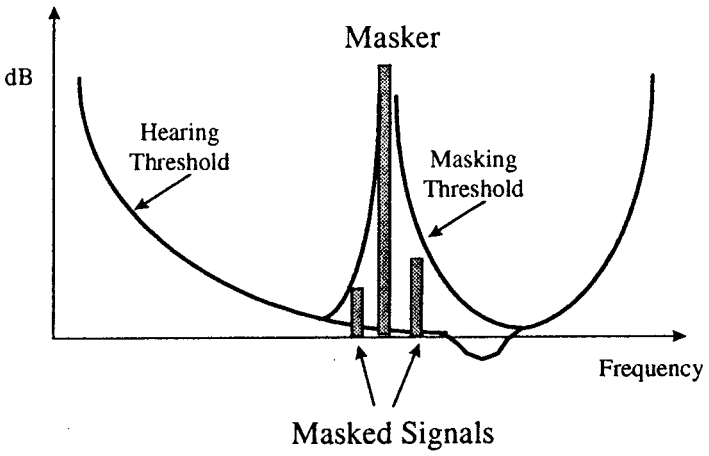

Figure 6. Example of frequency masking

6.2 Temporal Masking

In addition to simultaneous masking, masking phenomena can extend in time outside the period when the masker is present. Masking can occur prior to and after the presence of the masker. Accordingly, two types of temporal

[2] We shall refer to "masking thresholds" or "masking curves" to indicate the elevation of the hearing threshold due to the presence of one or more masker sounds. We define the "masked threshold" or "masked curve" as the combination of the hearing threshold and the masking threshold.

masking are generally encountered: pre-masking and post-masking. Pre-masking takes place before the onset of the masker; post-masking takes place after the masker is removed. Pre-masking is somewhat an unexpected phenomenon since it takes place before the masker is switched on. In general, temporal masking can be explained if we consider the fact that the auditory system requires a certain integration time to build the perception of sound and by the fact that louder sounds require longer integration intervals than softer ones.

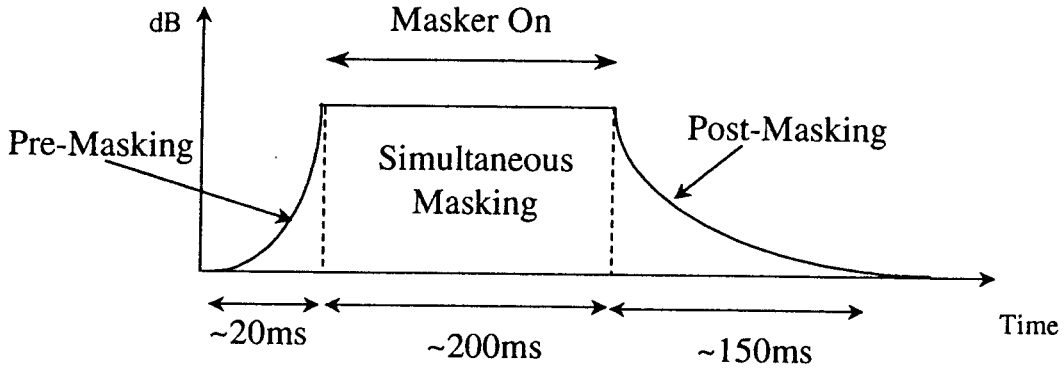

Figure 7. Example of temporal masking

In *Figure 7*, an example of temporal masking is shown [Zwicker and Fastl 90]. A 200 ms masker masks a short tone burst with very small duration relative to the masker. In the figure pre-masking lasts about 20 ms, but it is most effective only in the few milliseconds preceding the onset of the masker. There is no conclusive experimental data that link the duration of pre-masking effects with the duration of the masker. Although pre-masking is a less dramatic effect than post or simultaneous masking, it is nevertheless an important issue in the design of perceptual audio codecs since it is related to the audibility of "pre-noise" or "pre-echo" effects caused by encoding blocks of input samples. Pre-noise or pre-echo distortion occurs when the energy of the coded signal is spread in time prior to the onset of the attack. This effect is taken into consideration in the design of several perceptual audio coding systems both in terms of psychoacoustics models and analysis/synthesis signal adaptive filter design.

Figure 8 from [Bosi and Davidson 92] shows an example of a castanet signal (*Figure 8* (a)) in which encoding with a fixed block length led to a spread of energy in the 5 ms prior the onset of the transient (*Figure 8* (b)). This effect is perceived as a distortion sometimes described as a "double attack" and it is known in literature as pre-echo. Although some pre-masking effects can last on the order of tens of milliseconds, pre-masking is most effective only a few milliseconds. It should also be noted that pre-masking is less effective with trained subjects. In order to correct for pre-

echo distortion, adaptive filter banks (see Chapter 5) are often adopted in perceptual audio coding. *Figure 8* (c) shows the reduction in pre-echo distortion that resulted from using an adaptive filter bank to adjust the block length in the presence of the transient signal.

Post-masking is a better understood phenomenon. It reflects the gradual decrease of the masking level after the masker is switched off. Post-masking is a stronger effect than pre-masking and has a much longer duration. In *Figure 7* post-masking lasts about 150 ms. Post-masking depends on the masker level, duration, and relative frequency of masker and probe.

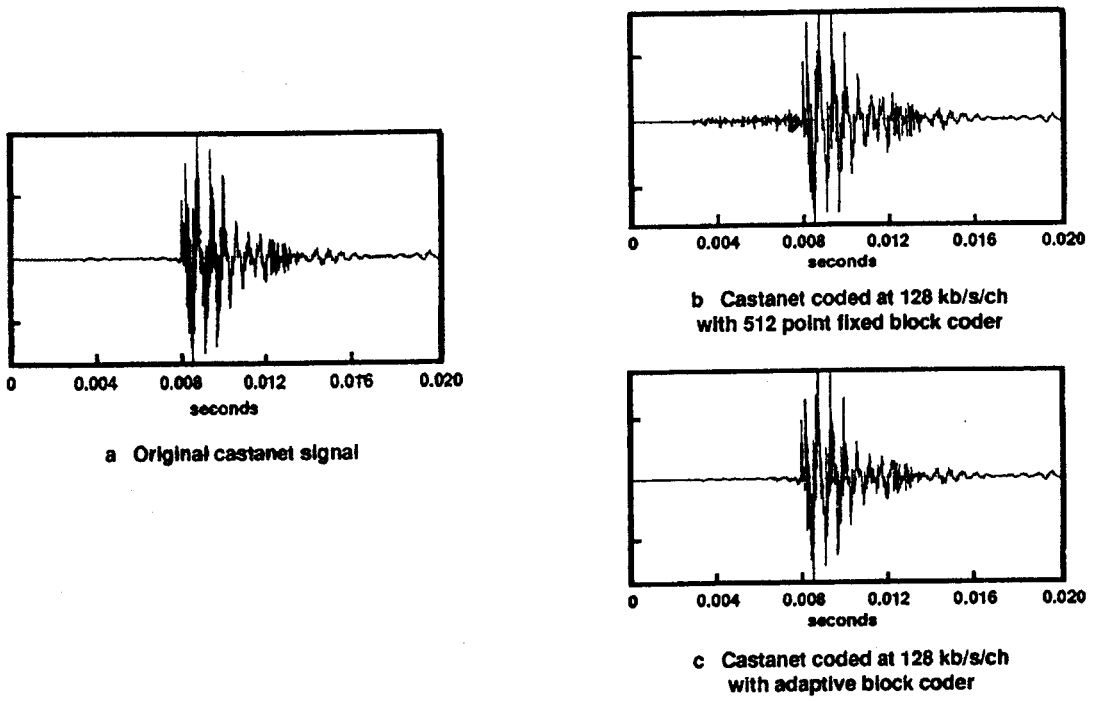

Figure 8. Example of pre-echo effects in a transient signal coded with a fixed (b) versus adaptive (c) resolution filter bank; the original signal is shown in (a). In (b) the amount of energy spread in time prior to the onset of the signal is perceived as pre-echo distortion and it is not temporally masked by the signal from [Bosi and Davidson 92]

