DAT335 – Music Perception and Cognition
Cogswell Polytechnical College
Spring 2009

Week 1 – Class Notes

The Nature of Sound and the Structure and Function of the Auditory System

The Physical Characteristics of Sound

The Nature of Sound

Sound originates from the vibration of an object. This type of motion disturbs the medium surrounding it (usually air) as a change in pressure. The air molecules are being moved together (compression) and displaced (rarefaction) from their normal position. The sound wave moves outward from the source, but the molecules are not actually moving along with the wave; rather, they move along an average resting place. If the sound moves along an axis that is in the same direction from which the sound is propagated, we can call the wave "longitudinal".

The wave as it moves along its projection path can be subject to energy loss, reflection, and refractions caused by objects in its path. The sound we hear will be slightly different from what was initially generated.

The simplest type of sound is the sine wave or sinusoid. In order to describe this wave we must note the following things:

1) Frequency - number of times per second the waveform repeats itself.
2) Amplitude - amount of pressure variation about a mean.
3) Phase – portion of the cycle through which the wave has advanced in relation to a fixed point in time.

Another thing we must describe is the period or the time it takes for one complete cycle of the waveform to be completed. Period if the reciprocal of the frequency, $T = 1/f$.

Periodic sounds are those that repeat regularly and can be simple as sinusoids, or more complex like instrumental sounds.

All of these sounds have pitches, which is defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale". Simply put, they can be described as those sounds that create a sense of melody. These sounds are called "tones".

Pitch is related to repetition rate or frequency. However, it can not be directly measured since it is a subjective attribute. One way of "measuring" pitch is to compare a complex sound with the pitch of a sinusoid at a certain frequency.

Fourier Analysis and Spectral Representations

For complex sounds it is far more convenient and meaningful to represent them in terms of their frequency content.

Fourier's theorem states that any complex waveform can be analyzed into a series of sinusoids with specific frequencies, amplitudes, and phases. This process is called the Fourier transform, and each sinusoid is called a Fourier component. The complex waveform can be synthesized by adding together all the component sinusoids. Thus, any complex tone can be formed by adding simple sinusoidal tones.

Periodic complex tones are the simplest types of tones that can be analyzed with Fourier analysis. These tones have a fundamental, or lowest frequency present, and a series of harmonics that are integer multiples of the fundamental frequency. In this case, the fundamental frequency is the one responsible for providing the repetition rate (pitch) of the tone as a whole.

It is helpful to represent sounds in the frequency domain because humans are able to hear to some extent the individual harmonic contents of a periodic sound. This fact is also know as Ohm's acoustical law. Normally, humans hear complex sounds as a single pitch corresponding to the repetition rate of the sound as a whole.

When listening to two simultaneous pure tones, whose frequencies are slightly different, it is possible to hear two distinct tones each with its own pitch. In this manner, we are perceiving the sound in terms of its Fourier components.

Sound in terms of its frequency components can be represented by its magnitude spectrum, or its sound amplitude, energy, or power as functions of frequency. They are called amplitude spectrum, energy spectrum, or power spectrum.

Sinusoidal or pure sounds have, by definition, a single frequency component. Thus, its spectrum will show a single element at the sounds given frequency.

Other complex sounds, like square waves, will show more frequency components in its spectrum. In the square wave's case, we will observe odd numbered harmonics that decrease in amplitude as the number of harmonics increase.

Pulse trains contain all harmonics at equal amplitude. But will show a spectrum with many harmonics with low amplitudes due to the fact that each pulse contains a small amount of energy.

Non-periodic sounds, such as white noise, can also be analyzed with Fourier analysis. In the frequency domain, these type of sounds have a mixture of non-harmonically related sinusoids. Their spectrum will be continuous since the energy is spread over certain frequency bands. Examples are the single impulse and white noise. Their spectrum is similar, but the single impulse has less energy in its amplitude spectrum than that of white noise.

Finally, sinusoidal or tone bursts, have a line magnitude spectrum only is the sinusoid lasts a very long time. Tone bursts of short duration have a magnitude spectrum containing energy over a range of frequencies around the nominal frequency of the sinusoid. The spread of energy increases as the tone burst is shortened.

