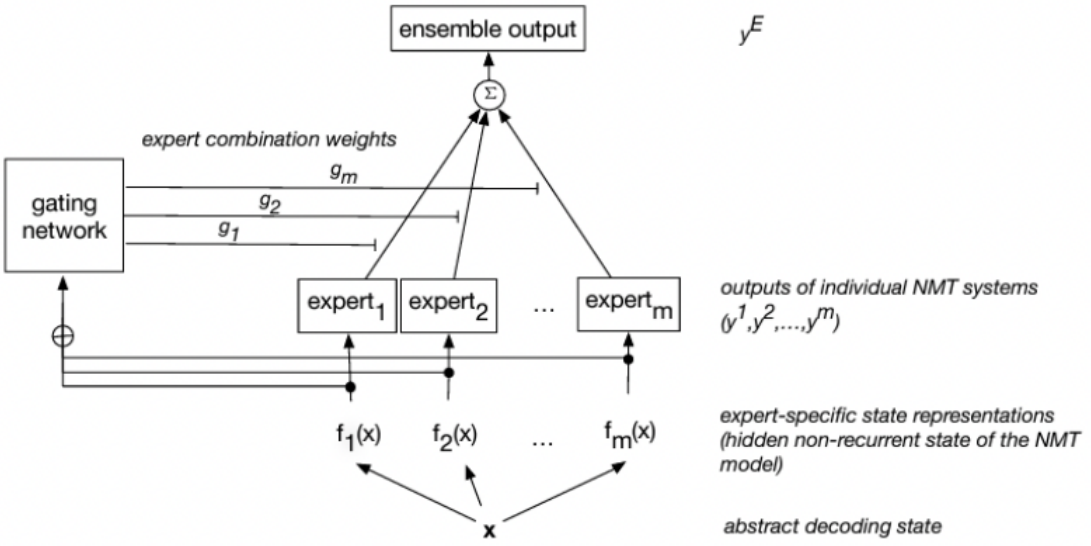
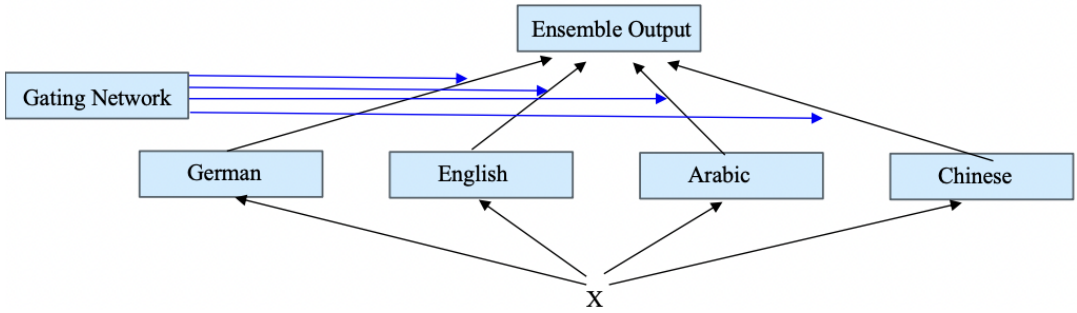
CAS CS 505
Lec 21

Conclusions on GPT; Audio Processing for Deep Learning

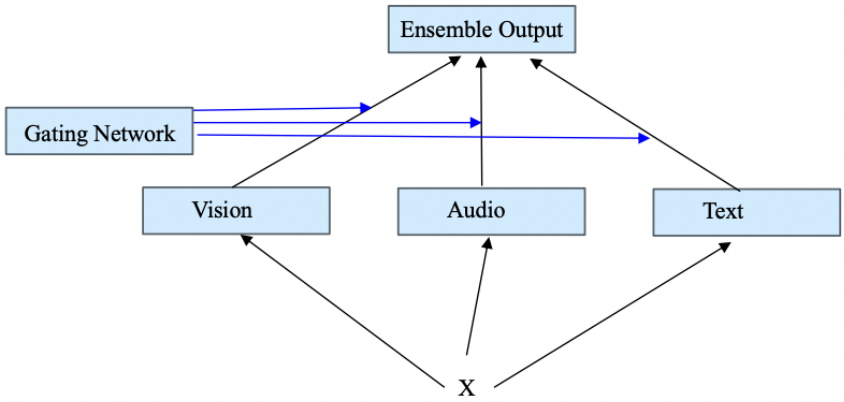
1. GPT-3 (2020)
    a. 175B parameters,
    b. 96 layers
    c. 96 masked-attention heads in each layer; used a more efficient alternation of dense and sparse attention patterns
    d. Input width: 2048 tokens
    e. Used softmax/temperature and Top-p sampling
    f. Training for GPT-3 was a scaled-up version of GPT-2, including a larger text corpus:
        i. Common Crawl of Web
        ii. Entire English Wikipedia
        iii. WebText2 (continuation of WebText from GPT-2)
        iv. Lots of miscellaneous books, technical manuals, encyclopedias, etc. etc. etc.
    g. GPT-3 also used an AI technique called Mixture of Experts (MoE):
        i. Multiple pretrained networks are used to divide the problem space into separate task regions;
        ii. Expert results are aggregated by a gating network, which combines one or more outputs into an ensemble output:



h. Examples:
    i. Multi-language models contain pretrained experts for distinct languages:



    ii. Multi-modal models contains distinct sub-networks for vision, audio, and text:



i. In the transformer architecture, the experts can also communicate with each other; typically only some layers are built from MoE:

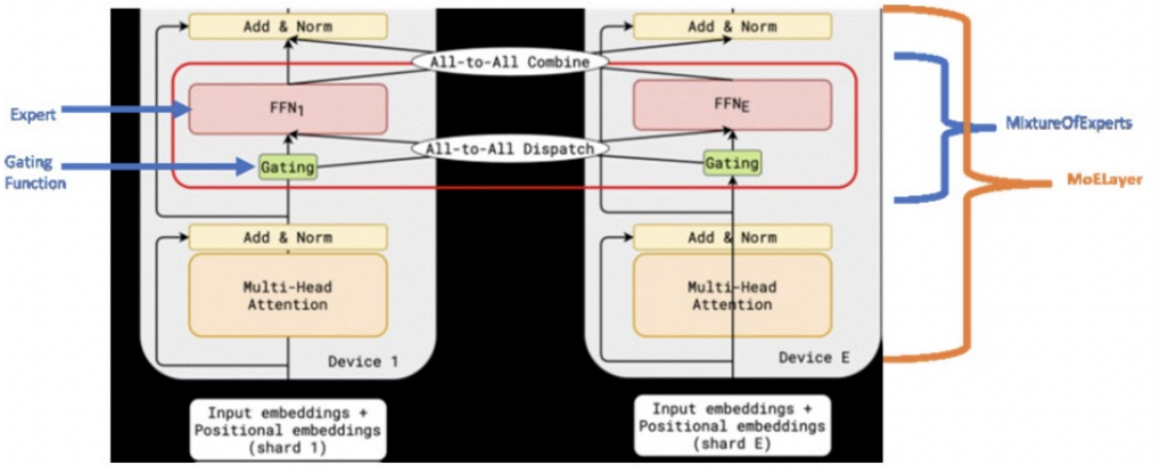
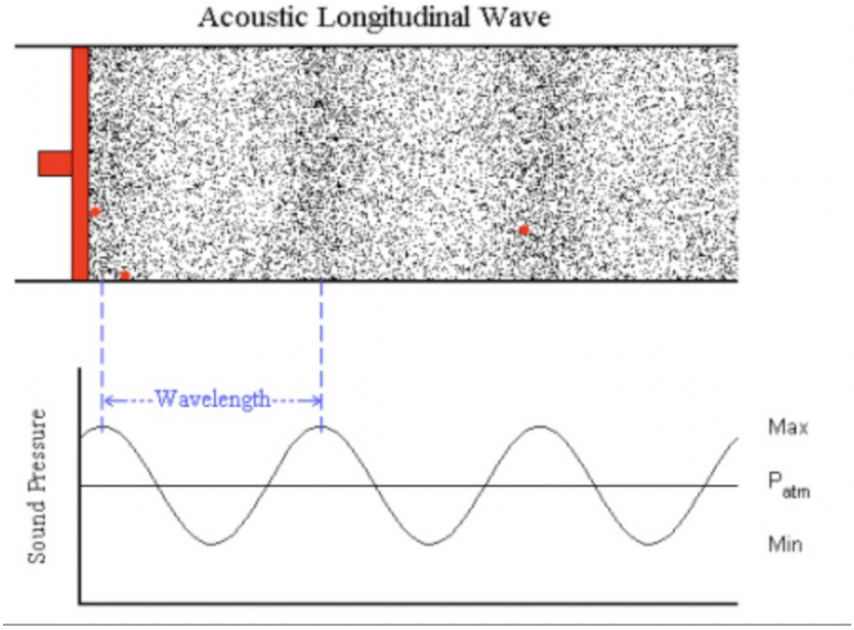


Figure 1: MoE components

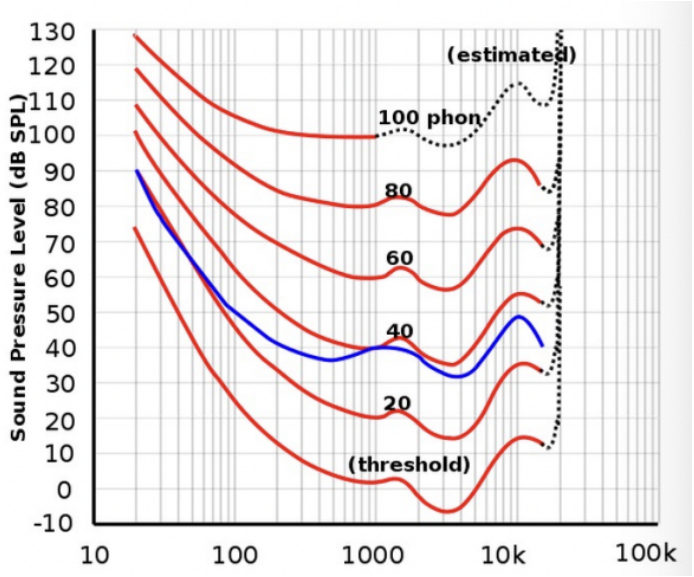
j. GPT-4 (March 2023)
    i. Multimodal
    ii. 1.8 T parameters
    iii. 120 layers
    iv. Input size of largest version is 32 K tokens
    v. Handles 26 languages
    vi. MoE with 16 experts, each with 111 B parameters

2. Audio Computing: Physical Basis of Sound
    a. Sound is produced by vibrating objects which produce pressure waves in air, traveling at 343.21 meters/sec (768 mph, or a mile in 4.69 seconds), which are sensed by the ear and interpreted by the brain.
    b. These waves are longitudinal waves (the motion is along the direction of travel), as opposed to transverse waves (motion is at right angles to the direction).
3. Physical Basis of Sound
    a. Time Domain Representation: If we record the atmospheric pressure over time, we get a curve with amplitude in pounds per square inch (psi) on the y axis and time in seconds over the x axis in the shape of the sin (or cos) function:



4. Properties of (Pure) Sine Waves
    a. Wavelength (λ): Distance between peaks of a wave (affects pitch -- high or low sounds); measured in meters.
    b. Period (p): The time between peaks, measured in seconds.
    c. Amplitude (A): magnitude of the wave above the midpoint (x axis).
    d. Frequency (f): the number of times a wave occurs in a second. Measured in cycles per second or Hertz (Hz) or KiloHertz (kHz).