An important question in the design of perceptual audio coders is how to account for masking effects. Masking curves are typically measured only for very simple maskers and maskees (either pure tones or narrow-band noise). In perceptual audio coding the assumption is that masking effects derived from simple maskers can be extended to a complex signal. Masking thresholds are computed by: a) identifying masking signals in the frequency domain representation of the data, b) developing frequency and temporal

masking curves based on the characteristics of each identified masker, and c) combining the individual masking curves with each other and with the threshold in quiet to create an overall threshold representing audibility for the signal. This overall audibility threshold or masked threshold is then used to identify inaudible signal components and to determine the number of bits needed to quantize audible signal components.

7. MEASURING MASKING CURVES

Masking curve data are collected by performing experiments on subjects that record what are the limits of audibility for a test signal (or probe) in the presence of a masking signal. The masking threshold varies dramatically depending on the nature and the characteristics of the masker and of the probe. Typically, for frequency masking measurements, the probe and the masker can be a sinusoidal tone or narrow band noise of extended duration. For temporal masking measurements, a short burst or sound impulse is used as a probe and the masker is of limited duration.

One way to measure a masking curve is to use a variant of the tracking method described for measuring the threshold in quiet. In this case, however, a masking signal will be played as the subject tries to identify the audibility limits for test signals. *Figure 9* shows an example of the masking curve that results from such an experiment. In this example, the masking signal is a pure tone a 1 kHz with an SPL of 60 dB. The lower zigzag line is the threshold in quiet for this test subject measured in the absence of the masking signal. The upper zigzag line is the audibility threshold when the masking signal is playing. Notice how masking in this case is strongest at frequencies near the masker's frequency and how it drops off quickly as the test signal moves away from the masker frequency in either direction – these features tend to be quite general results. Notice also that the highest masking level is roughly 15 dB below the masker level and that the drop-off rate is much quicker moving to low frequencies than moving to higher ones – these features tend to be very dependent on the specifics of both the masking signal and the test signal. In the following sub-sections, we summarize some of the main features of frequency masking curves as determined by similar experiments on test subjects.

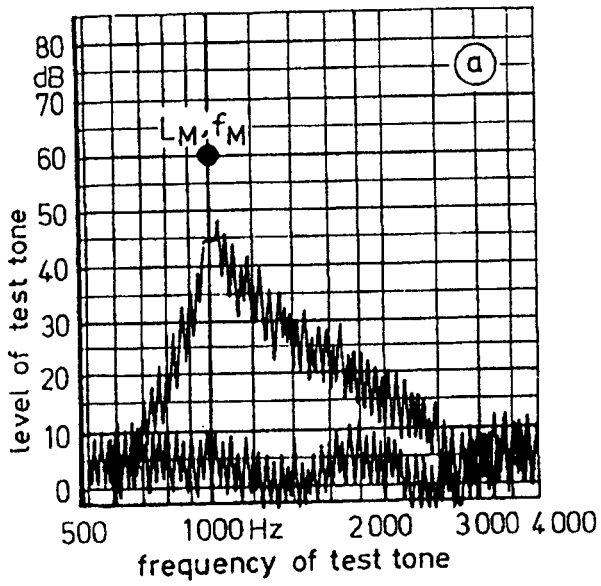

Figure 9. Sample results from experimental determination of a masking curve from [Zwicker and Fastl 90]

7.1 Narrow-Band Noise Masking Tones

In the case of narrow-band noise masking tones, the masker is noise with a bandwidth equal to or smaller than a critical band (see the definition of critical bandwidth in the next sections). *Figure 10* shows measured masking thresholds for tones masked by narrow-band noise centered at 250 Hz, 1 kHz, and 4 kHz [Zwicker and Fastl 90]. The noise bandwidths are 100 Hz, 160 Hz, and 700 Hz respectively. The slopes of the noise above and below the center frequency are very steep, dropping more than 200 dB per octave. The level of the masker is 60 dB, computed based on the noise intensity density and bandwidth. The horizontal dashed line shows the noise level in the figure. The solid lines in the figure show the levels of the pure tone probe in order to be just audible. The dashed curve at the bottom represents the threshold in quiet.

The masking threshold curves present different characteristics depending on the frequency of the masker. While the frequency dependence of the threshold masked by the 1 kHz and the 4 kHz narrow-band noise are similar, the 250 Hz threshold appear to be much broader. In general, masking thresholds are broader for low frequency maskers (when graphed, as is customary, using a logarithmic frequency scale). The masking thresholds reach a maximum near the masker center frequency. Their slopes can be very steep ascending from low frequencies (over 100 dB per octave), and present a somewhat gentler decrease after reaching the maximum. This steep rise creates the need for very good frequency resolution in the analysis of the

audio signals, otherwise errors will be made in the evaluation of masking effects.

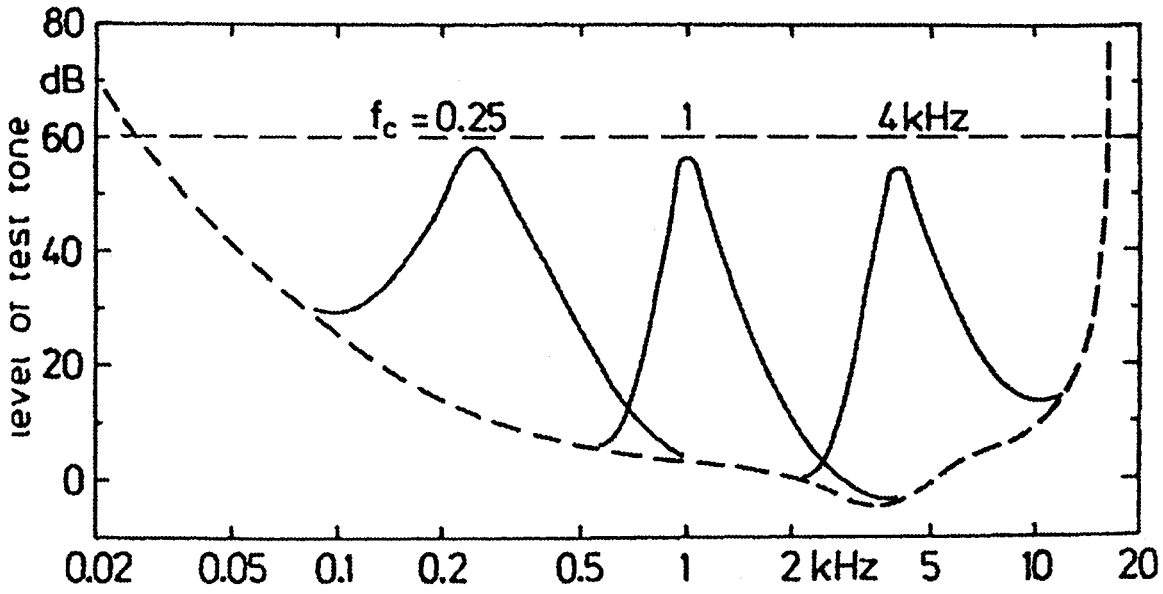

Figure 10. Masking thresholds for 60 dB narrow-band noise masking tones from [Zwicker and Fastl 90]

