# Audio Features

CS498

# Today's lecture

- Audio Features
- How we hear sound
- How we represent sound
  - In the context of this class

# Why features?

- Features are a very important area
  - Bad features make problems unsolvable
  - Good features make problems trivial

- Learning how to pick features is the key
  - So is understanding what they mean

# A simple example

- Compare two numbers:

$$x, y = \{3, 3\} \qquad x, z = \{3, 100\}$$

# A simple example

- Compare two numbers:

$$\left\|x - y\right\| = 0 \qquad \left\|x - z\right\| = 97$$

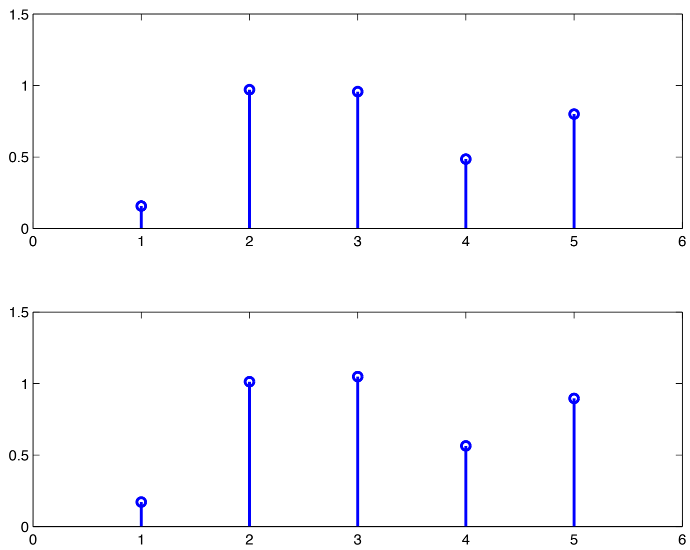
  - $x, y$ similar but $x, z$ not so much

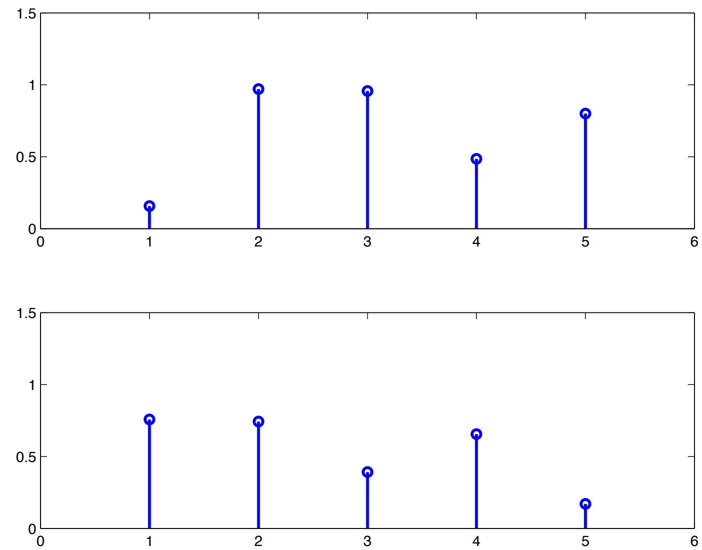
- Best way to represent a number is itself!

# Moving up a level

- Compare two vectors:

$$\mathbf{x}, \mathbf{y} \qquad\qquad \mathbf{x}, \mathbf{z}$$



# Moving up a level

- Compare two vectors:

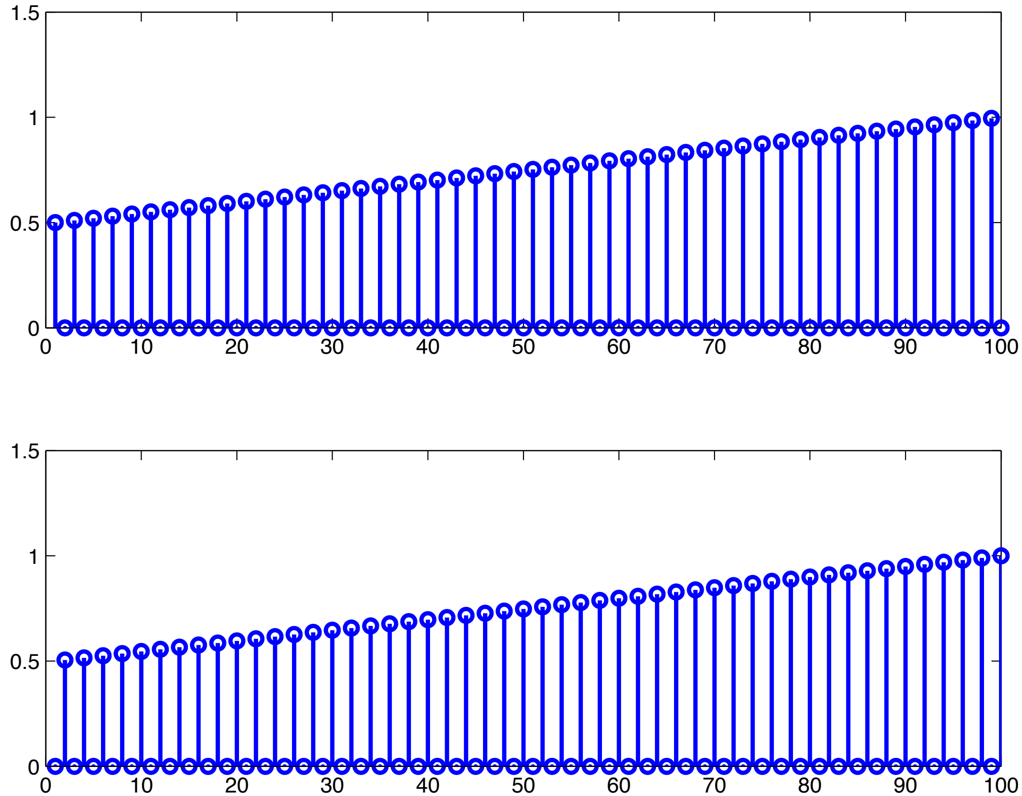
$$\angle \mathbf{x}, \mathbf{y} = 0.03 \text{ rad} \qquad \angle \mathbf{x}, \mathbf{z} = 0.7 \text{ rad}$$

$$\left\| \mathbf{x} - \mathbf{y} \right\| = 0.16 \qquad \left\| \mathbf{x} - \mathbf{z} \right\| = 1.07$$

  - Simply generalizing numbers concept

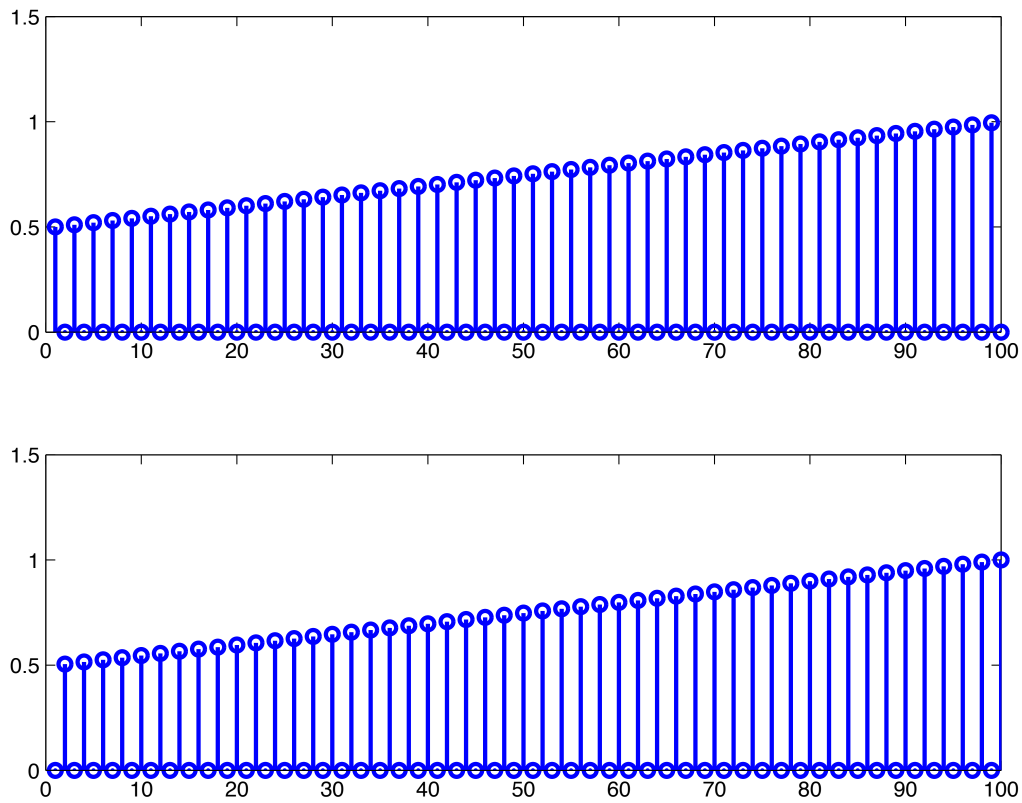
# Moving up again

- Compare two longer vectors:



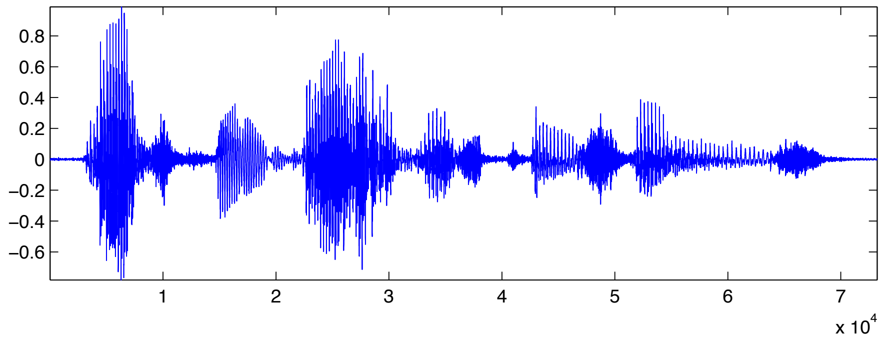
# Look similar but are not!