Because Fourier analysis requires in theory that the waveform to last an infinite amount of time, this is impossible in practical situations. Also, the human ear does not take an infinite time to process sounds. In this case, a short segment of the sound is taken and assumed that the waveform has zero value outside of these limits. This is known as “windowing” a waveform. The resulting spectrum will have its energy spread over a considerable range of frequencies because of the sharp discontinuities introduced by windowing the waveform. To reduce this effect, a window that has more taper near the edges and more weight given at the center of the window. The frequency resolution of a window depends on its duration and shape. Commonly, frequency resolution is equal to the reciprocal of the window duration.

For complex sounds, it is useful to analyze them as a succession of windows that overlap in time, so that the short-term spectrum with time can be calculated.

The Measurement of Sound Level

While the instruments (microphones) used to measure the magnitudes of sounds normally respond to air pressure fluctuations, sound magnitudes are usually expressed in terms of intensity. Sound intensity is defined as the sound energy transmitted per second (power) through a unit area in a sound field. In air, the relationship between pressure and intensity is the following: intensity is proportional to the square of the pressure variation.

Because the human ear has a very large range in sound intensity, it is convenient to express intensity in terms of logarithmic scales dealing with the ratio of two intensities. The first intensity, I_0, is a reference value and the other intensity, I_1, is relative to the other. The decibel is the unit used to convey sound intensity ratios.

$$number\ of\ decibels = 10\log(I0/I1)$$

The magnitude of a sound is usually referred as its “level”. Notice that the number of decibels represent a ratio, rather than an absolute intensity. In order to denote an absolute intensity, it is necessary to state that the intensity I_1 is n dB above or below a reference intensity I_0. The most commonly used reference intensity I_0 is equal to 2×10^{-15} W/m².

When a sound level is specified using this reference level its is referred to as a sound pressure level (SPL).

The reference sound level, 0dB SPL, was chosen as the average absolute threshold for humans for a 1000 Hz sinusoid. This is the minimum detectable sound level in the absence of other external sounds.

Finally, decibel notation can also express ratios of pressure. Since intensity is proportional to the square of the pressure, we can express this relation as such:

$$number\ of\ decibels = 10\log\left(\frac{I1}{I2}\right) = 10\log\left(\frac{P1}{P2}\right)^2 = 20\log\left(\frac{P1}{P2}\right)$$

Sound with line spectra can specify the sound level of either the overall (total) level or that of its individual components. The total level can be calculated from the total intensity of the sound, which is proportional to the mean square pressure. The pressure can be expressed as its root-mean-square (RMS) value. The RMS value is an average of the total pressure values.

Sound with continuous spectra can specify an overall level, but in order to specify the energy distribution we have to describe the energy, power, or intensity in terms of bandwidth (over a certain band of frequencies). The energy in contained in a band is known as the energy density. Power (energy per unit time) and intensity (energy per unit area) can also be described. When the intensity density of noise is expressed in decibels relative to the reference intensity I_0 it is know as the spectrum level.

For example, white noise has a spectrum level that does not change with frequency. Pink noise has a spectrum level that decreases by 3 dB for each doubling of frequency.

Beats

If two sinusoids with slightly different frequencies are played together, they will resemble a single sinusoid with a frequency equal to the mean of the two components, but the amplitude will fluctuate at a regular rate. This fluctuation in amplitude is known as “beats”. This phenomenon occurs because of the changing phase relationship between the two sinusoids, which cause them to reinforce or cancel each other. The resulting amplitude fluctuation rate is equal to the difference of the two frequencies.

The Concept of Linearity

For a system to be considered linear, the following conditions must be met:

1) If the input of a system is changed in magnitude by a factor *k,* then the output should also change in magnitude by the same factor, but other wise be altered.
2) The output of the system in response to a number of independent inputs presented simultaneously should be equal to the sum of the outputs that would have been obtained if each input were presented alone.

If a system is to be considered linear, it must be assumed that it is time-invariant. Meaning that the input-output function does not change over time. For example, for the input *I* and the output *O:*

$$O = cI$$

c being a constant that does not change with time.

Filters and their Properties

In order to manipulate the spectra of a complex stimuli, we must use an electronic device known as a filter. They alter an electronic signal by manipulating its spectrum. They are linear devices, for sinusoidal input, but they are designed to attenuate some frequencies of the output more than others. Examples of these are highpass, lowpass, bandpass, and bandstop filters.