The difference in level between a signal component and the masking threshold at a certain frequency is sometimes referred to as the signal to mask ratio, SMR. Higher SMR levels indicate less masking. The minimum SMR between an masker and the masking curve it generates is a very important parameter in the design of audio coders. The minimum SMR values for a given masker tend to increase as the masker frequency increases. For example, in *Figure 10* we have a minimum SMR value of 2 dB for a noise masker with 250 Hz center frequency, 3 dB for the 1 kHz masker, and 5 dB for the 4 kHz masker.

In *Figure 11*, the masking threshold for narrow-band noise centered at 1 kHz is shown for different masker SPLs. The minimum SMR stays constant at around 3 dB for all levels. At frequencies lower than the masker, each of the measured masking curves has a very steep slope that seems to be independent of the masker SPL. In contrast, the slope in the masking curve towards higher frequencies shows noticeable sensitivity to the level of the masking signal. Notice that the slope appears to get shallower as the masking level is increased. In general, the frequency dependence of the masking curves is level sensitive, i.e. non-linear. The dips in *Figure 11* are caused by non-linear effects in the hearing system driven by the high level of the noise masker and the probe.

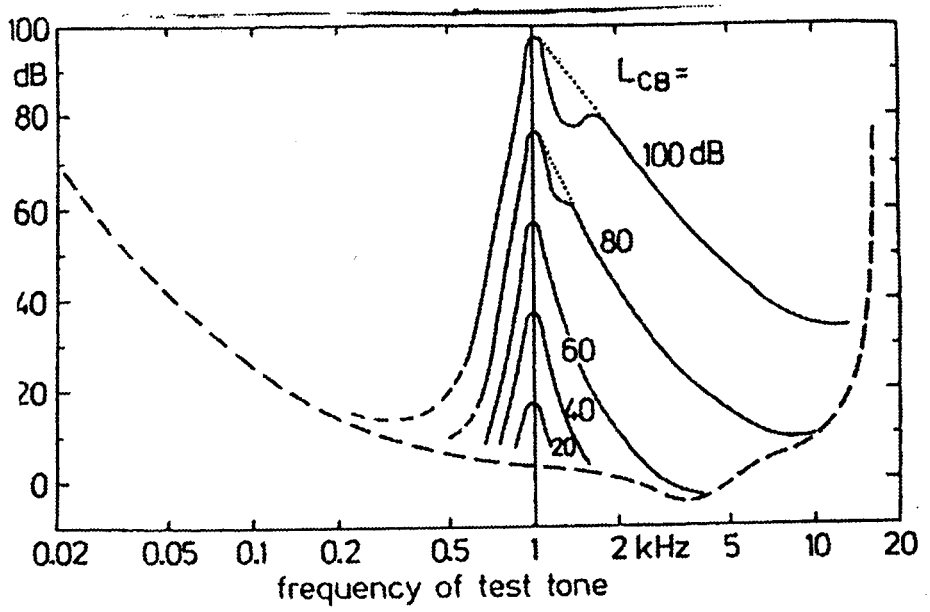

Figure 11. Masking thresholds for a 1 kHz narrow-band noise masker at different levels masking tones from [Zwicker and Fastl 90]

7.2 Tones Masking Tones

Although much of the early work on masking phenomena was based on measurements of pure tones masking pure tones, such masking experiments present greater difficulties than noise masking experiments due to the phenomenon of beating. In such experiments, the subjects also sometimes perceive additional tones besides the masker and probe. The most dominant effect, the beating effect, is localized in the neighborhood of the masker frequency and it depends on the masker level. *Figure 12* shows the results for a 1 kHz masker at different levels. In this particular experiment [Zwicker and Fastl 90] the probe was set 90 degrees out of phase with the masker when it reached the frequency of 1 kHz (equal to the masker frequency) to avoid beating in that area. It is interesting to notice that at low masking levels, there is a greater spreading of the masking curves towards lower frequencies than higher frequencies. The situation is reversed at high masking levels, where there is a greater spreading towards high frequencies than lower frequencies.

In general, the minimum masker SMRs are larger in experiments on tones masking tones than in experiments of noise masking tones. For example, we can see that the 90 dB masking curve in *Figure 12* peaks at roughly 75 dB implying a minimum SMR of roughly 15 dB. These types of results have been reproduced many times and the implication seems to be that noise is a better masker than tones. This phenomenon is referred to in the literature as the "asymmetry of masking" [Hellmann 72 and Hall 97].

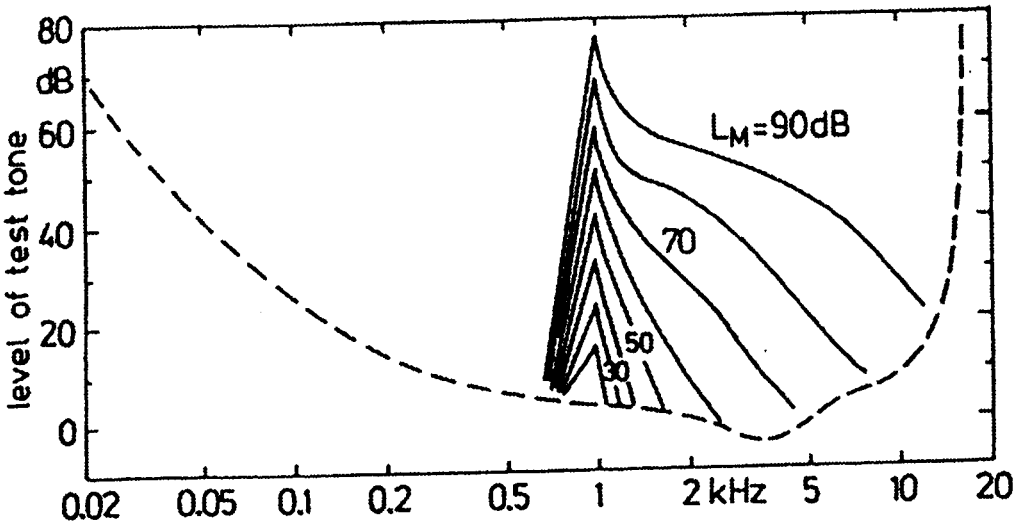

Figure 12. Masking thresholds for a 1 kHz tone masker at different levels masking tones from [Zwicker and Fastl 90]

7.3 Narrow-Band Noise or Tones Masking Narrow-Band Noise

Masking models exploited in perceptual audio coding rely upon the assumption that quantization noise can be masked by the signal. Often the codecs' quantization noise is spectrally complex rather than tonal. In this context, therefore, a suitable masking model might be better derived from experimental data collected in the case of narrow-band noise probes masked by narrow-band noise or tonal maskers. Unfortunately, there is very little data in the literature that address this issue. In the case of narrow-band noise probes masked by narrow-band noise maskers, phase relationships between the masker and the probe largely affect the results. According to [Hall 98] and based on [Miller 47], measurements for wide-band noise lead to minimum SMRs of about 26 dB.