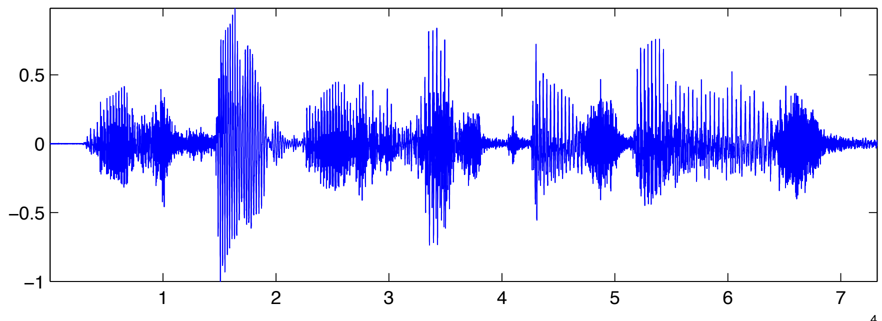
- Oops! $\angle \mathbf{x}, \mathbf{y} = 1.57 \text{ rad}, \quad \|\mathbf{x} - \mathbf{y}\| = 7.64$



# How about this?

- Are these two vectors the same?




  - Not if you look at their norm or angle …

# Data norms won't get you far!

- You need to articulate what matters
  - You need to know what matters
- Features are the means to do so
- Let's examine what matters to our ears
  - Our bodies sorta know best

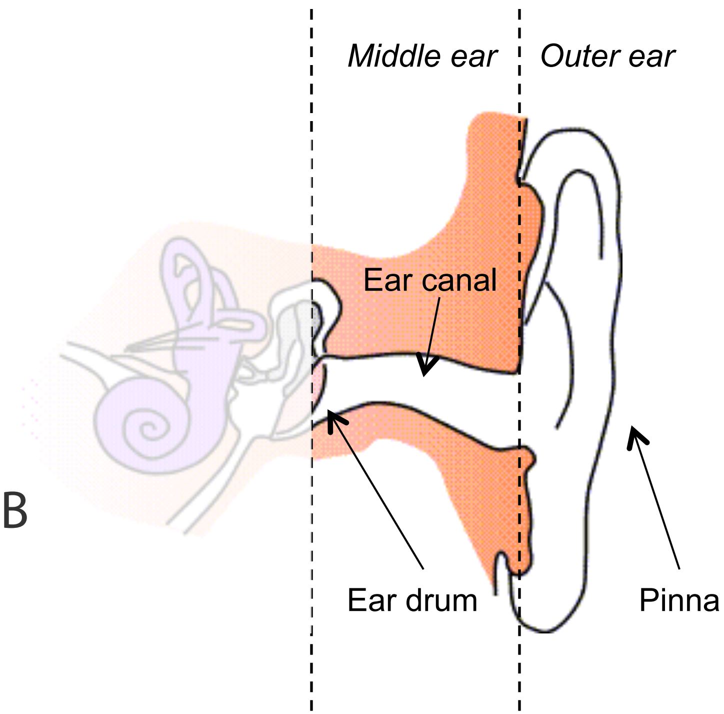
# Hearing

- Sounds and hearing
- Human hearing aspects
  - Physiology and psychology
- Lessons learned

# The hardware

(outer/middle ear)

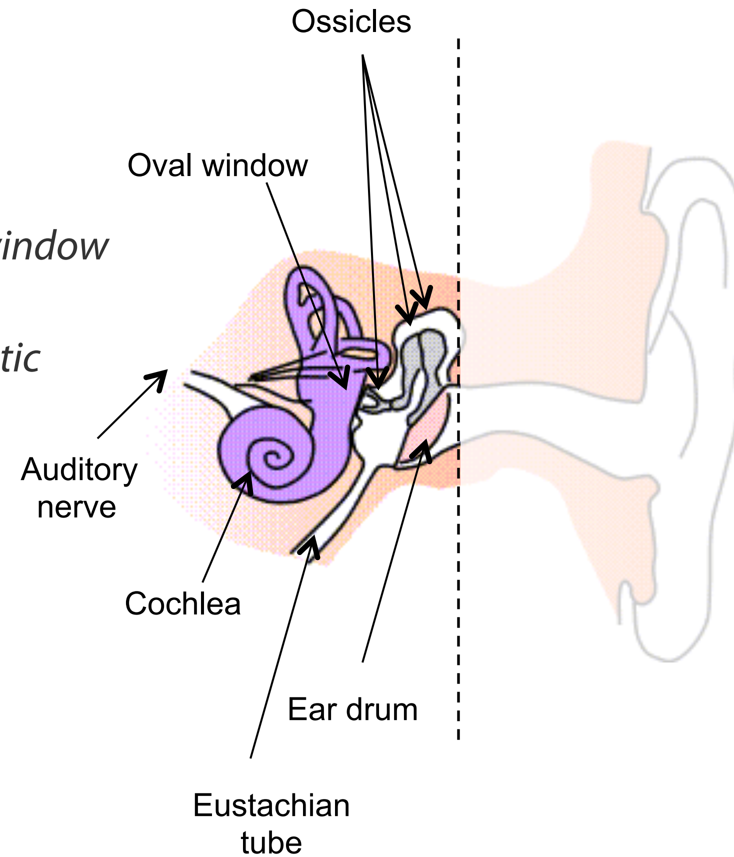
- The pinna (auricle)
  - Aids sound collection
  - Does directional filtering
  - Holds earrings, etc …
- The ear canal
  - About 25mm x 7mm
  - Amplifies sound at ~3kHz by ~10dB
  - Helps clarify a lot of sounds!
- Ear drum
  - End of middle ear, start of inner ear
  - Transmits sound as a vibration to the inner ear



# More hardware

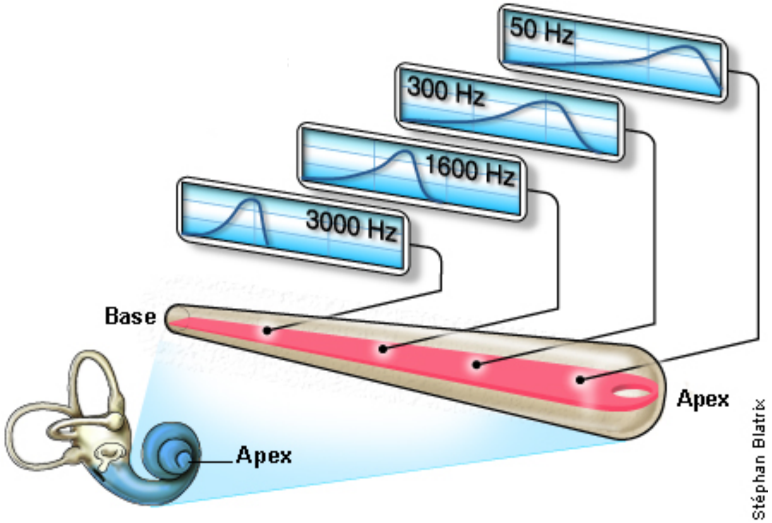
(inner ear)

- Ear drum (tympanum)
  - Excites the ossicles (ear bones)
- Ossicles
  - Malleus (hammer), incus (anvil), stapes (stirrup)
  - Transfers vibrations from ear drum to the *oval window*
  - Amplify sound by ~14dB (peak at ~1kHz)
  - Muscles connected to ossicles control the *acoustic reflex* (damping in presence of loud sounds)
- The oval window
  - Transfers vibrations to the *cochlea*
- Eustachian tube
  - Used for pressure equalization



# The cochlea

- The “A/D converter”
  - Translates oval window vibrations to a neural signal
  - Fluid filled with the *basilar membrane* in the middle
  - Each section of the basilar membrane resonates with a different sound frequency
  - Vibrations of the basilar membrane move sections of *hair cells* which send off neural signals to the brain
- The cochlea acts like the equalizer display in your stereo
  - Frequency domain decomposition
- Neural signals from the hair cells go to the auditory nerve

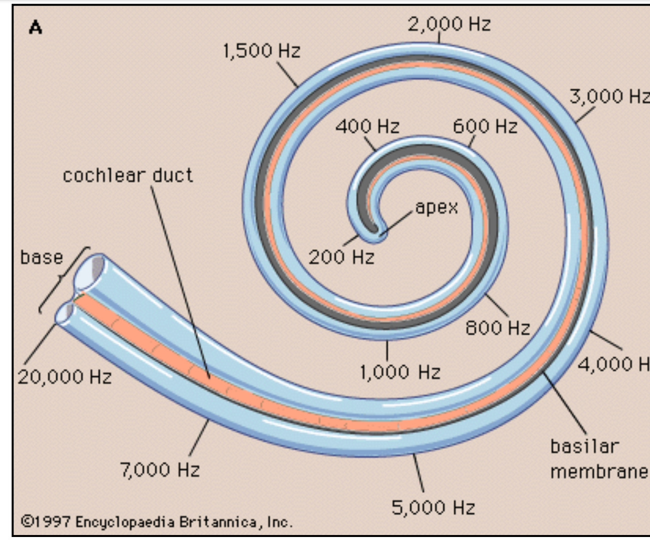



Microscope photograph of hair cells (yellow)

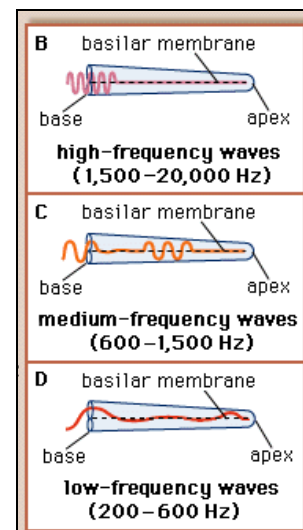
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# Masking & Critical bands

