Spectrograms (revisited)

We begin the lecture by reviewing the units of spectrograms, which I had only glossed over when I covered spectrograms at the end of lecture 19. We then relate the blocks of a spectrogram to auditory filters and spend the remainder of the lecture on the latter.

Recall that spectogram partitions a signal $I(t)$ into B blocks of length T samples, each and then takes the amplitude spectrum of each block. Each block is typically 10-100 ms. The spectrogram is meant to capture events at that time scale, such as the glottal pulses or parts of speech sounds (vowels versus consonants, voiced vs unvoiced, etc).

The units of spectrograms need to be treated carefully. The Fourier transform of a block uses frequency ω in units of cycles per block, that is, cylces per T samples. These frequency units can be converted to cycles per second by multiplying by ω_0, which is the number of blocks per second. We can think of ω_0 as the fundamental frequency that is represented by the spectrogram.

The block number b can be converted to time in seconds by multiplying by seconds per block, or $\frac{1}{\omega_0}$. The number of samples per block is then $\frac{1}{\omega_0}$ times the number of samples per second. High quality audio signals usually have 44,100 samples per second. To put this another way, if you choose T samples for each block, then dividing T by 44,100 samples per second gives the number of seconds per block.

Putting those conversions aside, it is important to realize that time scales of 10 ms to 100 ms are quite large, relatively to the time scales that we were discussing last lecture when we considered spatial localization. Sound travels at 340 ms^{-1} and so 10 ms sound duration corresponds to 3.4 meters. If a block of a spectrogram is 10 ms long, then this covers 3.4 m of a snapshot of sound. The component of the sound at such a wavelength does not play a role in spatial hearing since the two ears would be at nearly the same phase of the wave at any time and the shadowing by the head and the pinna effects are negligable for such long waves.

Auditory Filters

We have discussed filtering of sound by the outer ear, and last lecture we discussed filtering by the basilar membrane. Researchers have also examined the frequency response properties of ganglion cells in the cochlea by measuring spikes of axons in the cochlear nerve, and researchers have also measured cell responses in the brainstem of various animals. These experiments often use pure tone stimuli. An example of a plot showing different cells and their thresholds for responding to pure tone stimuli was shown in the slides. Typically cells in the cochlear or brainstem have a peak (or center) frequency to which they are tuned. Indeed this is what we meant last lecture when we discussed the tonotopic map of cells along the basilar membrane and in areas such as the cochlear nucleus and MSO and LSO.

Masking and Critical bands

It is also possible to measure and model auditory filters using human or animal psychophysics experiments. A common experiment is to ask how good are we at discriminating two different frequencies. For example (not discussed in class), consider an experiment in which two tones are played, one following the other, and the listener is asked to say whether the tones are the same or different. Another example is *masking experiments*: one tone is presented twice (called the masking

tone) one after the after, and another tone is presented just once, namely at the same time as one of the two masking tones. The question is, how loud does the second (called *test*) tone need to be for you to hear it i.e. to say which of the two intervals contains the test. One typically holds the test tone at some frequency and sound pressure level, and varies the frequency and loudness of the masking tone. We say that the masking tone *masks* the test tone.

Many masking experiments have been done, and consistently show that similar frequencies mask each other much more than different frequencies mask each other. This is consistent with the fact that the cochlea decomposes sounds into bands and then encodes the bands independently. If two frequencies are coded in different bands (or frequency "channels"), then they tend to mask each other less. One often speaks of *critical bands* that cover the range of temporal frequencies that our auditory system is sensitive to.

Models of auditory filters of sound that are based on masking experiments have characterized the bands as follows:

- Below 1000 Hz, human can discriminate two frequencies reliably when they differ by more than about 100 Hz. For this reason, many models of auditory processing begin by filtering the sound below 1 kHz by using about 10 channels, each 100 Hz wide.

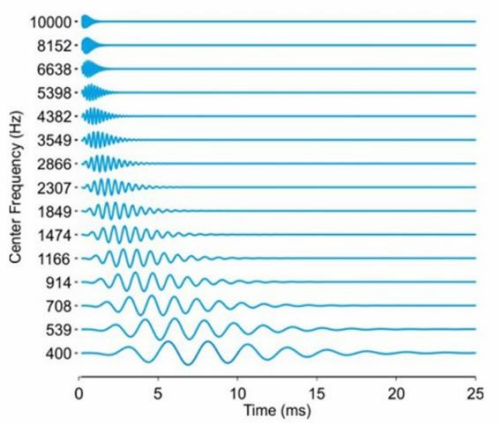
- Above 1000 Hz, humans can discriminate two frequencies reliably when they differ by more than $\frac{1}{3}$ of an octave. For this reason, many models of auditory processing filter the sound from 1,000 Hz to 22 kHz using about 14 channels. i.e. $3 \log_2 22 \approx 14$ since there are 3 bands per octave and $\log_2 22$ octaves from 1 to 22 kHz.