In the case of tones masking narrow-band noise, early work by Zwicker and later others [Schroeder, Atal and Hall 79 and Fielder 87], suggest that the minimum SMR levels are between 20 and 30 dB. In general, it appears that when the masker is tonal, the minimum SMR levels are higher than when the masker is noise-like.

8. CRITICAL BANDWIDTHS

In measuring frequency masking curves, it was discovered that there is a narrow frequency range around the masker frequency where the masking

threshold is flat rather than dropping off. For example, *Figure 13* shows the masking threshold for narrow-band noise at 2 kHz centered between two tonal maskers at 50 dB SPL as a function of the frequency separation of the two maskers. Notice how the masking threshold is flat at about 33 dB until the maskers are about 150 Hz away from the test tone (i.e. about 300 Hz away from each other) at which point it drops-off rapidly.

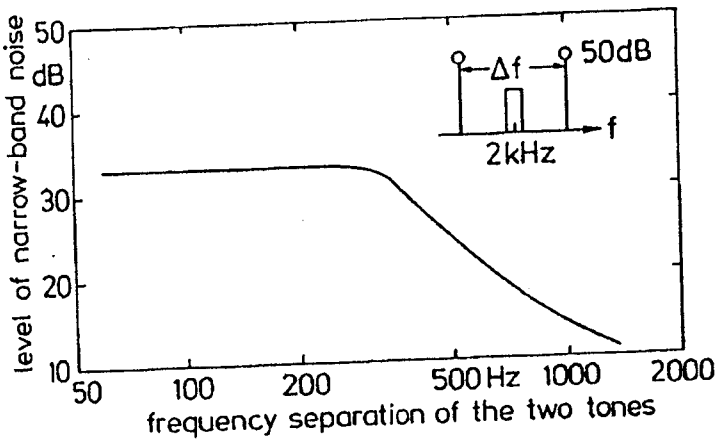

Figure 13. Threshold of a narrow-band noise centered between two sinusoidal maskers at a level of 50 dB as a function of the frequency separation between the two sinusoids from [Zwicker and Fastl 90]

Figure 14 shows analogous results for the case where the maskers are narrow-band noise and the test signal is tonal. Notice how the masking threshold is again flat until the maskers are about 150 Hz away from the maskee. Notice also that the level of masking at low frequency separations from these noise maskers is at roughly 46 dB (versus only roughly 33 dB when tonal maskers are employed), consistently with our earlier findings that noise-like maskers provide greater masking than tonal maskers. The main point, however, is that there is a so-called "critical bandwidth" around a masker that exhibits a constant level of masking regardless of the type of masker. The concept of critical bandwidth was first introduced by Harvey Fletcher in 1940 [Fletcher 40]. Fletcher's measurements and assumption led him to model the auditory system as an array of band-pass filters with continuously overlapping pass-bands of bandwidths equal to critical bandwidths. Experiments have shown that the critical bandwidth depends on the frequency of the masker. However, the exact form of the relationship between critical bandwidth and masker frequency is somewhat subject to controversy since differing results have been obtained using different types of measurements.

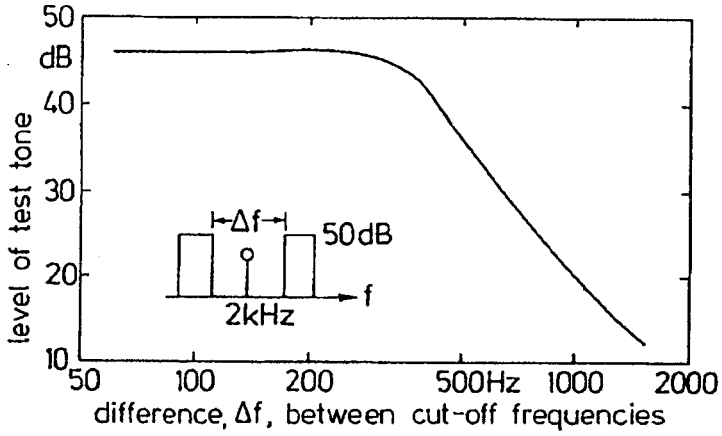

Figure 14. Threshold of a sinusoid centered between two narrow-band noise maskers at a level of 50 dB as a function of the frequency separation between the cut-off frequencies of the noise maskers from [Zwicker and Fastl 90]

Since the early work by Fletcher, different methods for measuring critical bandwidths have been developed and the resulting empirical data seem to differ substantially for frequencies below 500 Hz. In the pioneering work of Fletcher and later work by Zwicker [Zwicker 61], the critical bandwidth was estimated to be constant at about 100 Hz up to masker frequencies of 500 Hz, and to be roughly equal to 1/5 of the frequency of the masker for higher frequencies. An analytical expression that smoothly describes the variation of critical bandwidth Δf as a function of the masker center frequency f_c is given by [Zwicker and Fastl 90]:

$$\Delta f / \mathrm{Hz} = 25 + 75\left[1 + 1.4(f_c / \mathrm{kHz})^2\right]^{0.69}$$

This formula for critical bandwidths is widely accepted as the standard description of them.

8.1 Equivalent Rectangular Bandwidth

A number of articles including Greenwood [Greenwood 61], Scharf [Scharf 70], Patterson [Patterson 76], Moore and Glasberg [Moore and Glasberg 83] disagree in their estimation of the critical bandwidths with that of the standard formula, especially below 500 Hz. In particular, Moore and Glasberg measure a quantity they define called the "equivalent rectangular bandwidth", ERB, which should be equivalent to the critical bandwidth. Their experiments were designed to provide an estimate of the auditory filter shapes by detecting the threshold of a sinusoidal signal masked by notched noise as a function of the width of the notch. The ERB as defined by Moore

and Glasberg is about 11% greater than the -3 dB bandwidth of the auditory filter under consideration. The ERB, as a function of the center frequency f_c of the noise masker, is well fit by the function [Moore 96]:

$$\text{ERB/Hz} = 24.7\ (4.37\ f_c/\text{kHz} + 1)$$

The ERB function seems to provide values closer to the critical bandwidth measurements of Greenwood [Greenwood 61] than of Flecther or Zwicker at low frequencies. *Figure 15* compares the standard critical bandwidth formula with Moore's ERB formula and with other experimental measurements of critical bandwidth. Notice that the critical bandwidths predicted by the ERB formula are much narrower at frequencies below 500 Hz than implied by the standard critical bandwidth formula. Since the critical bandwidth represents the width of high-level masking from a signal, narrower critical bandwidth estimates put stronger requirements on a coder's frequency resolution.