- When two different sounds excite the same section of the basilar membrane one is masked
- This is observed at the micro-level
  - E.g. two tones at 150Hz and 170Hz, if one tone is loud enough the other will be inaudible
  - A tone can also hide a noise band when loud enough
- There are 24 distinct bands throughout the cochlea
  - a.k.a critical bands
  - Simultaneous excitation on a band by multiple sources results in a single source percept
- There is also some temporal masking
  - Preceding sounds mask what's next
- This is a feature which is taken into advantage by a lot of audio compression
  - Throws away stuff you won't hear due to masking

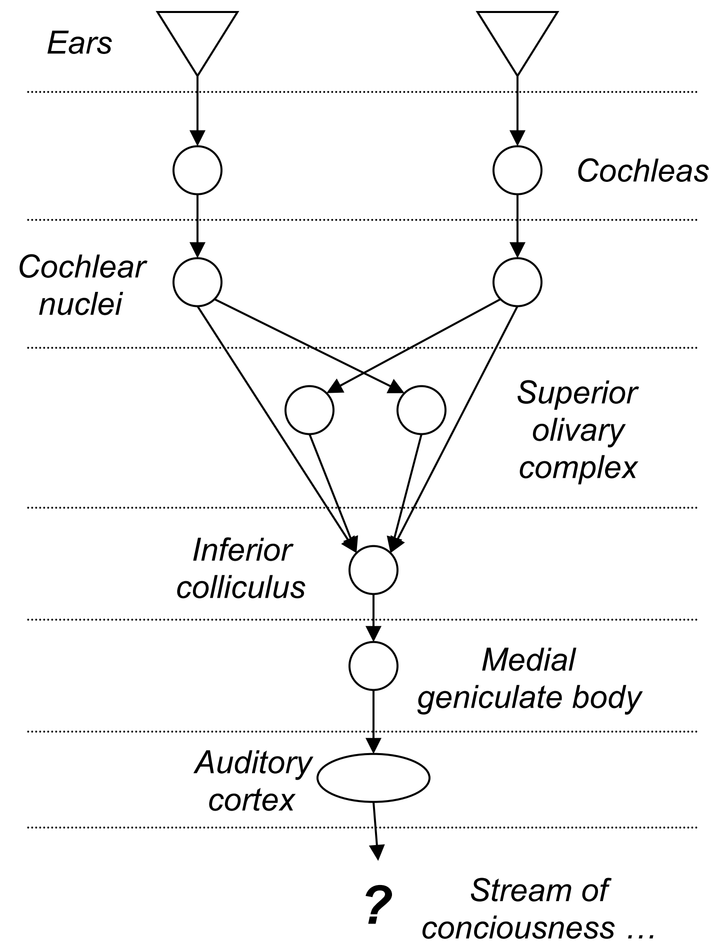



*Masking for close frequency tones vs distant tones*



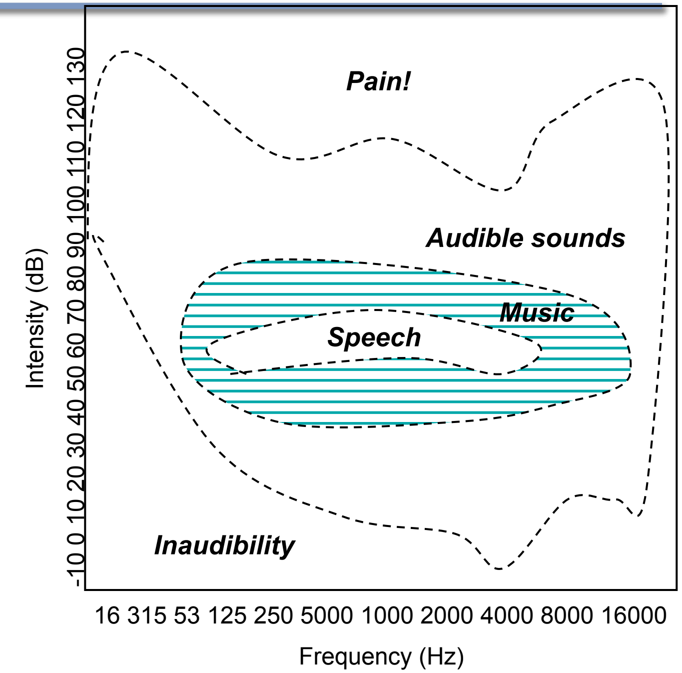
# The neural pathways

- A series of neural stops
- Cochlear nuclei
  - Prepping/distribution of neural data from cochlea
- Superior Olivary Complex
  - Coincidence detection across ear signals
  - Localization functions
- Inferior Colliculus
  - Last place where we have most original data
  - Probably initiates first auditory images in brain
- Medial Geniculate Body
  - Relays various sound features (frequency, intensity, etc) to the auditory cortex
- Auditory Cortex
  - Reasoning, recognition, identification, etc
  - High-level processing



# The limits of hearing

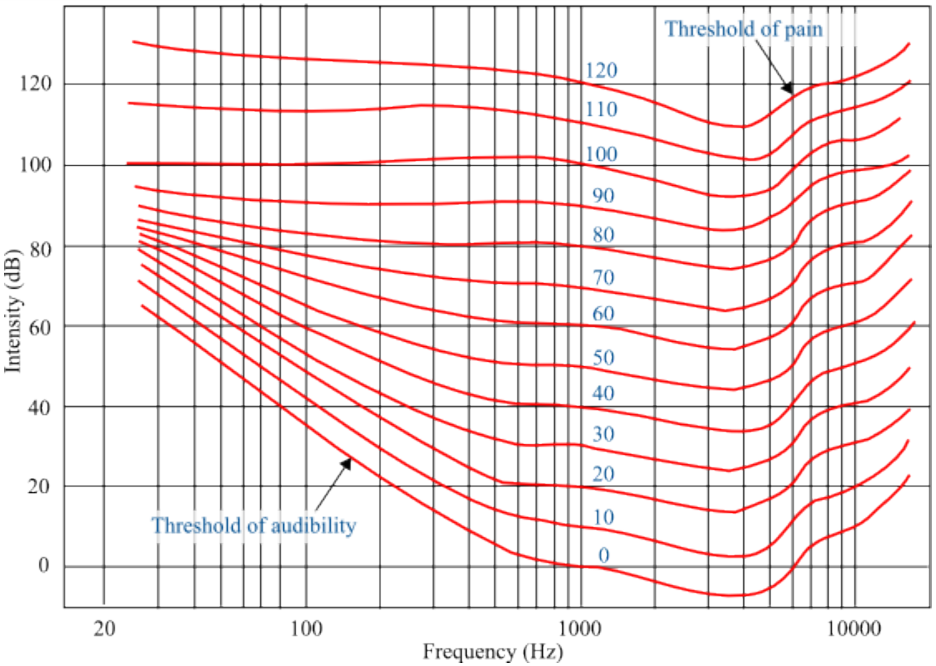
- Frequency
  - 20Hz to 20kHz (upper limit decreases with age/trauma)
  - Infrasound (< 20Hz) can be felt through skin, also as events
  - Ultrasound (> 20kHz) can be "emotionally" perceived (discomfort, nausea, etc)
- Loudness
  - Low limit is $2 \times 10^{-10}$ atm
  - 0dB SPL to 130dB SPL (but also frequency dependent)
    - A dynamic range of $3 \times 10^{6}$ to 1!
  - 130dB SPL threshold of pain
  - 194dB SPL is definition of a shock wave, sounds stops!

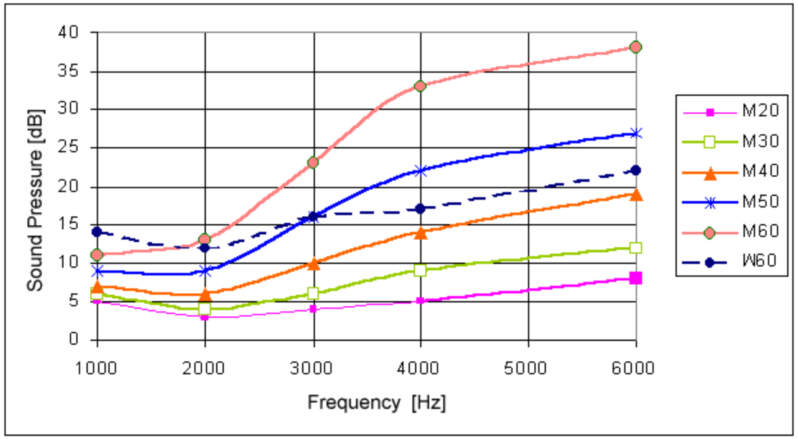



*Tones at various frequencies, how high can you hear?*

# Perception of loudness

- Loudness is subjective
  - Perceived loudness changes with frequency
  - Perception of “twice as loud” is not really that!
  - Ditto for equal loudness
- Fletcher-Munson curves
  - Equal loudness perception curves through frequencies
- Just noticeable difference is about 1dB SLP
- 1kHz to 5kHz are the loudest heard frequencies
  - What the ear canal and ossicles amplify!
- Low limit shifts up with age!