Bandpass filters are of particular interest because at one stage of the peripheral auditory system is often likened to a bank of bandpass filters.

Although modern digital filters come close to achieving a infinitely sharp cutoff, there is a range of frequencies in which some components are attenuated, but never removed. Usually, it is necessary to define the cutoff frequency and the slope of the filter response curve.

The cutoff frequency is defined as the frequency ay which the output of the filter has fallen 3 dB relative to the output in the passband. The frequencies contained within the cutoff frequency define the bandwidth of the filter.

The slope of a filter defines the attenuation over equal amounts of frequency range outside the cutoff frequency.

Stimulus whose spectrum covers a wide frequency range is referred to as broadband. If a signal with a “flat” spectrum, white noise or impulses, are passed through a filter, the magnitude spectrum of the output of the filter will have the same shape as the filter characteristic. Altering the spectrum of a signal by filtering will produce a corresponding alteration in the waveform of a signal, and in they way it is perceived.

The response of a filter to an impulse will produce an impulse response that defines the filter characteristics. The flat magnitude spectrum of the impulse will produce the output of the filter to have the same shape of the filter characteristic. The filter characteristic can now be obtained by calculating the Fourier transform of the impulse response.

Basic Structure and Function of the Auditory System

The Outer and Middle Ear

The outer ear is composed of the pinna (the visible part) and the auditory canal or meatus. The pinna significantly modifies the incoming sound, especially high frequencies, and this is important in sound localization. Sound then travels down the meatus and causes the eardrum , or tympanic membrane, to vibrate. These vibrations are then transmitted through the middle ear by three small bones, the ossicles, to the oval window of the cochlea. The three bones are called the malleus, incus, and stapes. The stapes actually makes contact with the oval window.

The middle ear has two functions:

1) Ensuring the efficient transfer of sound from the air to the fluids in the cochlea. Otherwise, if direct sound were impinged to the oval window it would be reflected back. This is because the oval ear has different resistance to movement than air (difference in acoustic impedance). The middle ear in this case functions as an acoustic impedance-matching device that improves sound transmission and minimizes reflection. By having different effective areas of the eardrum and the oval window and the lever action of the ossicle it is possible to achieve effective transmission.

2) Reducing the transmission of bone-conducted sound to the cochlea. If sounds created by skull vibrations were transmitted strongly to the cochlea, they would appear loud and would mask other external sounds. In this case, skull vibrations produce all of the ossicles to vibrate as well. There is little motion from the stapes relative to the oval window and thus, little bone-conducted sound is hardly transmitted to the cochlea.

The ossicle also serve in reducing the transmission of loud sounds to the cochlea. When the ear is exposed to an intense sound, the muscles attached to the ossicles contract and temporarily “break” the ossicle chain. This is know as the middle ear reflex. This reflex helps prevent damage in the cochlea, but only works for lower frequencies. Also, the reflex is too slow to protect against impulsive sounds like gunshots. However, two other functions have been suggested for the reflex. First, is the reduction of the audibility of self-generated sounds and two, is the reduction of the masking of middle and high frequencies by lower ones.

The Inner Ear and the Basilar Membrane

The cochlea is a spiral shaped structure located in the inner ear. It is filled with almost incompressible fluids, and has bony rigid walls. It is divided by two membranes: Reissner's membrane and the basilar membrane (BM). The start of the cochlea, where the oval window is situated, is called the base, while the tip is know as the apex. The apex has a small opening, the helicotrema, between the BM and the walls of the cochlea that connect the two outer chambers of the cochlea, the scala vestibuli and the scala tympani. The inward movement of the oval window will produce in a corresponding outward movement in the round window.

Motion in the oval window, produced by the stapes, will create a pressure difference across the BM causing it to move. The pressure difference and the pattern of motion on the BM take time to develop and vary along its length. Sinusoidal stimulation to the BM will take the form of a traveling wave that moves from the base to the apex of the BM.

The varying mechanical properties of the BM will affect the response to sounds of different frequencies. The base is narrow and stiff, while the apex is wider and much less stiff. As a result, the peak position pattern of vibration will differ according to the frequency of the stimuli.