5. Sound Wave Properties: Frequency
    a. Frequency is an absolute measure, and is strongly related to but not absolutely identical to the notion of pitch.
    b. Pitch = perceived frequency of a sound
    c. Pitch is subjective, and inaccurate at extremes of frequency or amplitude, and not measured precisely by the ear throughout the range of hearing. However, it is generally perceived in a log scale.
        i. Human: 20Hz – 20 kHz
        ii. Dogs: 67 Hz – 44 kHz
        iii. Cats: 55 Hz – 79 kHz
        iv. Bats: 1Hz – 200 kHz
    d. Sensitivity of human ear at various frequencies; curves represent impressions of equal loudness at various frequencies:



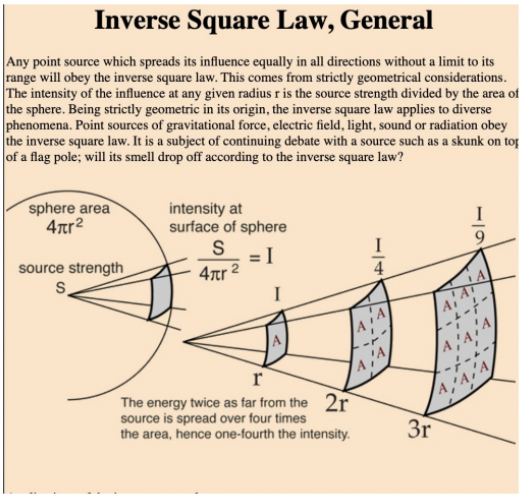
6. Sound Wave Properties: Intensity and Loudness

Sound Intensity

Notice that sound waves carry energy. We define the intensity I as the rate at which energy E flows through a unit area A perpendicular to the direction of travel of the wave.

Intensity = Power / Area

$I = P / A = E / (At)$



a.

b. Thus, Intensity is proportional to the square of the amplitude:
    i. I = c * A2 (c is a constant depending on properties of medium)
c. Since ear drums tend to be similar in area, often Intensity and Power are used as equivalent terms.

Loudness is not simply sound intensity!

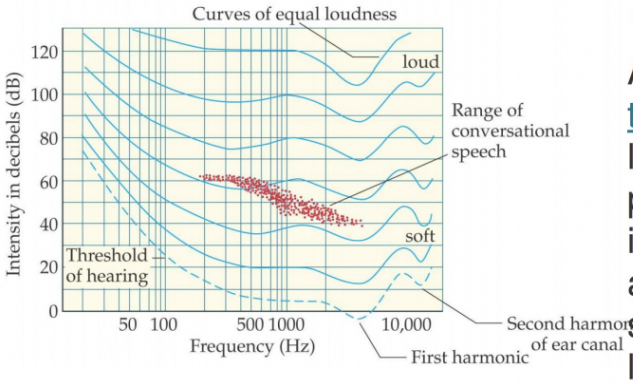
Loudness of a sound is measured by the logarithm of the intensity.

log here is to base 10

The threshold of hearing is at an intensity of 10^{-12} W/m².

Sound intensity level is defined by $\beta = (10\text{dB})\log\frac{I}{I_0}$

dB are decibels



A general "rule of thumb" for loudness is that the power must be increased by about a factor of ten to sound twice as loud.