When we refer to *critical bands*, we often think of a partioning up of the frequency range. Note, however, that a 'partition' (mutually exclusive ranges of frequencies) is a convenient model, but does not describe the coding that occurs. There is no partition or boundary between frequency bands, but rather the bands form a continuum of frequencies.

Gammatone filters

The frequency behavior of auditory filter models are similar – whether we are referring to a basilar membraine mechanical response, a ganglion cell or brainstem cell response, or even a psychophyscal response namely critical bands. For this reason, one often conceptually does not distinguish which mechanism we are talking about.

Keeping it general, therefore, let's think of auditory filters as defining an impulse response function (or its Fourier transform, a transfer function). We can model these filters using Gabor functions of various center frequencies and bandwidths. One limitation with gabor function is that they have (Gaussian) tails that go off to infinity. The filter will have some peak sensitivity at some time in the past, but the tail of the filter will reach into the future which of course is impossible since a cell cannot respond to a sound that hasn't occured yet. (This same issue of "causality" came up with motion cells in vision.) The usual way around this is to use a slightly different window than a Gaussian, namely one which is asymmetric and goes to 0. In audition, one often uses a *gammatone filter*. See `https://en.wikipedia.org/wiki/Gammatone_filter` for the formula.



Examples are shown above. The lowest curve shows a filter with center frequency 400 Hz, so it is most sensitive to sine component whose period is 2.5 ms. You can verify for yourself that the lowest curve has roughly this period for its waves. As the center frequency increases for different curves shown, the period of the waves decreases. In addition, note that for lower frequency filters, the peak of the envelope occurs at a greater time in the past. One way to think of this is to imagine the cochlea and remember that the low frequency components are represented at the far end (the apex) and high frequencies are represented at the near end (base). If you think of the sound as a wave travelling through the cochlea, then this corresponds qualitatively[1] to the response at the near end occuring before that of the far end.

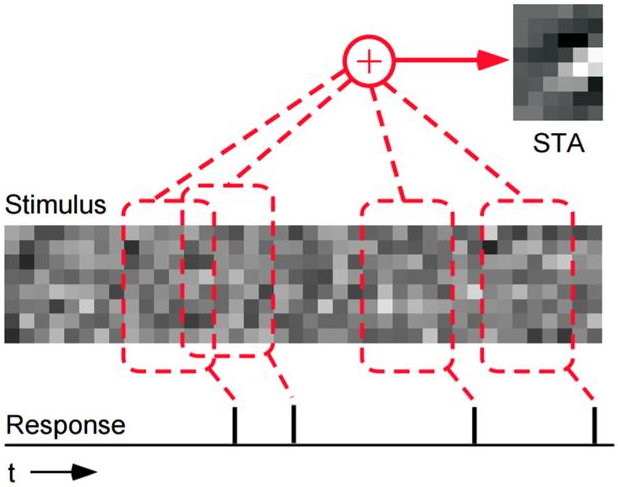
Spike triggered averaging: How to measure a cell's receptive field profile?

Several times in the course I have referred to a cell's receptive field profile. In vision, we saw center-surround cells in the retina and LGN, and we saw oriented cells in V1. In audition, I just mentioned gammatone filters. When I discussed the vision experiments from the 1950's of retinal cells and Hubel and Wiesel's measurement of V1 cells, I described their process as 'trial and error'. Present a stimulus over diffent positions in the visual field and perhaps at different sizes and orientations, and by hand determine which is the preferred stimulus. Then mark out the excitatory and inhibitory regions. This method is fine for some experiments. However, more systematic approaches have also been developed too.

One common method is the *spike triggered average*. The idea is use a random noisy signal as input, and to examine what specific values the signal takes which leads to the cell responding to the noise. The idea is that noise will occasionally by chance present a structure close to what the cell is tuned for, and when it does the cell will be more likely to spike. Spike triggered averaging takes two signals: the noise stimulus signal and the spike train response of the cell. For each spike, it considers a fixed block of time (say 300 ms) in the source signal *prior to that spike*. It then sums up these source signals. The idea is that if *something* in the signal caused the cell to spike at that time, then this something should be revealed by the spike triggered average. This approach has been quite successful.