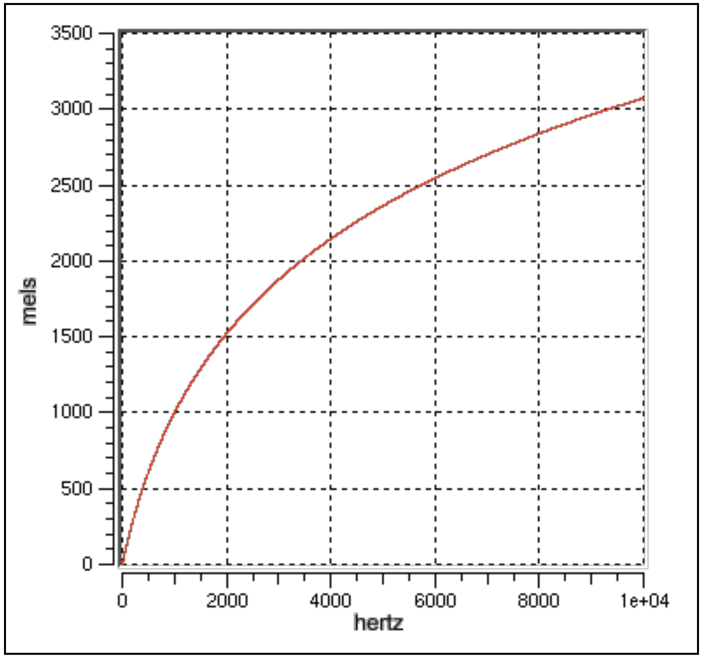



ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# Perception of pitch

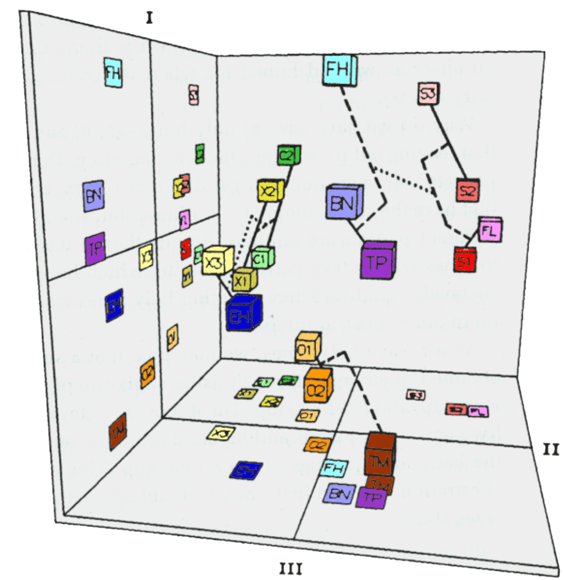
- Pitch is another subjective (and arbitrary) measure
- Perception of pitch doubling doesn’t imply doubling of Hz
  - Mel scale is the perceptual pitch scale
  - Twice as many Mels correspond to a perceived pitch doubling
- Musically useful range varies from 30Hz to 4kHz
- Just noticeable difference is about 0.5% of frequency
  - Varies with training though




*“Pitch is that attribute of auditory sensation in terms of which sounds may be ordered from low to high”*
- American National Standards Institute

# Perception of timbre

- Timbre is what distinguishes sounds outside of loudness & pitch
  - Another bogus ANSI description
- Timbre is dynamical and can have many facets which can often include pitch and loudness variations
  - E.g. music instrument identification is guided largely by intensity fluctuations through time
- There is not a coherent body of literature examining human timbre perception
  - But there is a huge bibliography on computational timbre perception!




*Gray's timbre space of musical instruments*

*Examples of successive timbre changes.  Loudness and pitch are constant*

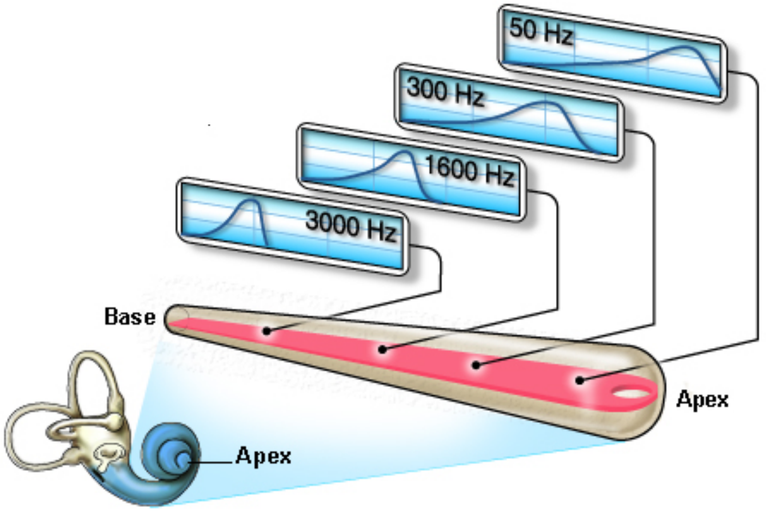
# So how to we use all that?

- All these processes are meaningful
  - They encapsulate statistics of sounds
  - They suggest features to use

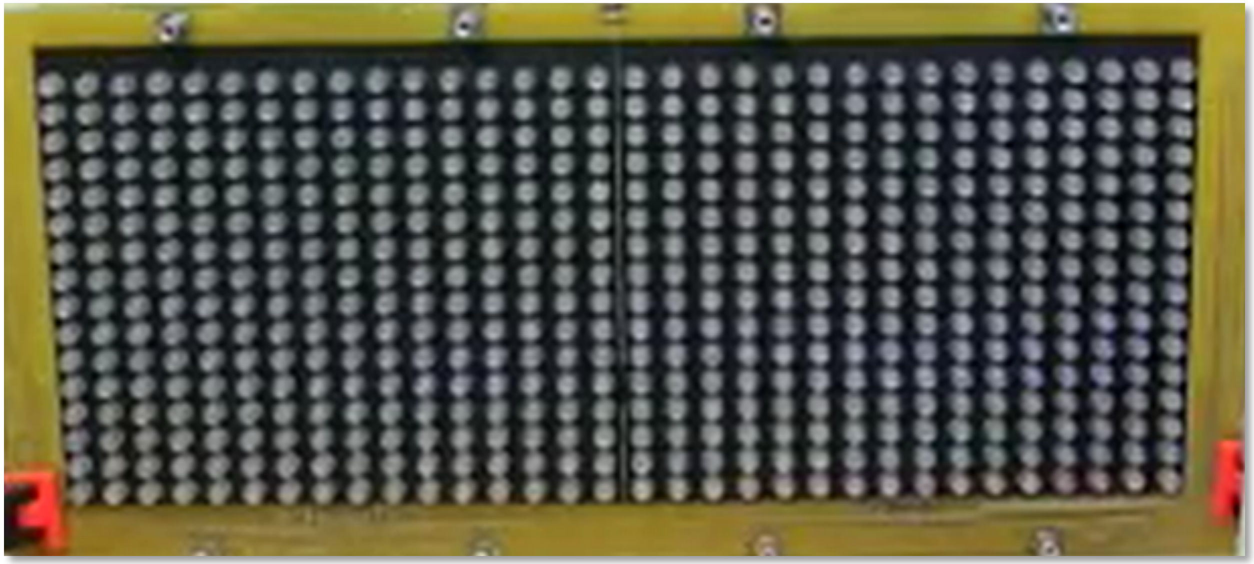
- To make machines that cater to our needs
  - We need to learn from our perception

# A lesson from the cochlea

- Sounds are not vectors
- Sounds are “frequency ensembles”
- That’s the “perceptual feature” we care about



# Like this!

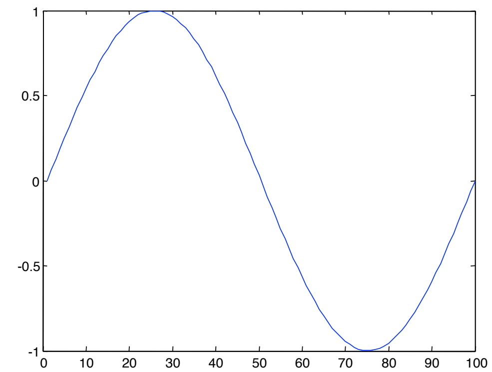


– But how do we get this?

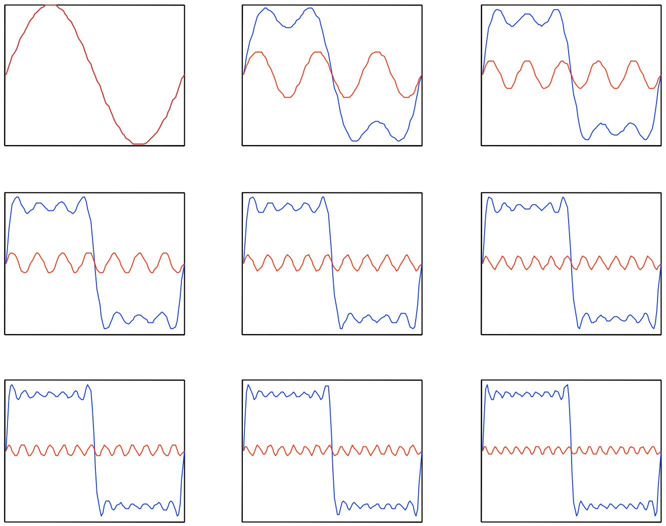
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# The “simplest” sound

- Sinusoids are special
  - Simplest waveform
  - An isolated frequency
- A sinusoid has three parameters
  - Frequency, amplitude & phase
    - $s(t) = a(t)\ \sin(\ f\ t + \varphi)$
- This simplicity makes sinusoids an excellent building block for most of time series