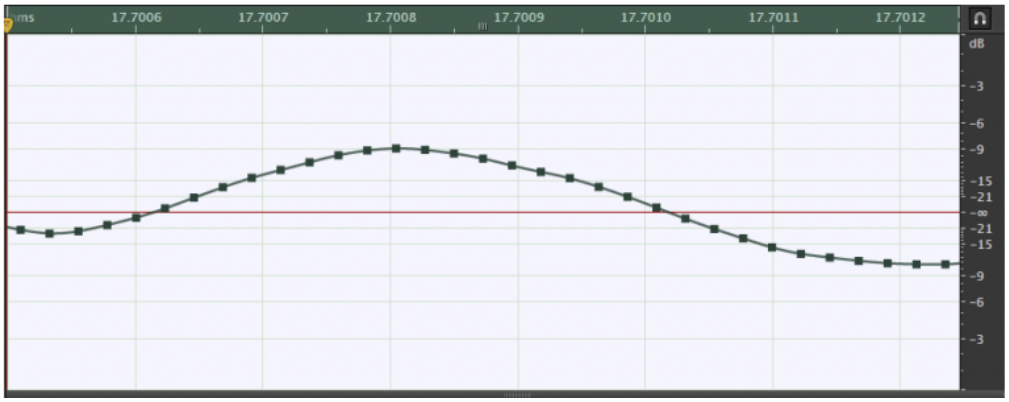
32

13

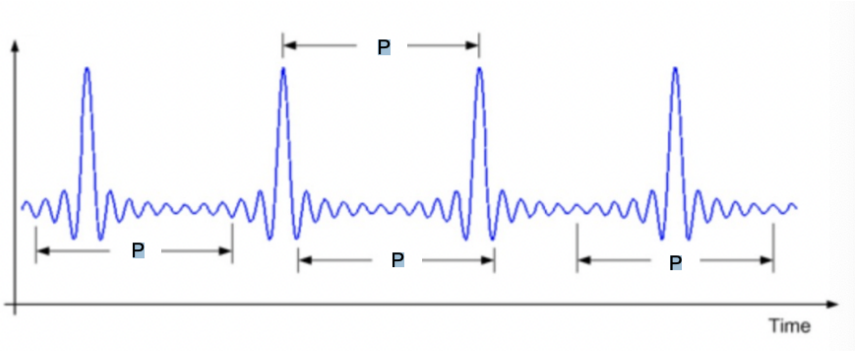
d.

7. Digital Audio: Analog vs digital signals
    a. An analog signal continually fluctuates in up and down in the real domain (time is a real number and amplitude is a real number).
    b. A digital signal has a discrete number of amplitudes over a discrete number of time steps, T = 0, 1, 2, ….
    c. Typically, represented as samples taken at a regular sample rate, e.g.,
    d. 16-bit integers 44,100 times a second.

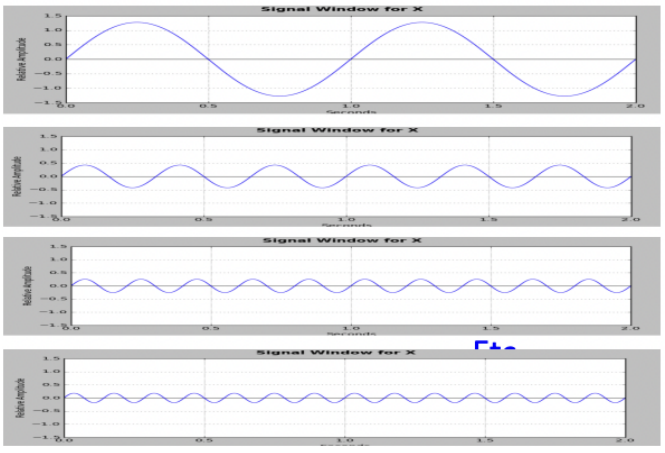
e.



8. Jean-Bastiste Fourier
    a. Jean-Baptiste Joseph Fourier (1768 – 1830) was a French mathematician and scientist who studied heat transfer in metals (among other things--for example he was the first to explain the Greenhouse Effect). In this investigation he realized the following remarkable fact:
        i. Any periodic function with period P can be constructed by adding together a sequence (possibly infinite) of simple sine waves of various amplitudes and phases, with frequencies that are integer multiples of the fundamental frequency f = 1/P.
        ii. A periodic function is one which repeats its values at a period P:

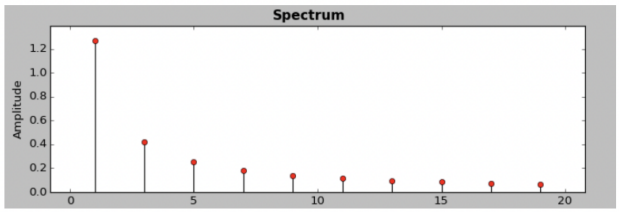


9. Fourier/Additive Synthesis
    a. Example of Fourier/Additive Synthesis
        i. A square wave of frequency f and amplitude 1.0 can be created from the following infinite Fourier Series:
            1. $(4/\pi)(\sin(2\pi f t) + (1/3) \sin(2\pi 3f t) + (1/5) \sin(2\pi 5f t) + \ldots.)$
            2. that is: [(1,1.27,0), (3,0.42,0), (5,0.25,0), (7,0.18,0), (9,0.14,0), …]
            3. shown here graphically (duration of 2 sec):

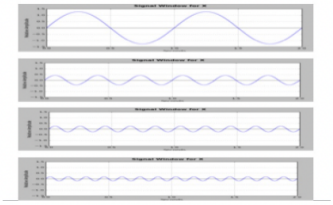


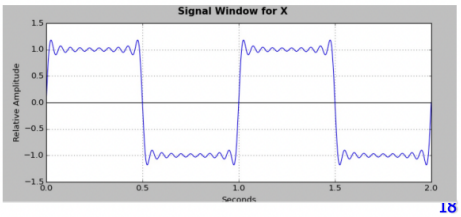
        ii. Spectrum: A graph of the amplitude and frequency of the components of a sinusoid is called a spectrum; here is a spectrum of a wave consisting of

the first ten terms of the Fourier Series for a square wave



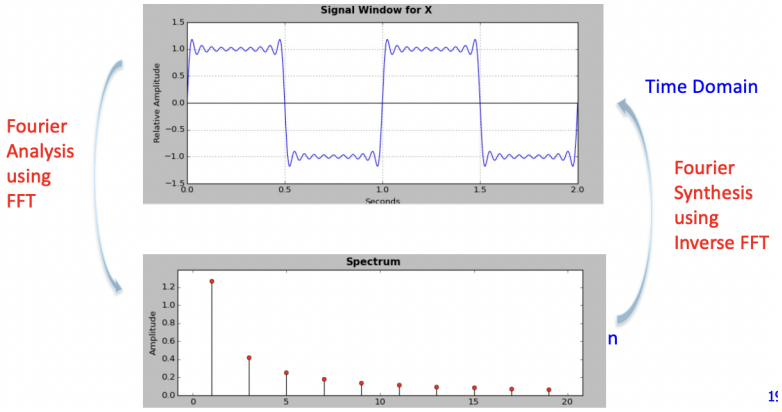
(freq, amp, phase)