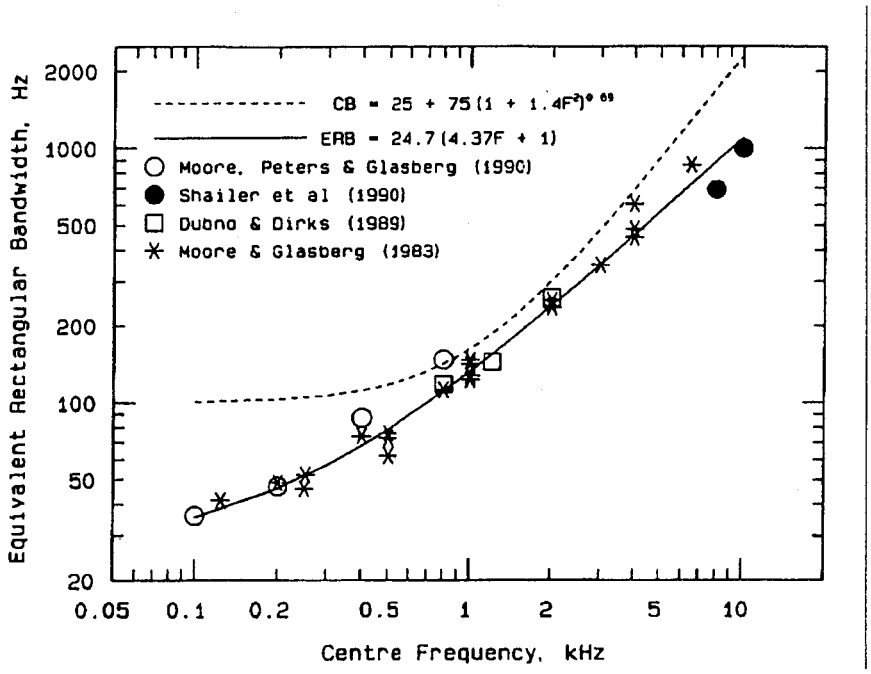

Figure 15. Critical bandwidth function and the ERB function plotted versus different experimental data for critical bandwidth from [Moore 96]

In summary, we have found that we can measure frequency masking curves for various masking and test signals. In all cases, we find that the masking curve levels are highest at frequencies near the masker frequency and drop off rapidly as the test signal frequency moves more than a critical bandwidth away from the masker frequency. We have seen that the shape of the masking curves depend on the frequency of the masker and its level. We have also seen that the masking curves depend strongly on whether or not

the masker is tonal or noise-like, where we have seen that much greater masking is created by noise-like maskers. We now turn to describe how hearing works to help us interpret the empirical data we have just seen and create models that link them together.

9. HOW HEARING WORKS

A schematic diagram of the human ear is shown in *Figure 16*. The outer, middle, and inner ear regions are shown. The main role of the outer ear is to collect sound and funnel it down the ear canal to the middle ear via the eardrum. The middle ear translates the pressure wave impinging on the eardrum into fluid motions in the inner ear's cochlea. The cochlea then translates its fluid motions into electrical signals entering the auditory nerve.

We can distinguish two distinct regions in the auditory system where audio stimuli are processed:

1. The peripheral region where the stimuli are pre-processed but retain their original character
2. The sensory cells which create the auditory sensation by using neural processing.

The peripheral region consists of the proximity zone of the listener where reflections and shadowing take place through the outer ear and ear canal to the middle ear. The sensory processing takes place in the inner ear.

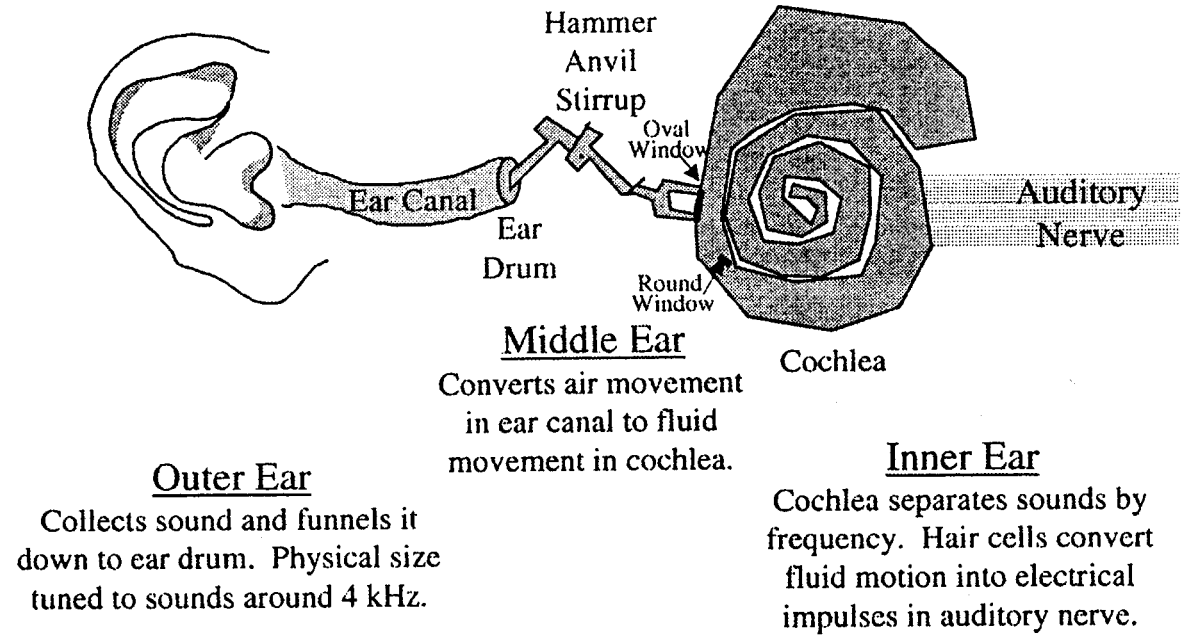

Figure 16. Outer, middle, and inner ear diagram.

9.1 Outer Ear

A sound field is normally approximated by a plane wave as it approaches the listener. The presence of the head and shoulders then distorts this sound field prior to entering the ear. They cause shadowing and reflections in the wave at frequencies above roughly 1500 Hz. This frequency corresponds to a wavelength of about 22 cm, which is considered a typical head diameter. The outer ear and ear canal also influence the sound pressure level at the eardrum. The outer ear's main function is to collect and channel the sound down to the eardrum but some filtering effects take place that can serve as an aid for sound localization. The ear canal acts like an open pipe of length roughly equal to 2 cm, which has a primary resonant mode at 4 kHz (see *Figure 17*). One can argue that the ear canal is "tuned" to frequency near its resonant mode. This observation is confirmed by the measurements of the threshold in quiet, which shows a minimum, i.e. maximum sensitivity, in that frequency region.