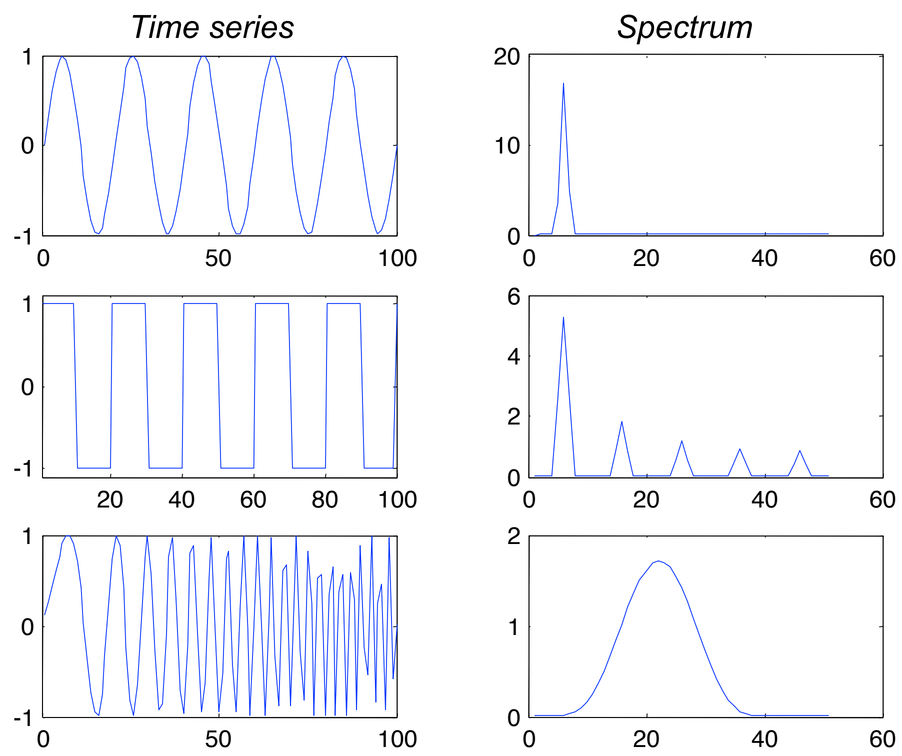



*Making a square wave with sines*



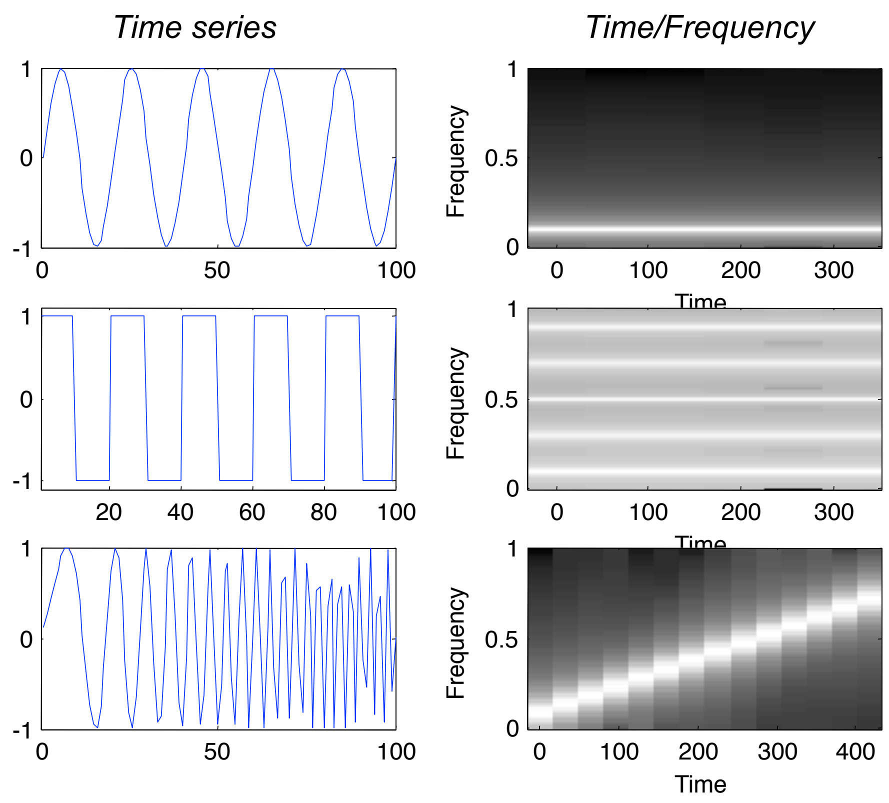
# Frequency domain representation

- Time series can be decomposed in terms of "sinusoid presence"
  - See how many sinusoids you can add up to get to a good approximation
  - Informally called the spectrum
- No temporal information in this representation, only frequency information
  - So a sine with a changing frequency is a smeared spike
- Not that great of a representation for dynamically changing sounds



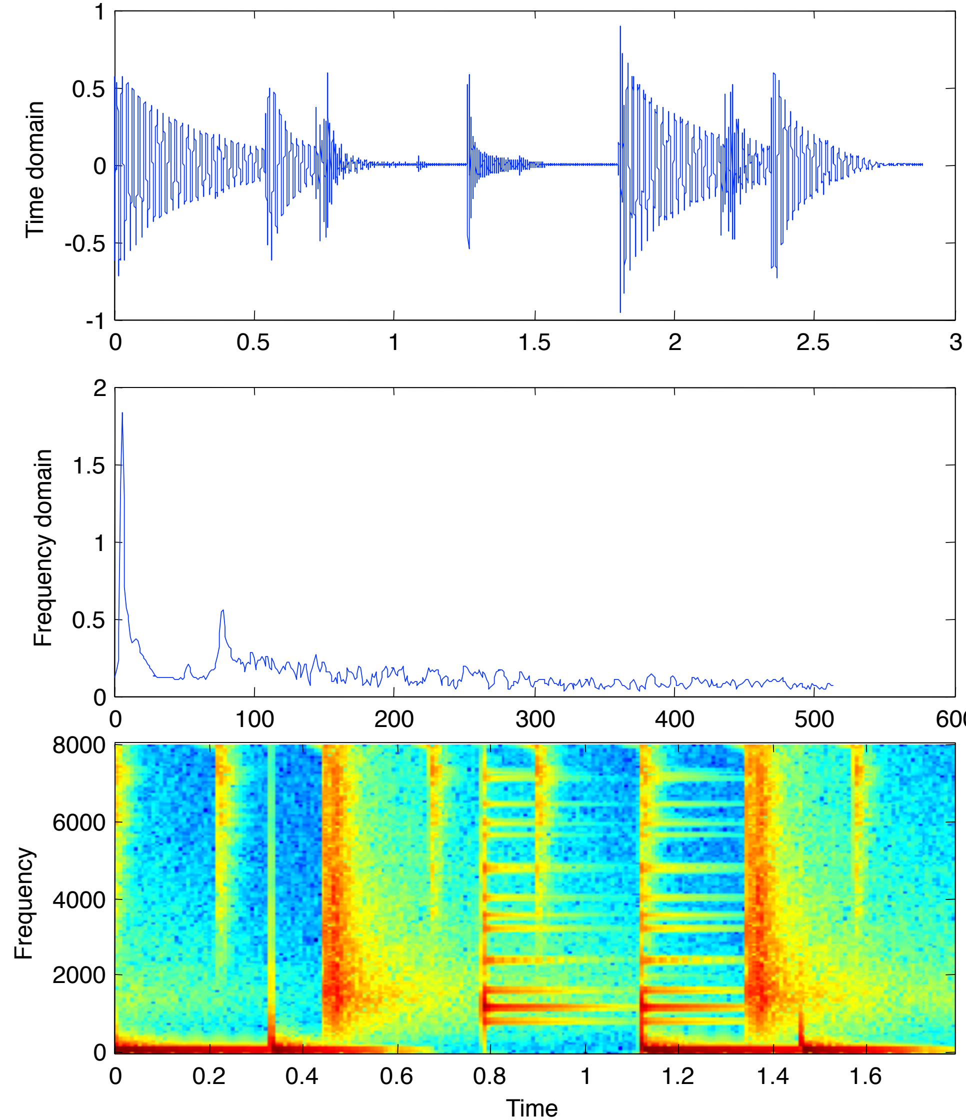
# Time/frequency representation

- Many names/varieties
  - Spectrogram, sonogram, periodogram, …
- A time ordered series of frequency compositions
  - Can help show how things move in both time and frequency
- The most useful representation so far!
  - Reveals information about the frequency content without sacrificing the time info



# A real example

- Time domain
  - We can see the events
  - We don’t know how they sound like though!
- Spectrum
  - We can see a lot of bass and few middle freqs
  - But where in time are they?
- Spectrogram
  - We can “see” each individual sound
  - And we know how it sounds like!



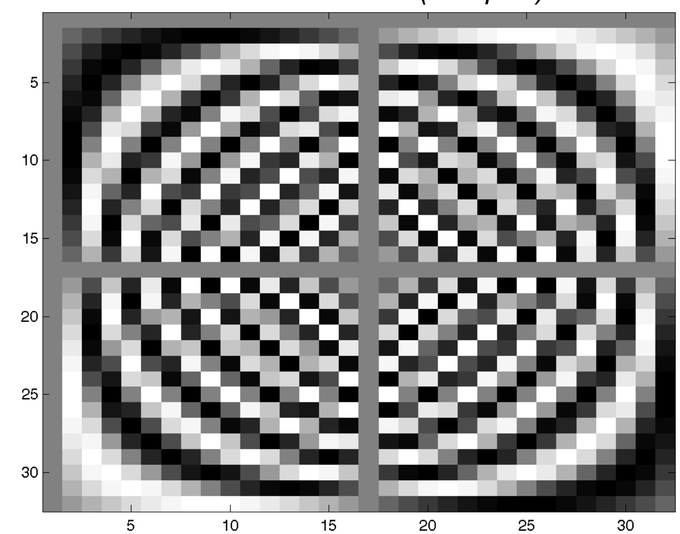
# The Discrete Fourier Transform

- So how do we get from time domain to frequency domain?
  - It is a matrix multiplication (a rotation in fact)
- The Fourier matrix is square, orthogonal and has complex-valued elements

$$F_{j,k} = \frac{1}{\sqrt{N}} e^{ijk\frac{2\pi}{N}} = \frac{1}{\sqrt{N}}\left(\cos\frac{jk2\pi}{N} + i\sin\frac{jk2\pi}{N}\right)$$

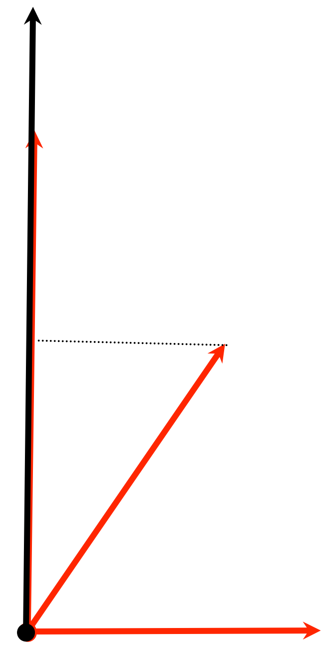
- Multiply a vectorized time-series with the Fourier matrix and voila!