18

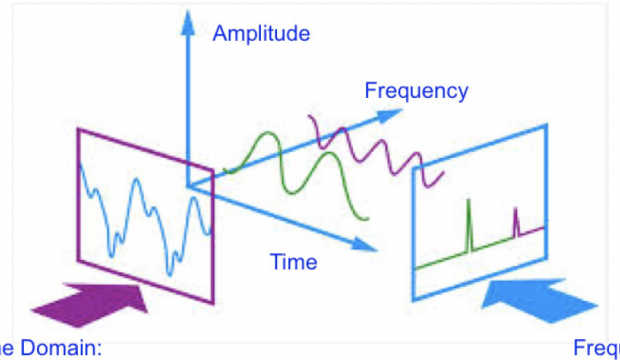
b. Fourier provided a way of looking at a signal in two equivalent ways, as a graph of amplitude of the signal vs time, the Time Domain, or as a graph of frequencies vs amplitudes (the Frequency Domain):



n

19

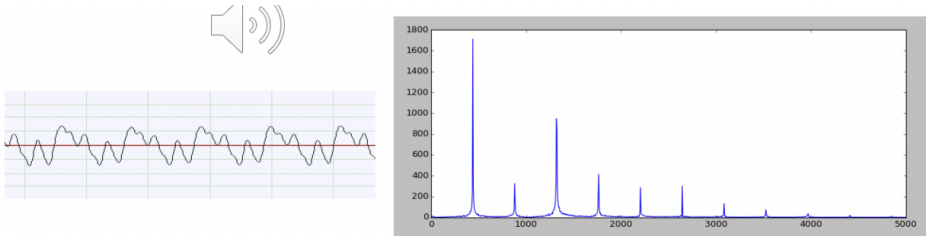
c. It can NOT be overemphasized how important these results are for understanding sound (which has three dimensions); these are two different, but completely equivalent ways of viewing the same phenomenon, and the pair of transforms are lossless and very efficient: O(N log N).



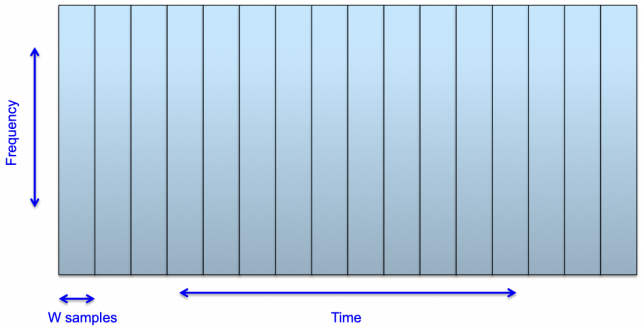
Time Domain:
Amplitude vs Time

Frequency Domain:
Amplitude vs Frequency

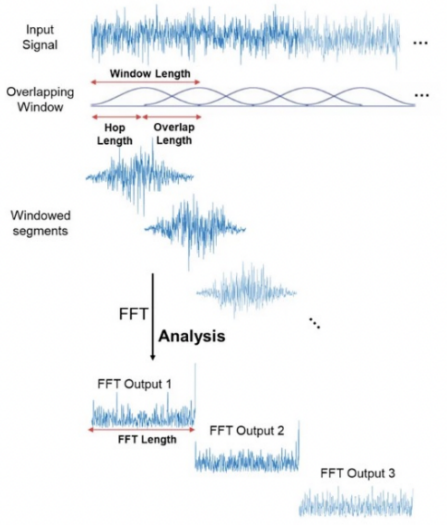
10. Harmonic Series
    a. Harmonic Series: A series of integer multiples of a particular lowest, fundamental frequency is called a Harmonic Series and each component is called a Harmonic or Overtone:
    b. Example (in Hz): 440, 880, 1320, 1760, 2200, 2640, 3080, 3520, 3960, ….
    c. A periodic signal can always be constructed from such a sequence, although in additive synthesis of musical signals we will not restrict ourselves only to such sequences. Real musical signals always have more complex spectra.
    d. Still, musical sounds are in large part based on such spectra; here is a clarinet playing A 440 Hz:



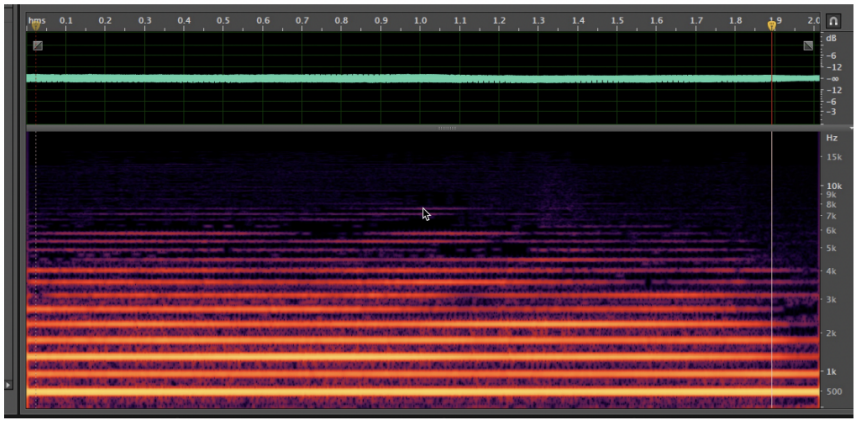
11. Digital Audio Fundamentals: The Discrete Fourier Transform
    a. Spectrogram:
    b. A spectrogram is a 2 D array of spectral data over time; hence it is a 3 D object over frequency, amplitude, and time:



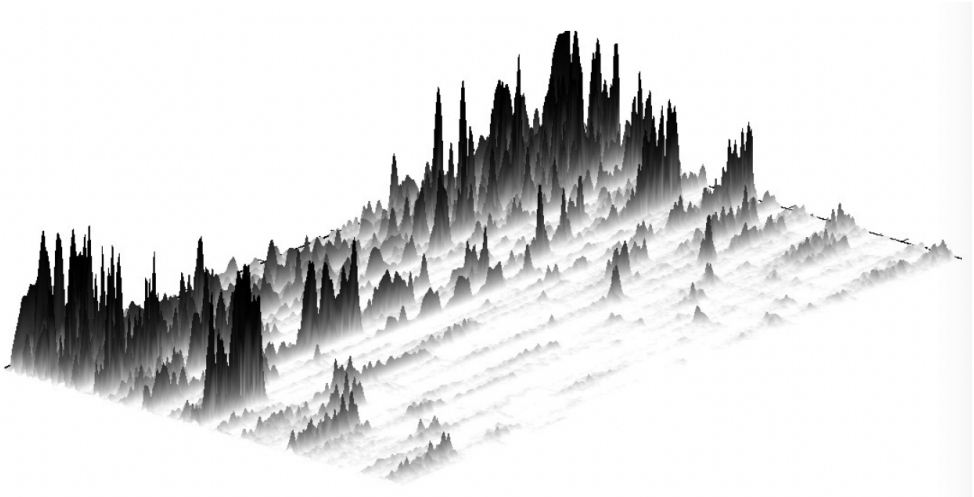
    c.



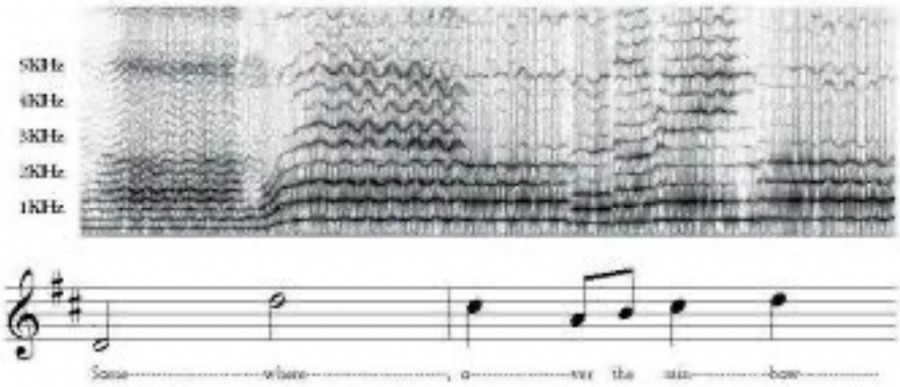
12. Fourier Analysis and Synthesis
    a. Spectrograms: Audio software such as Adobe Audition provides for spectrographic views of signals:



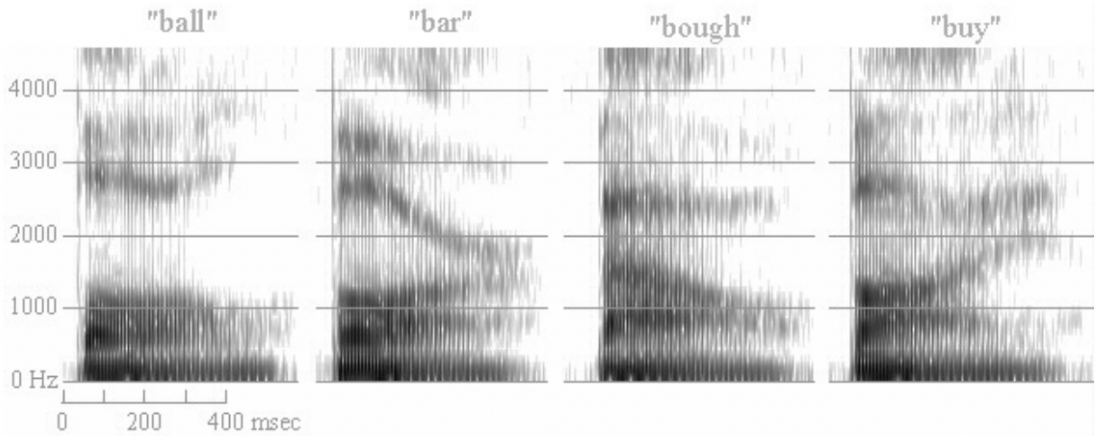
13. Digital Audio Fundamentals: The Discrete Fourier Transform
    a. Viewing 2D data can be done using faux-3D plots:



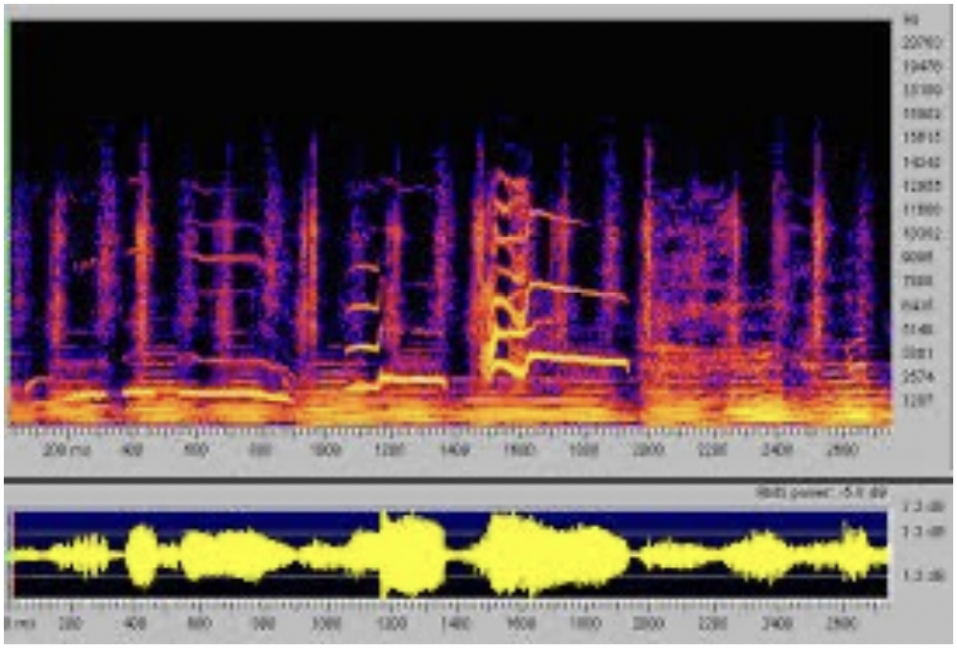
    b. But is more commonly done by “heat-maps” where the amplitude is indicated by greyscale or color:



14. Fourier Analysis and Synthesis
    a. But is more commonly done by “heat-maps” where the amplitude is indicated by greyscale or color:



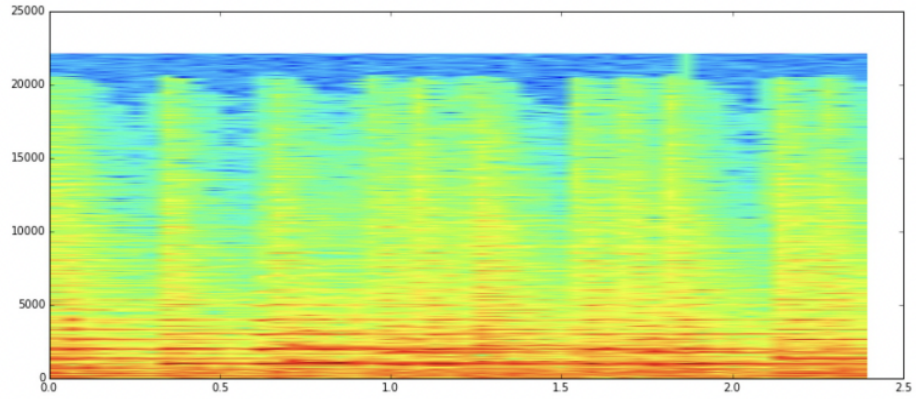
15. Digital Audio Fundamentals: The Discrete Fourier Transform
    a. But is more commonly done by “heat-maps” where the amplitude is indicated by greyscale or color:



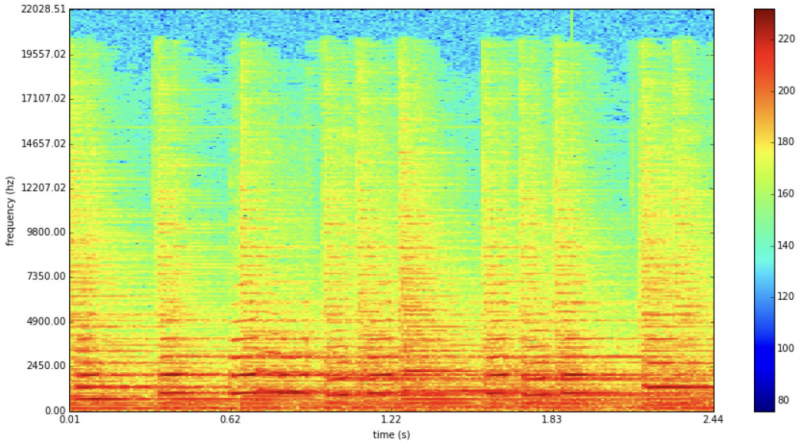
    b. In matplotlib, we have a library function to create spectrograms from a signal directly

```
specgram(x, NFFT=256, Fs=2, Fc=0, detrend=mlab.detrend_none,
         window=mlab.window_hanning, noverlap=128,
         cmap=None, xextent=None, pad_to=None, sides='default',
         scale_by_freq=None, mode='default', scale='default',
         **kwargs)
```

c. (spectrum, freqs, t, im) = plt.specgram(X,NFFT=2048, Fs=44100, noverlap=0)

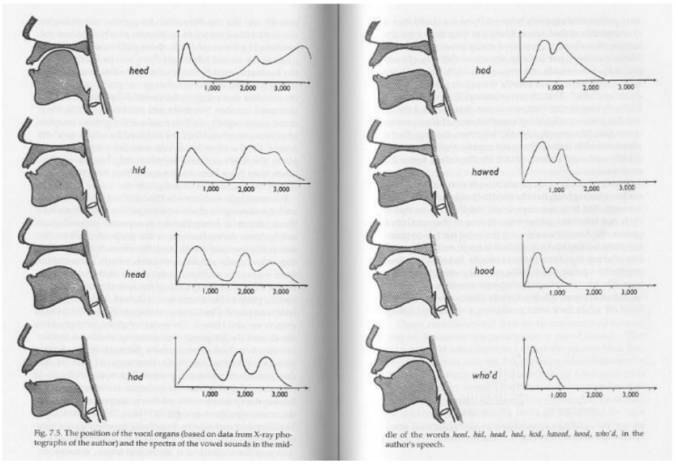


d. By adding a few bells and whistles, we can get log scale and proper axis measurements:



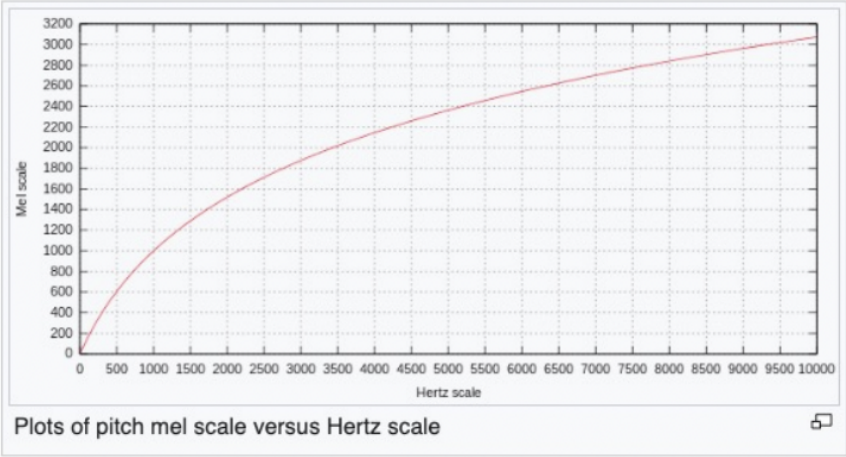
16. Musical Acoustics: Vocal Tract

a. Even if you do not play a musical instrument, you use spectra all the time to communicate! The reason you recognize different vowel sounds is that they have different spectra:



17. MEL Scale
    a. Humans perceive both loudness and pitch on a log scale;
    b. For pitch, the relationship between pitch f in Hz, and our human perception m (in Mels) this is called the Mel Scale:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

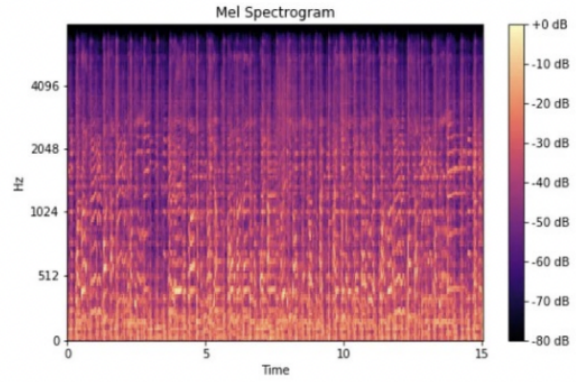


Plots of pitch mel scale versus Hertz scale

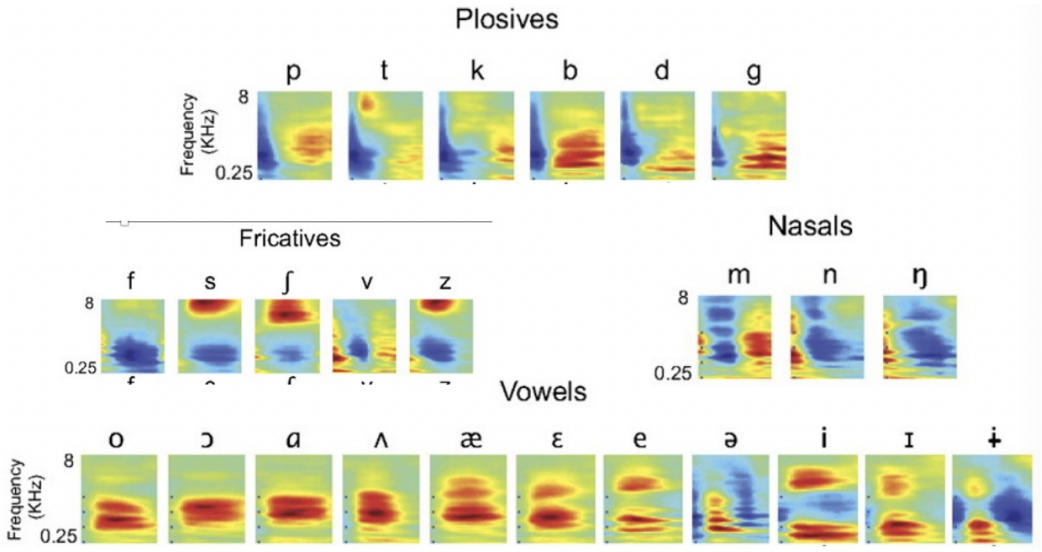
18. MEL Spectrogram
    a. Therefore, to capture the human experience of sound, we typically use a Mel Spectrogram, where
        i. Pitch is given in Mels
        ii. Loudness is given in Decibels:

```
mel_spect = librosa.feature.melspectrogram(y=y, sr=sr, n_fft=2048,
hop_length=1024)
mel_spect = librosa.power_to_db(spect, ref=np.max)

librosa.display.specshow(mel_spect, y_axis='mel', fmax=8000,
x_axis='time');
plt.title('Mel Spectrogram');
plt.colorbar(format='%+2.0f dB');
```



19. Human Vocal Signals
    a. Each phoneme in human language has a rather distinct spectrogram:



    b.