High frequency sounds will produce maximum displacement near the base of the BM, while low frequency sounds will produce a maximum near the apex. It can be observed then, that the cochlea is functioning in a similar manner as a Fourier analyzer. The frequency that produces a maximum response on a particular point of the BM is called the characteristic frequency (CF). Each point in the BM will vibrate with the same frequency as a response to a sinusoidal stimuli.

A method to describe the "sharpness" of tuning of the pattern of BM vibration is to describe the frequency resolution at each point in the BM as that of a bandpass filter with a center frequency corresponding to the CF. The slopes of the filter can be measured as 10 dB/octave. Bandwidth in the BM is not constant and increases in proportion with the CF.

However, regardless of the measuring technique used to determine the tuning sharpness of the BM, it must be noted that the sharpness critically depends on the physiological condition of the subject; thus, the better condition, the sharper the tuning. A healthy ear will posses a high sensitivity to a limited range of frequencies, and will require higher sound intensities to produce responses as the signal frequency is moved outside that range.

It has been shown that the BM vibration is nonlinear and that the magnitude of the response does not grow in direct proportion with the magnitude of the input. For low input sound levels (below 20 dB SPL) the CF function is quasi-linear and approaches linearity at high input sound levels (above 90 dB SPL), but at mid range a shallow slope is shown. This slope indicates a compressive nonlinearity in which a large range of input sound levels is compressed into a smaller range of responses on the BM.

A sinusoidal stimulation will produce a pattern of maximum displacement whose position depends on the frequency of the sound. If for example two tones are presented simultaneously, the BM will present two different patterns of vibration. Each with its own maximum at the place in which the BM is being excited. In this light, the BM functions as a Fourier analyzer by decomposing the complex sound into its individual components. However, when the two frequencies are close together, the BM vibration patterns will interact so that some points in the BM respond to both tones. This interference will produce a (non-sinusoidal) complex response and no distinct maximum will be show. Instead, a single broadband maximum will appear and as a result, the BM has failed in resolving the individual frequency components.

The Transduction Process and the Hair Cell

The organ of Corti possesses hair cells that lie between the BM and the tectorial membrane. These hair cells are divided into two groups by the tunnel of Corti. Hair cells close to the outside of the arch are called outer hair cells (hairs- stereocilia). Those on the other side of the arch are called the inner hair cells. The tectorial membrane lies above the stereocilia and actually makes contact with them. Because the tectorial membrane is hinged, as the BM moves, a shearing motion will be created between the BM and this membrane. This will result in a displacement of the stereocilia.

The inner hair cells act as transducers of mechanical motion to neural activity. When stereocilia are deflected they will open a "transduction channel" that allows potassium ions to flow into the hair cell causing a voltage difference between the inside and outside of the hair cell. This results in the release of neurotransmitter and initiates the action potential in the neurons of the auditory nerve. The great majority of afferent neurons carrying information from the cochlea to the higher levels of the auditory system connect to the inner hair cells.

Thus, most of the information about sounds is conveyed via the inner hair cells.
Outer hair cells are responsible for the sharp tuning and sensitivity of the cochlea.

Otoacoustic Emissions

Otoacoustic emissions refer to sounds emitted by the ear itself. The ear in response of a click will emit at first the reflected sound, but following a certain delay it will emit a sound. This process serves as proof of cochlear activity. Each ear will have its own characteristic response to stimuli. Responses tend to be stronger for frequencies between 500 – 2500 Hz. These emissions can be measured for brief tone bursts, clicks, and pure tones. Usually, otoacoustic emissions are observed in healthy ears. Any moderate pathology or exposition to drugs that affect the cochlea will show no detectable emissions.

If presented with two tones with different frequencies, a combination tone echo will be detected.

The hearing of sounds in the absence of external stimuli is called tinnitus. It is thought that this phenomenon is caused by abnormal activity in the auditory system, mainly the cochlea. Sounds emitted from the ear in the absence of input is called “spontaneous otoacoustic emissions”. These sounds indicate the presence of a source of energy within the cochlea capable of generating sounds.

Neural Response in the Auditory Nerve

Studies of auditory nerve activity show that the nerve fibers present spontaneous firing in the absence of stimulation. Also, these fibers show frequency selectivity. Finally, neural spikes occur at particular phase stimulation of the waveform.