The Fourier matrix (real part)



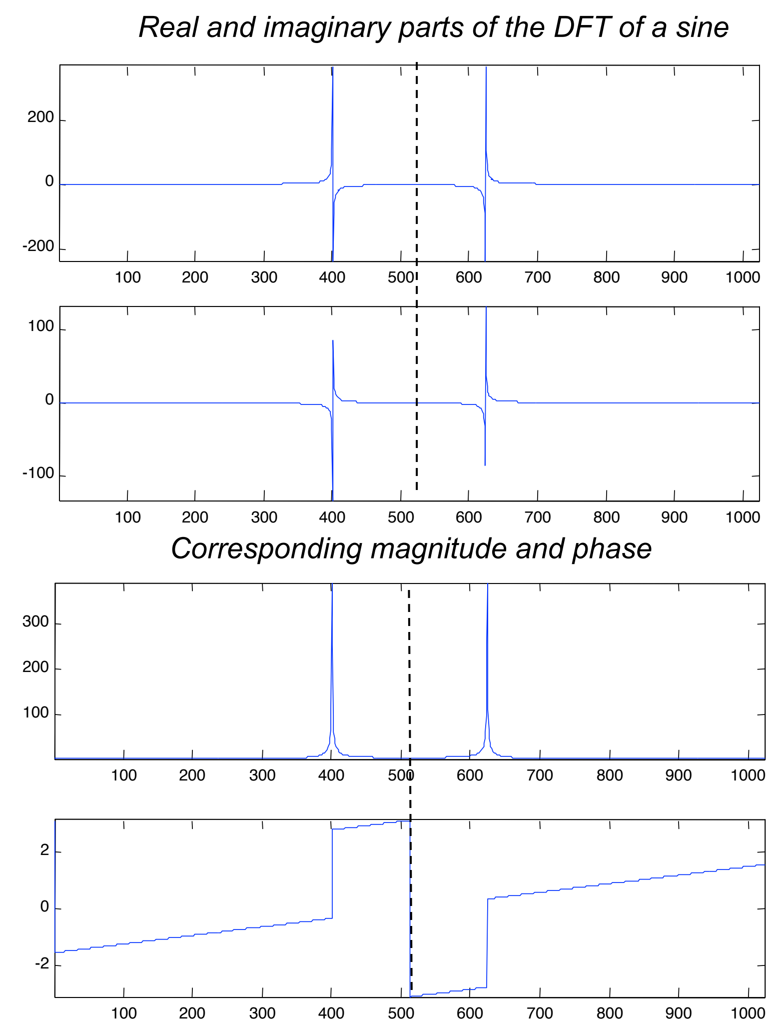
# How does the DFT work?

- Multiplying with the Fourier matrix
  - We dot product each Fourier row vector with the input
  - If two vectors point the same way their dot product is maximized
- Each Fourier row picks out a single sinusoid from the signal
  - In fact a complex sinusoid
  - Since all the Fourier sinusoids are orthogonal there is no overlap
- The resulting vector contains how much of each Fourier sinusoid the original vector had in it



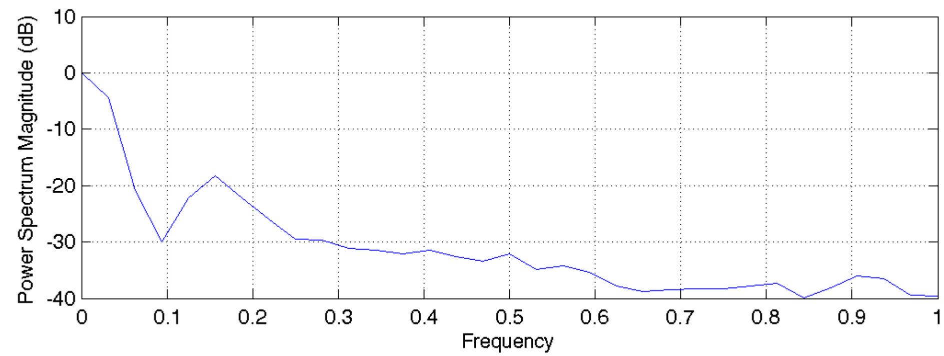
# The DFT in a little more detail

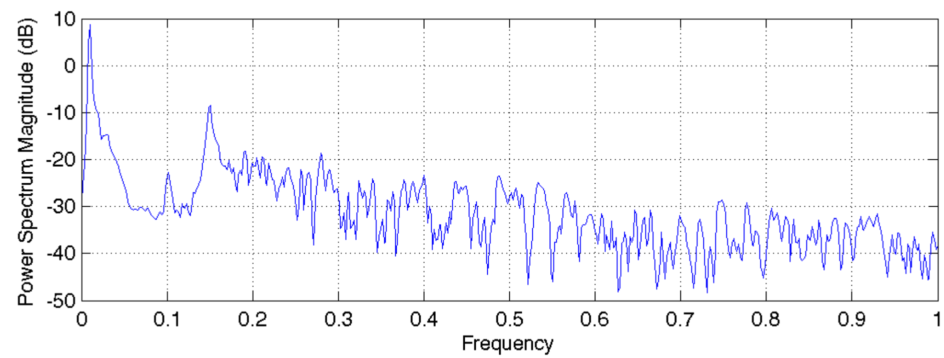
- The DFT features complex numbers
  - Doesn't have to, but it is convenient for other things
- The DFT result for <u>real</u> signals is *conjugate symmetric*
  - The middle value is the highest frequency (Nyquist)
  - Working towards the edges we traverse all frequencies downwards
  - The two sides are mutually conjugate complex numbers
- The interesting parts of the DFT are the magnitude and the phase
  - Abs( F) = || F ||
  - Arg( F) = ∡ F
- To go back we apply the DFT again (with some scaling)



# Size of a DFT

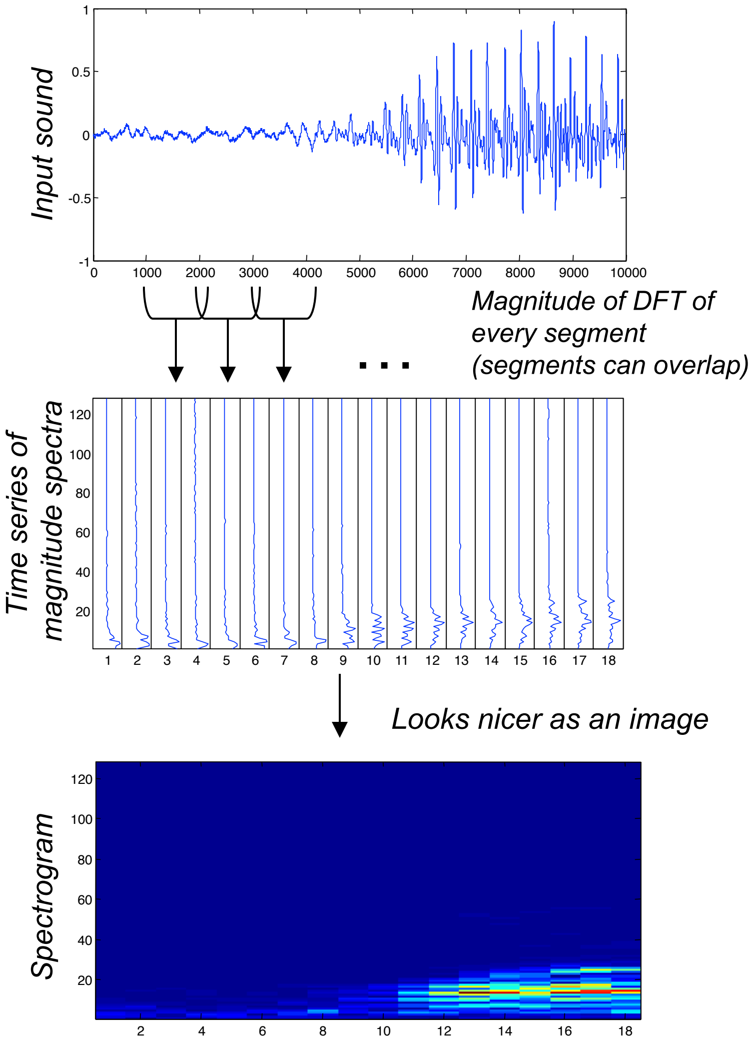
- The bigger the DFT input the more frequency resolution
  - But the more data we need!
- Zero padding helps
  - Stuff a lot of zeros at the end of the input to make up for few data
  - But we don't really infuse any more information we just make prettier plots



