Units of sound

As mentioned in lecture 22, sound consists of waves of air pressure that are measured by the ear. Air pressure is always positive. It oscillates about some mean value X_a which we call *atmospheric pressure.* The units of air pressure are atmospheres and the mean air pressure around us is approximately "one atmosphere". When we talk about an audio signal, we are typically refering to the perturbation of the air pressure about its mean. Then the time varying air pressure that we call the sound is $X_a + X(t)$. The audio signal is the $X(t)$ part, that is, the time varying perturbation about the mean value.

These perturbations are quite small. The quietest sound that we can hear is about 10^{-9} atmospheres. The loudest sound that we can tolerate without pain is about 10^{-3} atmospheres. Thus, we are sensitive to 6 orders of magnitude of pressure changes. (An *order of magnitude* is a factor of 10.)

A common measure of "physical loudness" of a sound is a ratio of the pressure of the sound pressure level $X(t)$ relative to some standard – call it X_0 – which is very soft, namely on the threshold of hearing. We cannot talk about the instantaneous loudness of a sound $X(t)$. Instead we take the mean squared value over some short time interval (such as a block in a spectrogram). Call this $\overline{X^2}$.

One typically measures loudness by the log of the ratio of the above two quantities. Define:

$$\text{Bels} = \log_{10} \frac{\overline{X^2}}{X_0^2}$$

or more commonly

$$\text{Decibels} \equiv 10 \log_{10} \frac{\overline{X^2}}{X_0^2}$$

The reason one multiplies by 10 is that the human auditory system is limited in its ability to discriminate sounds of different loudnesses, such that (roughly speaking) we can typically discriminate sounds that differ from each other by a loudness of about 1 dB.

Here are a few examples of sound loudness:

Sound	dB
jet plane taking off (60 m)	120
noisy traffic	90
conversation (1 m)	60
middle of night quiet	30
recording studio	10
threshold of hearing	0

Human hearing: frequency sensitivity

The human auditory system can hear over a certain range of frequencies (20 Hz to 44 kHz, roughly) but is not equally sensitive to all frequencies in this range. This is important for lossy audio compression because an encoder shouldn't spend as many bits on frequency components that the human ear is less sensitive to. But what do we mean by sensitivity? How do we measure this? Let me explain this in terms of a concrete experiment.

[ASIDE: The following deviates slightly from what I set up in class. Here I try to relate more closely to language of spectrograms and the method that I will discuss next class.]

Suppose a person (known as an experimental “subject”) is seated in a relatively quiet room and is given the task of deciding in which of two short time intervals a sound is played. You can think of these intervals as single blocks of a spectrogram.

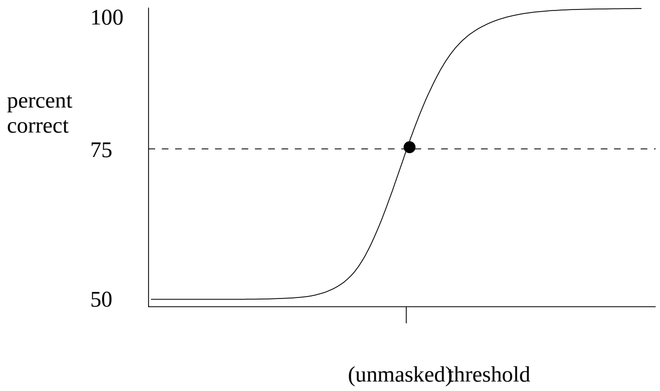
Say the two intervals are specified by a light that turns on at the beginning of the interval and off at the end of the interval. The sound in question is played in one and exactly one of the intervals. The experiment is repeated many times, and the sound is randomly played in either the first or the second interval (randomly chosen) and at intensity values that are randomly chosen. For each pair of intervals, the subject needs to respond either “interval 1” or “interval 2”.

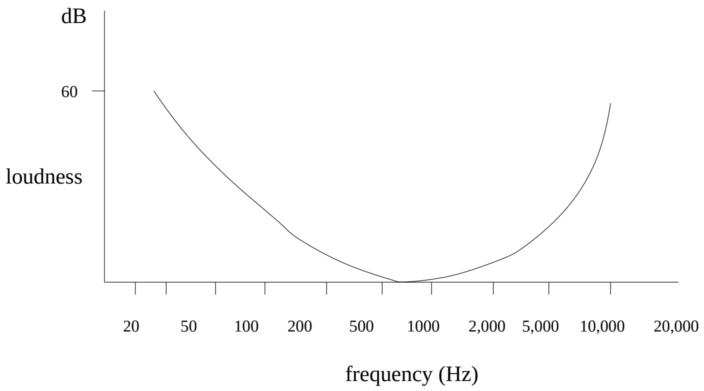
The experimentor then plots the percentage of intervals in which the subject makes the correct response as a function of the physical loudness (pressure amplitude) of the sound. This gives a curve which ramps up from 50% (guessing, when the sound level is too low) to 100% (easy, when sound level is sufficiently high). The experimentor chooses some arbitrary score (say, 75% correct). The pressure corresponding to this score is called it the *threshold.*

Many such experiments have been done. For example, suppose the sound frequency to be detected is a single *target* frequency k_T played over a block size of $m = 8192$ samples (about one fifth of a second), and the two blocks (intervals) in the experiment are separated by some *interstimulus interval*, say 1 second. If you repeat the experiment with various target frequencies k_T, then you can measure how the threshold varies with frequency. You get an inverted U shaped curve which has its minimum around 2000-5000 Hz. (See right sketch below.) Note that this plotted on a log scale (dB) so the sensitivity differences are quite huge.