[1](This 'travelling wave' turns out only to be qualitative, however, as the delay in the peaks of the curves shown doesn't correspond to the speed at which the sound wave propagates in the cochlear; rather it has more to do with the mechanics of the basilar membrane.

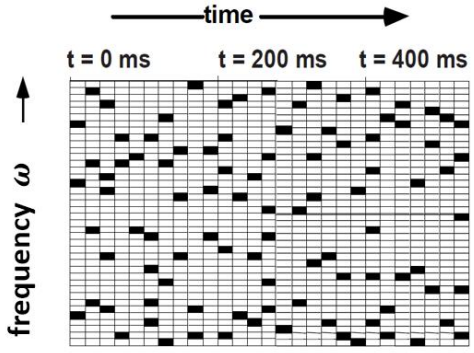
The example below shows an XT stimulus. The spike triggered average (STA) over four spikes is shown. In general one takes the average of the stimulus over thousands of spikes. A real example (for a V1 neuron) is shown in the slides.



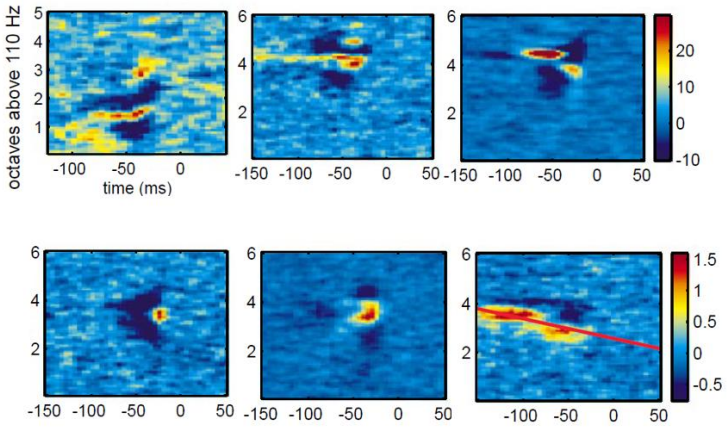
Auditory Cortex (A1)

Can spike triggered averaging be used to discover receptive fields in auditory cortex? (A1 is the audition analogue of V1, namely it is where the auditory signals are first processed in the cortex.) In principle, yes. However, in practice it has been difficult to do because many cells do not respond well to pure tones, regardless of the frequency. Moreover, spike triggered averaging doesn't work well either, if one uses a white noise stimulus e.g. the sound 'sssssssss'.

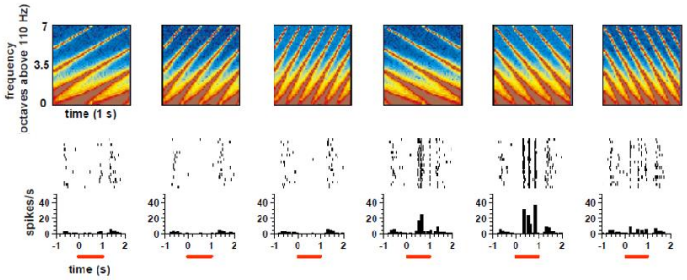
In the late 1990's, another idea for using spike triggered average was tested. Rather than using pure noise ("ssssssss"), instead random 'chords' were used which consisted of a sequence of short duration intervals of bandlimited noise. See illustration of a spectrogram of this random chord noise.



Examples of spike-triggered averages of some A1 cells are shown below. The axes are frequency versus time. Negative values of time indicate that the spikes (which were at time 0) responded to the parts of the sound that occured before the spike.



Notice that these cells are not simply excitatory for some particular band of frequencies. Rather, these A1 cells have both excitatory and inhibitory regions of the receptive field. The cell in the bottom right corner, in particular, seems to be sensitive to *frequency modulation* as indicated by the diagonal lines.



For this cell, the authors then confirmed that it was indeed sensitive to frequency modulation. They examined the responses of the cell to various FM modulated stimuli (see above). The plot shows six 'orientations' of FM modulated stimuli. The black dots below show rows of cell responses, namely spike trains. Each row is one trial i.e. one example where the sound plays. Each black dot is a spike. The plots the bottom show histograms where the rows of spikes are summed up – called a *peristimulus time histogram*. The main point here is that you get more spikes from the cell when the sound is FM modulated such that the frequencies decrease over time, as in the spike triggered average shown above (bottom right receptive field profile, with red line drawn on it).

At the end of the lecture, I briefly mentioned a few applications, namely cochlear implants and MP3 compression. Both of these applications are based on the theory of auditor filtering. I am leaving that discussion out of these lecture notes for now.