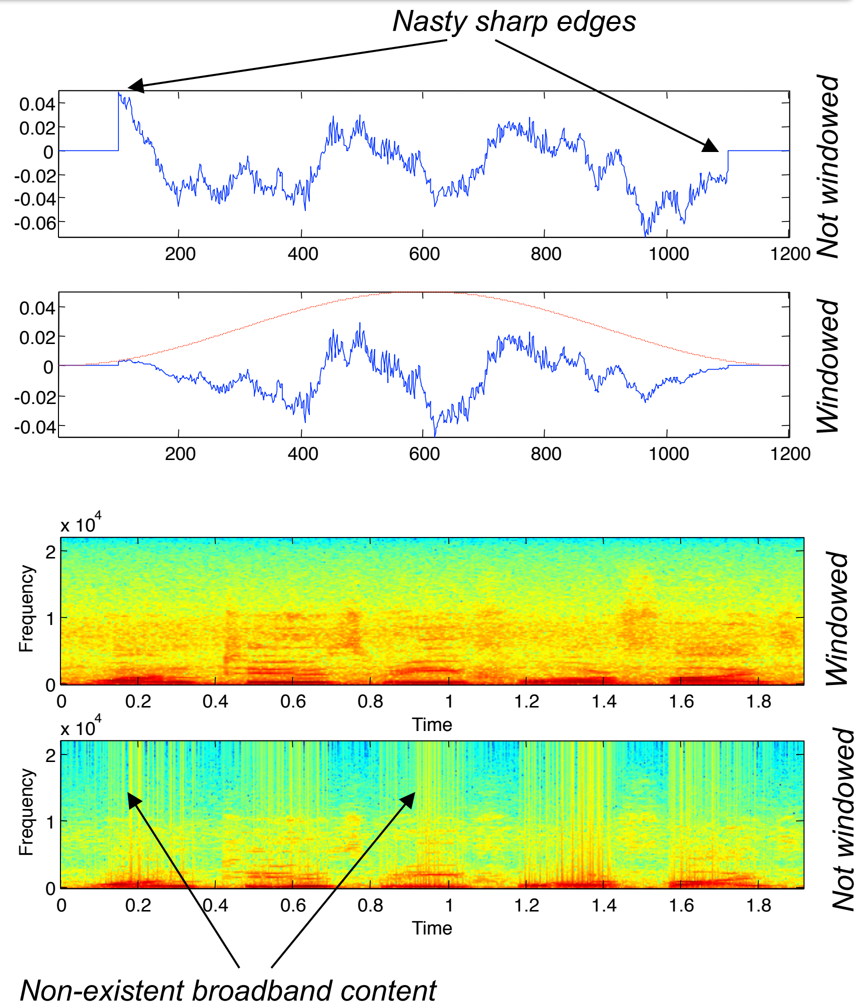
# From the DFT to a spectrogram

- The spectrogram is a series of consecutive magnitude DFTs on a signal
  - This series is taken off consecutive segments of the input
- It is best to taper the ends of the segments
  - This reduces "fake" broadband noise estimates
- It is wise to make the segments overlap
  - Due to windowing
- The parameters to use are
  - The DFT size
  - The overlap amount
  - The windowing function



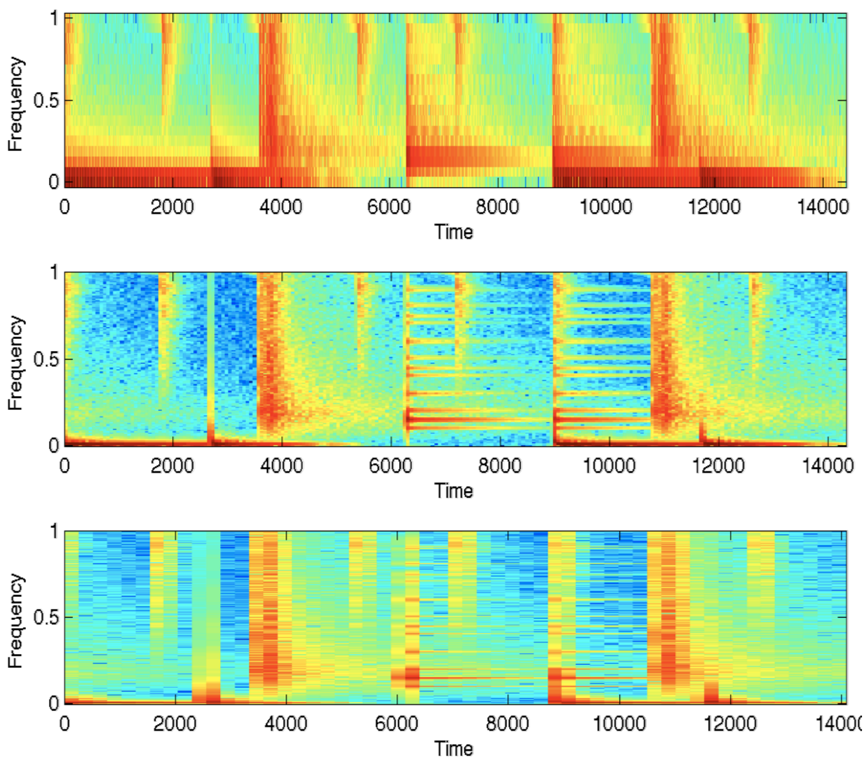
# Why window?

- Discontinuities at ends cause noise
  - Start and end point must taper to zero
- Windowing
  - Eliminates the sharp edges that cause broadband noise
- Overlap
  - Since we have windowed we need to take overlapping segments to make up for the attenuated parts of the input



# Time/Frequency tradeoff

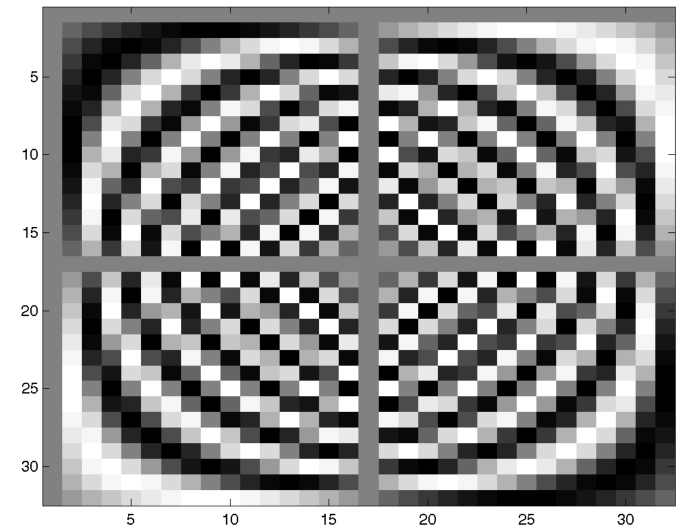
- Heisenberg's uncertainty principle
  - We can't accurately know both the frequency and the time position of a wave
  - Also in particle physics with speed and position of a particle
- Spectrogram problems
  - Big DFTs sacrifice temporal resolution
  - Small DFTs have lousy frequency resolution
- We can use a denser overlap to compensate
  - Ok solution, not great

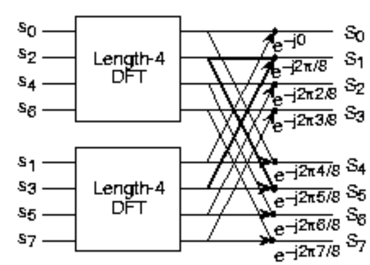


# The Fast Fourier Transform (FFT)

- The Fourier matrix is special
  - Many repeating values
  - Unique repeating structure
- We can decompose a Fourier transform to two Fourier transforms of half the size
  - Also includes some twiddling with the data
  - Two Fourier smaller transforms are faster than one big one
  - We keep decomposing it until we have a very small DFT
- This results into a really fast algorithm that has driven communications forward!
  - The constraint is that the transform size is best if a power of two so that we can decompose it repeatedly