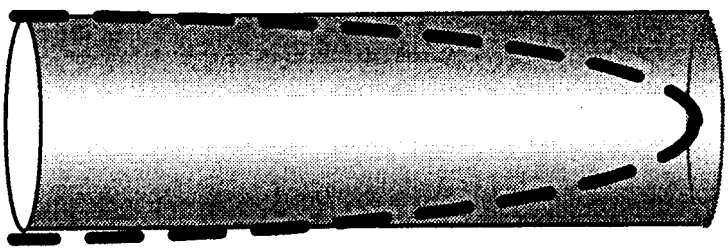

Figure 17. Outer ear model as an open pipe of length of about 2 cm

9.2 Middle Ear

The middle ear converts air movement in the ear canal into fluid movement in the cochlea. The hammer, anvil, and stirrup combination acts as lever and fulcrum to convert large, low-force displacements of air particles against the eardrum into small, high-force fluid motions in the cochlea. To avoid loss of energy transmission due to impedance mismatch between air and fluid, the middle ear mechanically matches impedances through the relative areas of the eardrum and stirrup footplate, and with the leverage ratio between the hammer and anvil arm. This mechanical transformer provides its best match in the impedances of air and cochlear fluid at frequencies of about 1 kHz. The stirrup footplate and a ring-shaped membrane at the base of the stirrup called the oval window provide the means by which the sound waves are transmitted into the inner ear. The frequency response of the filtering caused by the outer and middle ear can be described by the following function [Thiede et al. 00]:

$$A'(f)/dB = 0.6 * 3.64(f/kHz)^{-0.8} - 6.5e^{-0.6(f/kHz-3.3)^2} + 10^{-3}(f/kHz)$$

9.3 Inner Ear

The main organ in the inner ear is the cochlea. The cochlea is a long, thin tube wrapped around itself two and a half times into a spiral shape. Inside the cochlea there are three fluid-filled channels called "scalae" (see *Figure 18* for a cross sectional view): the scala vestibuli, the scala media, and the scala tympani. The scala vestibuli is in direct contact with the middle ear through the oval window. The scala media is separated from the scala vestibuli by a very thin membrane called the Reissner membrane. The scala tympani is separated from the scala media by the basilar membrane. From the functional point of view, we can view the scala media and the scala vestibuli as a single hydro-mechanical medium. The important functional effects involve the fluid motions across the basilar membrane. The basilar membrane is about 32 mm long and is relatively wide near the oval window while it becomes only one third as wide at the apex of the cochlea where the scala tympani is in direct fluid contact with the scala vestibuli through the helicotrema. The basilar membrane supports the organ of Corti (see *Figure 18*), which contains the sensory cells that transform fluid motions into electrical impulses for the auditory nerve.

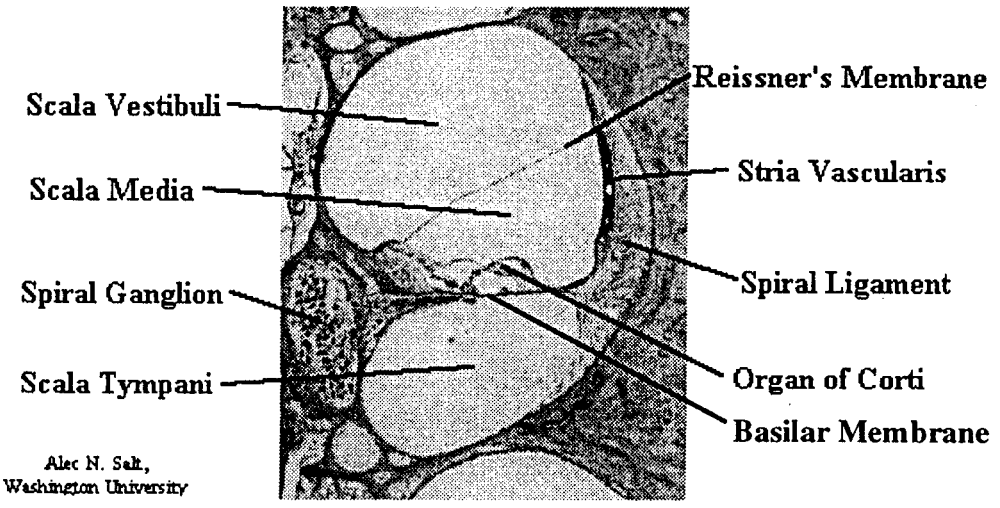

Figure 18. Cross section of the cochlea showing the scalae and organ of Corti. (Courtesy of Professor Alec N. Salt of Washington Unversity. Used with Permission.)

Figure 19 shows a functional diagram of the (unwrapped) cochlea [Pierce 83]. Fluid is displaced in the scala media/scala vestibuli by movements in the oval window driven by the middle ear. This fluid displacement is

equalized by movement of the basilar membrane or, for low frequencies, by fluid flow into the scala tympani through the helicotrema. Finally, the scala tympani fluid flow is equalized by offsetting movements of the round window, which is localized at the base of the scala tympani. The delay between the presentation of the signal at the oval window and the response of the basilar membrane increases with distance from the oval window. Such delay varies between less than 1 ms for high frequencies to above 5 ms for low frequencies.

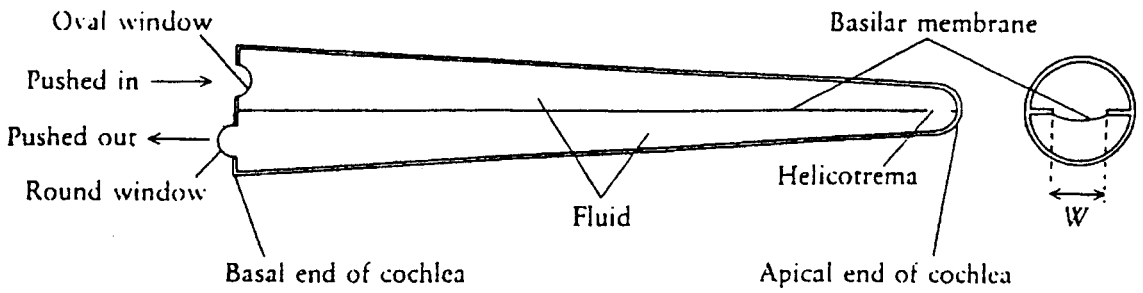

Figure 19. Functional diagram of the cochlea from [Pierce 83]

Georg von Békésy [von Békésy 60] experimentally studied fluid motions in the inner ear and proved a truly remarkable result previously proposed by von Helmholtz: the cochlea acts as a spectral analyzer. Sounds of a particular frequency lead to basilar membrane displacements with a small amplitude displacement at the oval window, increasing to peak displacements at a frequency-dependent point on the basilar membrane, and then dying out quickly in the direction of the helicotrema. *Figure 20* shows the displacement envelope that results from the motion of the basilar membrane in response to a 200 Hz frequency tone.