Spontaneous Firing Rates and Thresholds

Auditory fibers can be classified into three groups depending on the spontaneous firing rates: High spontaneous firing rates(18-250 spikes/sec), medium firing rates (0.5-18 spikes/sec), and low firing rates (<0.5 spikes/sec). The spontaneous rates correlate with the position and the size of the synapse on the inner hair cells. For example, high firing rates have large synapses and are located on the side of the inner hair cells facing the outer hair cells.

The threshold of a neuron is the lowest level at which a change in response of the neuron can be measure. In this case, higher rates have lower thresholds and vice versa.

Tuning Curves and Isorate Contours

Tuning curves help illustrate the frequency selectivity of a single nerve fiber. This type of curves show the fiber's thresholds as a function of frequency. A single nerve fiber derives its output from a particular part in the BM. CF's are distributed in the auditory nerve such that high CF's are located in the periphery of the nerve bundle, while a decrease in CF means a change of location towards the center. Meaning that the place representation of frequency in the BM is preserved as place representation in the auditory nerve. Also, the sharpness of tuning of the BM is the same as the sharpness of tuning of single neurons in the auditory nerve.

Isorate contours describes the intensity of a sinusoidal stimulation required to produce a predetermined firing rate as a function of frequency.

Rate versus Level Functions

This schematic shows the rate of single neural discharges as a function of stimulus levels. Neurons respond to increasing sound levels by increasing their firing rate. After a certain sound level the neuron can no longer respond to increases of sound level with increases in firing rates. At this point, the neuron is said to be saturated. The range of sound levels between saturation levels is called the dynamic range.

Neural Excitation Patterns

Excitation patterns are the representation of the effective amount of excitation produced by a stimulus as a function of the CF. It is plotted as effective level (dB) against CF. Excitation patterns can be considered as an internal representation of the spectrum of the stimulus.

Phase Locking

The temporal pattern of neural spikes carries information about the stimulus. These nerve spikes tend to be synchronized or phase locked to the stimulating waveform (pure tone). While nerve fibers do not necessarily fire on every cycle of the stimulus, but when they do, they occur each time at the same phase of the waveform.

Neural firing does not occur at regular intervals. However, the information about the period of the stimulating waveform is carried unambiguously in the temporal pattern of firing of single neurons. Also, the distribution of time intervals between spikes is closely related on the frequency of the stimulating waveform. If the first peak of an interspike histogram does not occur at the period corresponding to the frequency of the stimulus, it is said that the neuron is in refractory period. Meaning that a neuron can not fire a spike within a determined period of time.

Phase locking does not occur over the entire audible frequency range, and it is said that the upper limit lies around 4-5 kHz. This limit is determined by the precision in which the initiation of a nerve impulse is linked to the phase of a waveform.

Two-tone Suppression

Auditory nerve response studies have shown that the activity of a single fiber in response to a single tone can be suppressed by the presence of a second tone. When the frequency and intensity of the second tone within the excitation area bounded by a tuning curve, it usually produces an increase in firing rate. However, if it falls outside of that area, the response to the first tone is reduced or suppressed.

For non-harmonically related tones, it has been found that neural discharges maybe phase locked to one tone, or to the other, or to both tones simultaneously. The dominant tone appears to “capture” the response of the neuron. It is likely that this “capture effect” underlies the masking of one sound by another.

Phase Locking to Complex Sounds

Studies in auditory nerve responses for stimuli pairs with simple frequency ratios have shown that the effective stimulating waveform can be approximated by addition of the component sinusoids. Although the amplitude and phase relations cannot be taken form the stimulus parameters.

Other similar studies show that, for complex tone stimuli, neural activity is phase locked to the overall repetition rate of the stimulus. In other words, the response is equal to the absent fundamental frequency. This temporal coding may be responsible for perceiving the pitch of complex tones.

Neural Responses at Higher Levels in the Auditory System

It is likely that the cortex is concerned with analyzing complex aspects of the stimuli than frequency or intensity. Many cortical neurons will not respond to pure tones, and those that do respond have different tuning properties from those found in the primary auditory neurons.

Some neurons have different tuning properties in the narrow, broad, and multi-range of preferred frequencies. This suggests a hierarchical organization, with several narrow range units converging into a single multi-range unit. The cortex is also organized so that multiple “maps”, concerned with different stimulus features, exist. For example, some neurons are concerned with frequency changes, while others are particularly sensitive to sound localization in space.