This difference in sensitivity could be used for audio compression. For example, the encoder might use fewer bits to encode the very low and very high frequencies, and more bits for the middle frequencies. It could do so by using a relatively large Δ for low and high frequencies, and a smaller Δ for medium frequencies.

Notice that the above scheme would choose the Δ values independently of the audio file itself. A more sophisticated scheme would base the Δ values on the audio file itself. To understand how this might work, let’s generalize the above experiment and address the phenomenon of *masking.*



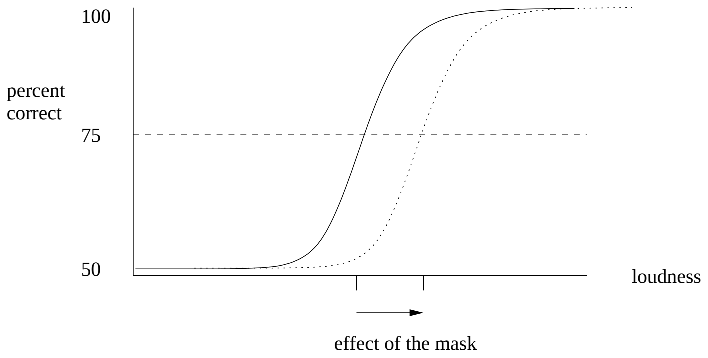


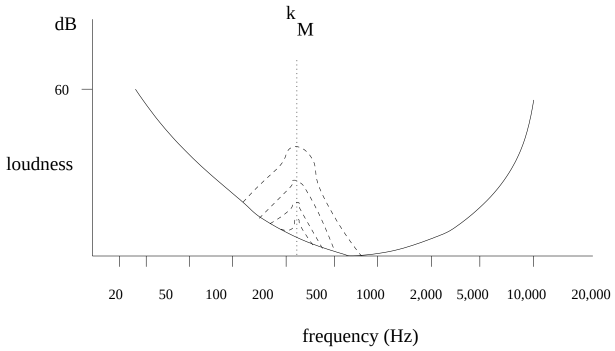
Masking

Masking is the familiar phenemonon that it is easier to hear the details of a sound when it occurs by itself than when it is accompanied by a second sound. If you are listening to one sound (music) while another sound is present (say someone talking), then it is difficult to hear the details of either one of them.

Let's generalize the experiment I mentioned above, and now consider two frequencies, k_T and k_M, where T is again "target" and now M stands for "mask". Suppose that k_M is played in both intervals at some loudness level $Y_j(k_M)$, and the target frequency k_T is played in only one of the two intervals. The task is to say which of the two intervals contains k_T. Again, for any fixed $Y_j(k_M)$, the experimentor varies $Y_j(k_T)$ and examines how large $Y_j(k_T)$ needs to be in order for the subject to identify the interval correctly (1 vs. 2) in 75 percent of the cases.

Basically, this experiment measures how well you can hear one frequency (k_T) in the presence of a second frequency (k_M, the mask). Not surprising, people become worse at detecting k_T when the mask gets louder i.e. $Y_j(k_M)$ increases. Specifically, when you plot "percent correct" for detecting k_T in the presence of the mask, the detection curve shifts to the right, relative to what I mentioned above (see below, top sketch) and the amount by which it shifts increases with $Y_j(k_M)$. This means that the listener needs the loudness of the target i.e. $Y(k_T)$ to be greater in order to able to detect the prescence of k_T with the same probability. We say that *frequency k_M masks frequency k_T*, and we say the *amount of masking is the amount of the rightward shift in the curve at threshold level (e.g. 75 percent correct).*





One general finding is that a frequency k_M will mask another frequency k_T more when $\mid k_M - k_T \mid$ is small. This effect is illustrated above-right. Each curve shows the amount of masking by frequency k_M on frequency k_T, for a given loudness value of k_M, that is, for fixed $Y(k_M)$. Different dashed curves correspond to different loudness values of the masking frequency. The louder the masking sound the greater is the masking effect. The key point is that the masking effect is restricted to frequencies k_T near k_M.

The above fact suggests that, to encode an audio signal using lossy compression, we could choose the Δ values based on the audio signal $X(t)$ itself. For example, if two frequencies k_M are k_T similar and $|Y_j(k_M)| \gg |Y_j(k_T)|$, then there is no point in encoding the non-zero value of $|Y_j(k_T)|$ since this component of the sound won't be heard. We would set $\Delta_j(k_T)$ large in this case.

A few interesting things to consider: First, we can understand these effects in terms of what happens in the ear, and what we discussed last class. Different positions along the basilar membrane respond to different frequencies of sound, and if two positions are sufficiently far from each other, then there is basically no interaction in their responses. The cells at these different parts of the basilar membrane thus encode the sound levels at the corresponding different frequencies, and they do so independently of each other.

Second, masking doesn't just occur for simultaneous sounds. It can also occur for sounds that occur nearby in time, i.e. in nearby blocks of the signal. For example, in terms of the spectrogram representation, $Y_j(k_M)$ can mask $Y_{j+1}(k_T)$ which is called *forward masking*, and strangely enough, $Y_{j+1}(k_M)$ can mask $Y_j(k_T)$ which is called *backwards masking.* Many experiments have been done to quantify these and other masking effects. The reason why forward and backwards masking is not so obvious (in terms of the underlying anatomy) but it is a strong effect nonetheless.

To summarize: what are the implications of these masking effects? For any interesting audio signal $X(t)$, the spectrograms $Y_j(k)$ will contain non-zero values at many frequencies k. Since nearby pairs of frequencies/times (j,k) will mask each other, one can use a larger Δ at frequencies/times when the neighbors have large values. Two questions arise. First, how to decide what are the $\Delta_j(k)$? Second, how can these $\Delta_j(k)$ be encoded? We will turn to these questions next lecture (in particular, the second question).