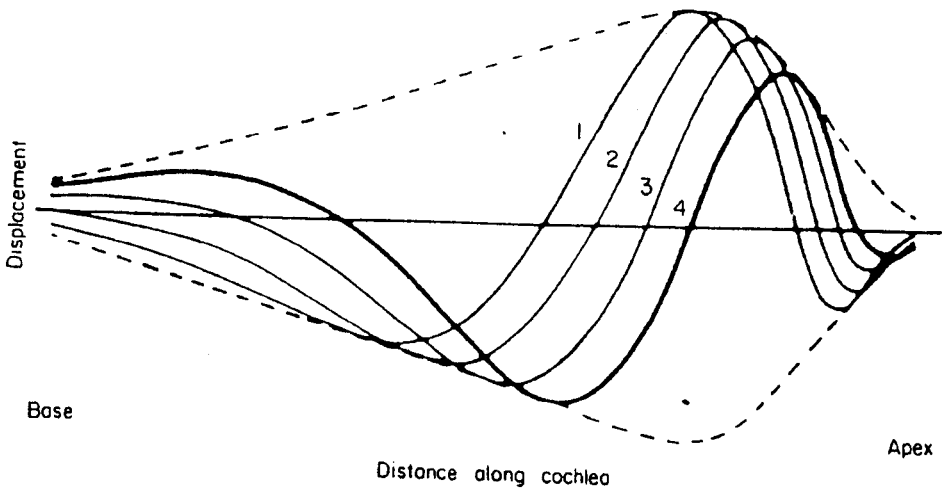

Figure 20. Traveling wave amplitude of the basilar membrane displacement relative to a 200 Hz frequency tone; the solid lines indicate the pattern at different instants in time; the dotted line indicates the displacement envelope from [von Békésy 60]

The experiments by von Békésy showed that low frequency signals induce oscillations that reach maximum displacement at the apex of the basilar membrane near the helicotrema while high frequency signals induce oscillations that reach maximum displacement at the base of the basilar membrane near the oval window. *Figure 21* shows the relative displacement envelopes of the basilar membrane for several different frequencies (50, 200, 800, 1600 Hz tones). *Figure 22* shows the locations of the displacement envelope peaks for differing frequencies along the basilar membrane from [Fletcher 40]. In this sense, it is often said that the cochlea performs a transformation that maps sound wave frequencies onto specific basilar membrane locations or a "frequency-space" transformation. The spectral mapping behavior of the cochlea is the basis for our understanding of the frequency dependence of critical bandwidths, which are believed to represent equal distances along the basilar membrane.

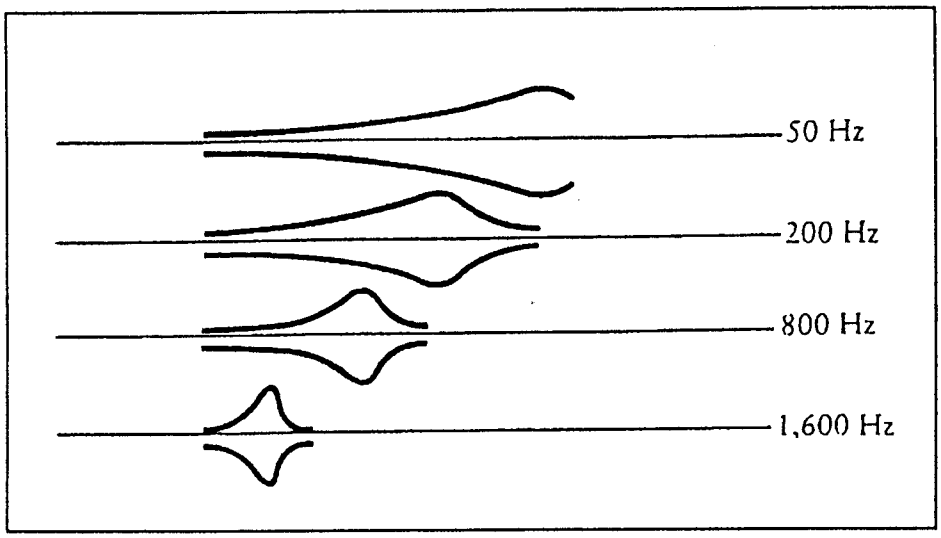

Figure 21. Plots of the relative amplitude of the basilar membrane response as a function of the basilar membrane location for different frequency tones; the left side of the plot is in proximity of the oval window, the right side of the plot is in proximity of the helicotrema from [Pierce 83]

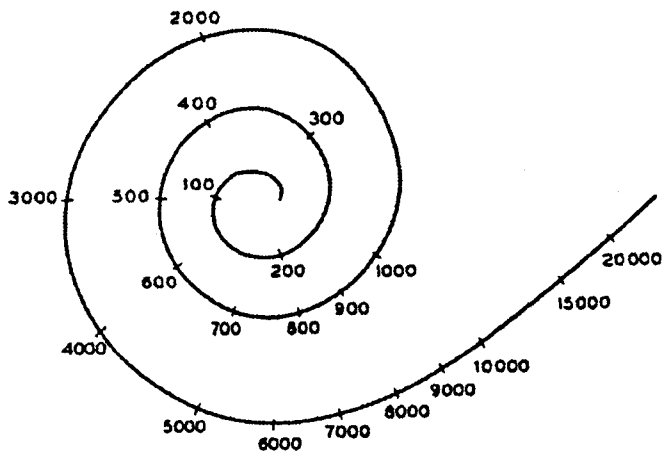

Figure 22. Frequency sensitivity along the basilar membrane from [Fletcher 40]. Copyright 1940 by the American Physical Society

On the basilar membrane, the organ of Corti transforms the mechanical oscillations of the basilar membrane into electrical signals that can be processed by the nervous system. The organ of Corti contains specialized cells called "hair cells" that translate fluid motions into firing of nerve cells in the auditory nerve. In the organ of Corti two types of sensory cells are contained: the inner and outer hair cells. Each hair cell contains a hair-like bundle of cilia that move when the basilar membrane oscillates. When the cilia move, ions are released into the hair cell. This release leads to neurotransmitters being sent to the attached auditory nerve cells. These nerve cells then send electrical impulses to the brain, which lead to the hearing sensation. The inner ear is connected to the brain by more than

30,000 auditory nerve fibers. The characteristic frequency of a fiber is determined by the part of the basilar membrane where it innervates a hair cell. Since the nerve fibers tend to maintain their spatial relation with one another, this results in a systematic arrangement of frequency responses according to location in the basilar membrane in all centers of the brain.

At high intensity levels, the basilar movement is sufficient to stimulate multiple nerve fibers while much fewer nerve fibers are stimulated at lower intensity levels. It appears that our hearing process is able to handle a wide dynamic range via non-linear effects (i.e., dynamic compression) in the inner ear. Structural differences between the inner and the outer hair cells indicate different functions for the two types of sensory cells. The inner hair cells play the dominant role for high-level sounds (the outer hair cells being mostly saturated for these levels). The outer hair cells play the dominant role at low levels, heavily interacting with the inner hair cells. In this case, the outer hair cells act as a non-linear amplifier to the inner hair cells with an active feedback loop and symmetrical saturation curves, allowing for the perception of very soft sounds.

It should be noted that in the inner ear a certain level of neural suppression of internal noise takes place. The effects of this noise suppression can be modeled by the following filtering of the signal [Thiede et al. 00]:

$$\text{Internal Noise} / \text{dB} = 0.4 * 3.64(\text{f} / \text{kHz})^{-0.8}$$

Summing this expression with that of the transfer function for the outer and middle ear, A'(f), one can derive the analytical expression A(f) that fits the experimental data for the threshold in quiet.

Finally, it is worth mentioning that at low frequencies, the nerve fibers respond according to the instantaneous phase of the motion of the basilar membrane while at frequencies above 3500 Hz there is no phase synchronization. Comparing intensity, phase, and latency in each ear, we are provided physical clues as to a sound source's location.

10. SUMMARY

In this chapter we have learned that the human ear can only hear sound louder than a frequency dependent threshold. We have seen that we can hear very little below 20 Hz and above 20 kHz. We extensively discussed the phenomenon of masking. Masking is one of the most important psychoacoustics effects used in the design of perceptual audio coders since it

identifies signal components that are irrelevant to human perception. Masking depends on the spectral composition of both the masker and maskee, on their temporal characteristics and intensity, and it can occur before and after the masking signal is present (temporal masking) and simultaneously with the masker. The experiments we have reviewed show that frequency masking is most pronounced at the frequency of the masker with rapid drop off as the frequency departs from there and that the ear has a frequency dependent limit to its frequency resolution in that masking is flat within a "critical band" of a masker. We discussed how the auditory system can be described as a set of overlapping band-pass filters with bandwidths equal to critical bandwidths. Examining how the hearing process works, we found that air oscillations at the eardrum are converted into oscillations of the basilar membrane, where different parts of the basilar membrane are excited depending on the frequency content of the signal, and then into auditory sensation sent to the brain. In the next chapter, we will show how to put these observations to use in audio coding.

11. REFERENCES

[Bosi and Davidson 92]: M. Bosi and G. A. Davidson, "High-Quality, Low-Rate Audio Transform Coding for Transmission and Multimedia Applications", Presented at the 93rd AES Convention, J. Audio Eng. Soc. (Abstracts), vol. 40, P. 1041, Preprint 3365, December 1992.

[Fielder 87]: Louis D. Fielder, "Evaluation of the Audible Distortion and Noise Produced by Digital Audio Converters", J. Audio Eng. Soc., Vol. 35, no. 7/8, pp. 517-535, July/August 1987.

[Fletcher 40]: H. Fletcher, "Auditory Patterns", Rev. Mod. Phys., Vol. 12, pp.47-55, January 1940.

[Fletcher and Munson 33]: H. Fletcher and W. A. Munson, "Loudness, Its Definition, Measurement and Calculation ", J. Acoust. Soc. Am., Vol. 5, pp. 82-108, October 1933.

[Greenwood 61]: D. Greenwood, "Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane", J. Acoust. Soc. Am., Vol. 33 no. 10, pp. 1344-1356, October 1961.

[Hall 97]: J. L. Hall, "Asymmetry of Masking Revisited: Generalization of Masker and Probe Bandwidth", J. Acoust. Soc. Am., Vol. 101 no. 2, pp. 1023-1033, February 1997.

[Hall 98]: J. L. Hall, "Auditory Psychophysics for Coding Applications", in *The Digital Signal Processing Handbook*, V. Madisetti and D. Williams, CRC Press, pp. 39.1-39.25, 1998.

[Hellman 72]: R. Hellman, "Asymmetry of Masking Between Noise and Tone", Percep. Psychphys., Vol. 11, pp. 241-246, 1972.

[Miller 47]: G. A. Miller, "Sensitivity to Changes in the Intensity of White Noise and its Relation to Masking and Loudness", J. Acoust. Soc. Am., Vol. 19 no. 4, pp. 609-619, July 1947.

[Moore 96]: B. C. J. Moore, "Masking in the Human Auditory System", in N. Gilchrist and C. Gerwin (ed.), *Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 9-19, AES 1996.

[Moore and Glasberg 83]: B. C. J. Moore and B. R. Glasberg, "Suggested Formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns ", J. Acoust. Soc. Am., Vol. 74 no. 3, pp. 750-753, September 1983.

[Patterson 76]: R. D. Patterson, "Auditory Filter Shapes Derived with Noise Stimuli", J. Acoust. Soc. Am., Vol. 59 no. 3, pp. 640-650, March 1976.

[Pierce 83]: J. Pierce, *The Science of Musical Sound*, W. H. Freeman, 1983.

[Scharf 70]: B. Scharf, "Critical Bands", in *Foundation of Modern Auditory Theory*, New York Academic, 1970.

[Terhardt 79]: E. Terhardt, "Calculating Virtual Pitch", Hearing Res., Vol. 1, pp. 155-182, 1979.

[Thiede et al. 00]: T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg and B. Feiten, "PEAQ-The ITU Standard for Objective Measurement of Perceived Audio Quality", J. Audio Eng. Soc., Vol. 48, no. 1/2, pp. 3-29, January/February 2000.

[von Békésy 60]: G. von Békésy, *Experiments in Hearing*, McGraw-Hill, 1960.

[Zwicker 61]: E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," J. Acoust. Soc. of Am., Vol. 33, p. 248, February 1961.

[Zwicker and Fastl 90]: E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin Heidelberg 1990.

12. EXERCISES

Masking Curves Framework:
In this exercise you will develop the framework for computing the masking curve for a test signal. We will return to this test signal in the next chapters to complete the masking curve calculation and utilize these results to guide the bit allocation for this signal.

1. Use an FFT to map a 1 kHz sine wave with amplitude equal to 1.0 into the frequency domain. Use a sample rate of 48 kHz and a block length of N = 2048. Do your windowing using a sine window. How wide is the peak? What is the sum of the spectral density $|X[k]|^2$ over the peak? Try dividing this sum by $N^2/8$, how does the result relate to the amplitude of the input sine wave? (Check that you're right by changing the amplitude to ½ and summing over the peak again.) If we define this signal as having an SPL of 96 dB, how can you estimate the SPL of other peaks you see in a test signal analyzed with the same FFT?
2. Use the same FFT to analyze the following signal:

$$\begin{aligned} x[n] = {} & A_0 \cos(2\pi 440n / F_s) + A_1 \cos(2\pi 554n / F_s) \\ & + A_2 \cos(2\pi 660n / F_s) + A_3 \cos(2\pi 880n / F_s) \\ & + A_4 \cos(2\pi 4400n / F_s) + A_5 \cos(2\pi 8800n / F_s) \end{aligned}$$

   where $A_0 = 0.6$, $A_1 = 0.55$, $A_2 = 0.55$, $A_3 = 0.15$, $A_4 = 0.1$, $A_5 = 0.05$, and F_S is the sample rate of 48 kHz. Using the FFT results, identify the peaks in the signal and estimate their SPLs and frequencies. How do these results compare with what you know the answer to be based on the signal definition?
3. Apply the threshold in quiet to this spectrum. Create a graph comparing the test signal's frequency spectrum (measured in dB) with the threshold in quiet.