*The Fourier matrix, N = 32*

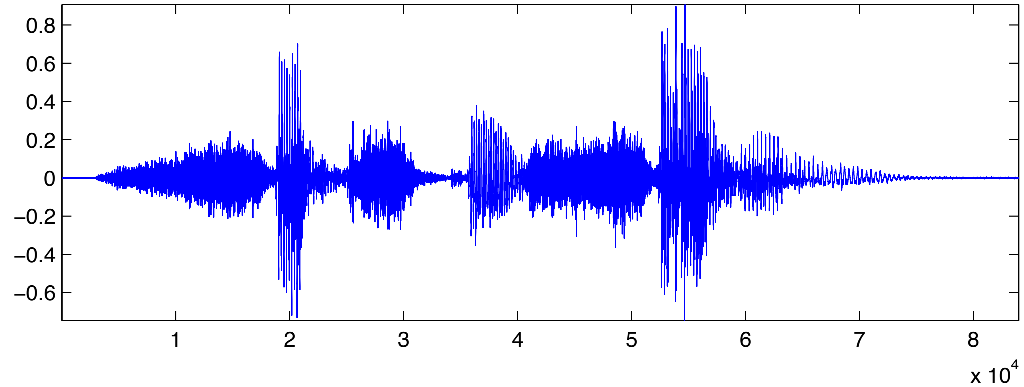


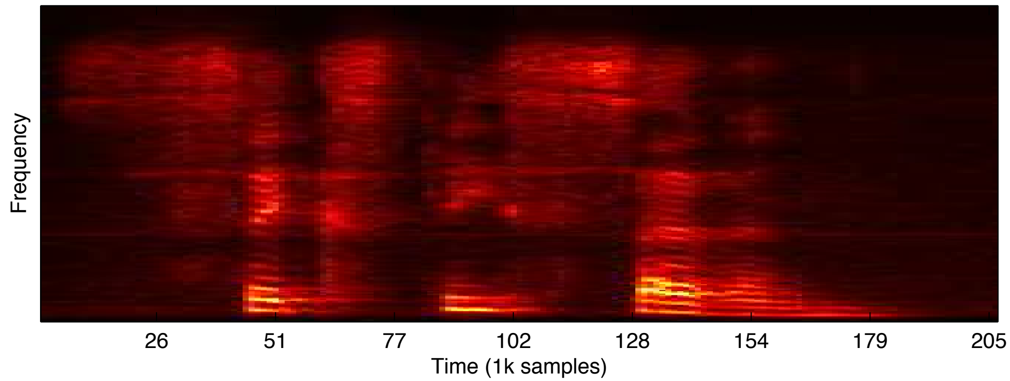




*Example FFT, N = 8*

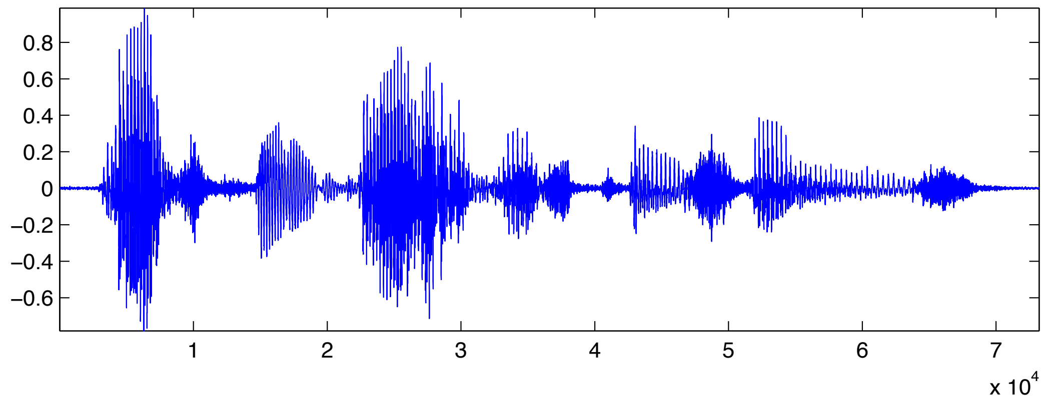
# Emulating the cochlea

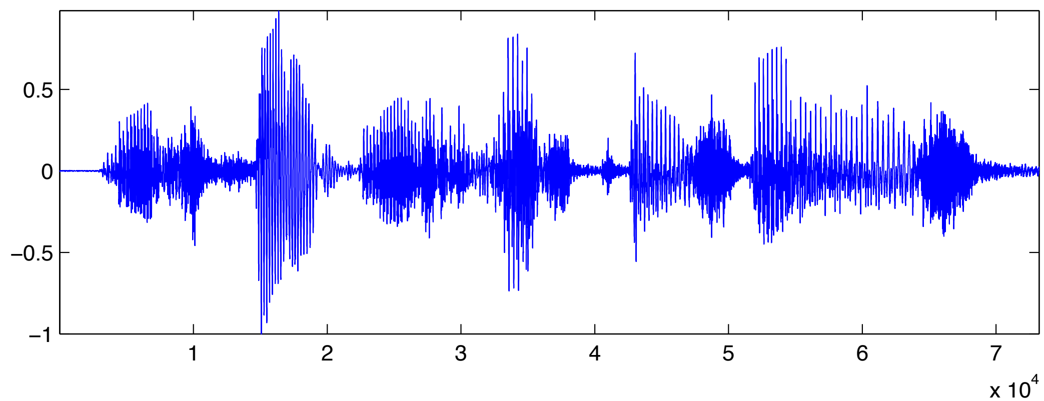
- Using the time/frequency domain
  - Take successive Fourier transforms
  - Keep their magnitude
  - Stack them in time
  - Now you can visually compare sounds!



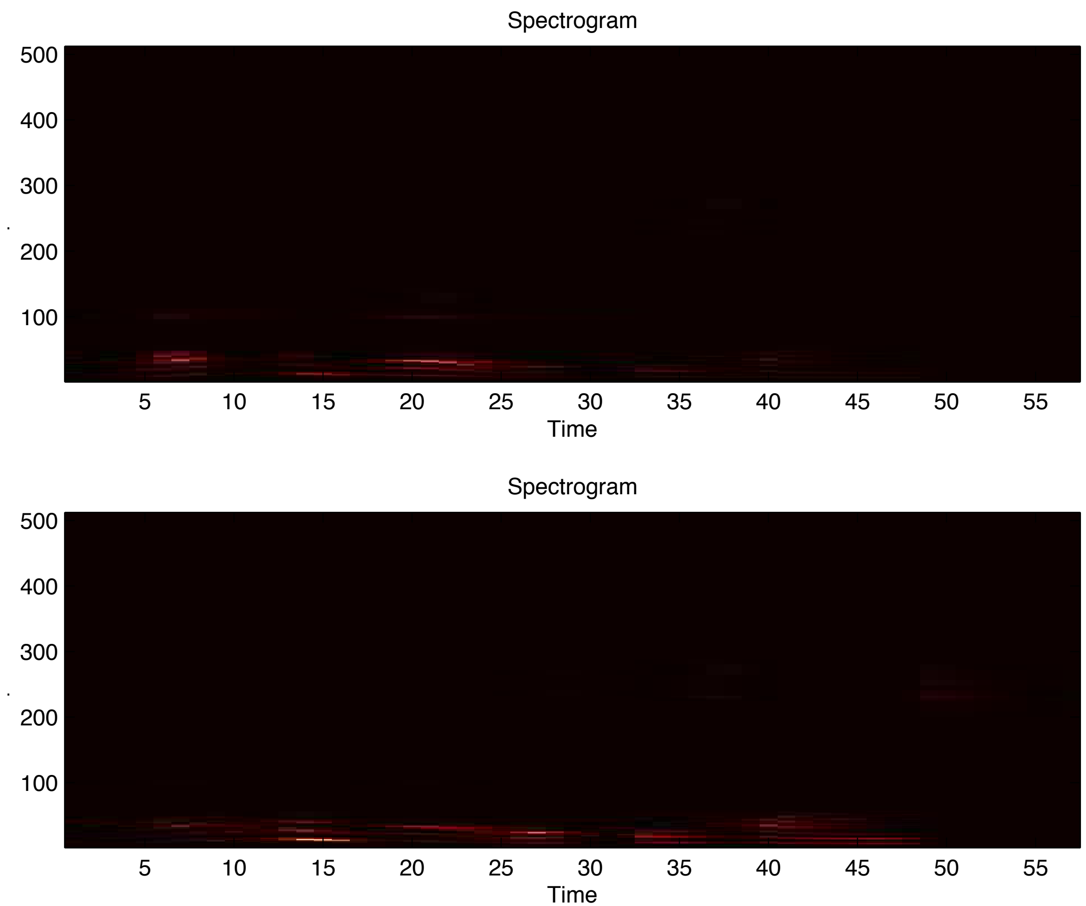



# Back to our example



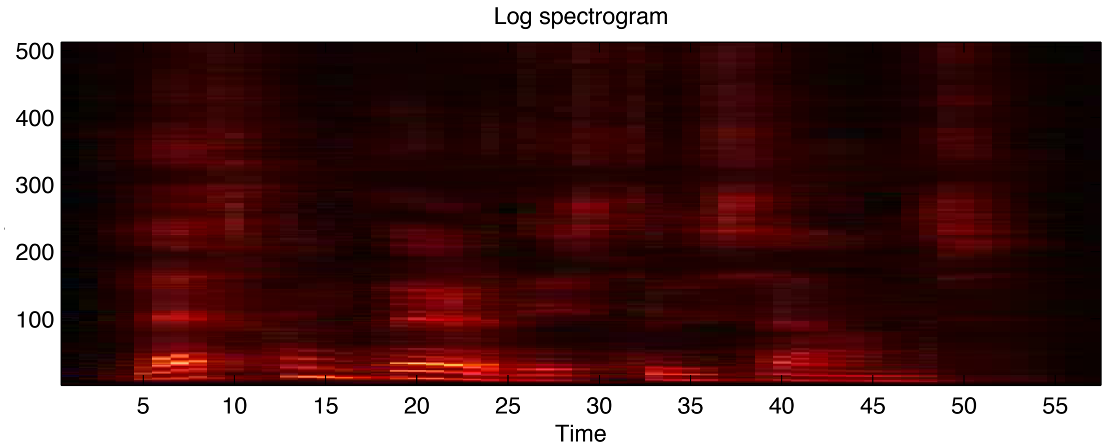
# Corresponding spectrograms

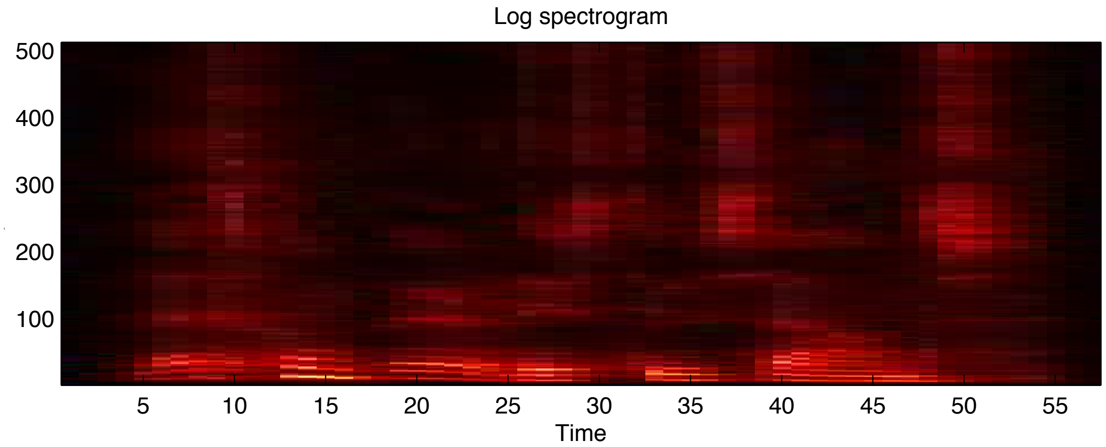


# A lesson from loudness perception

- We don’t perceive loudness linearly
- How much louder is the second “test”?
- The magnitude we plot should be logarithmic, not linear

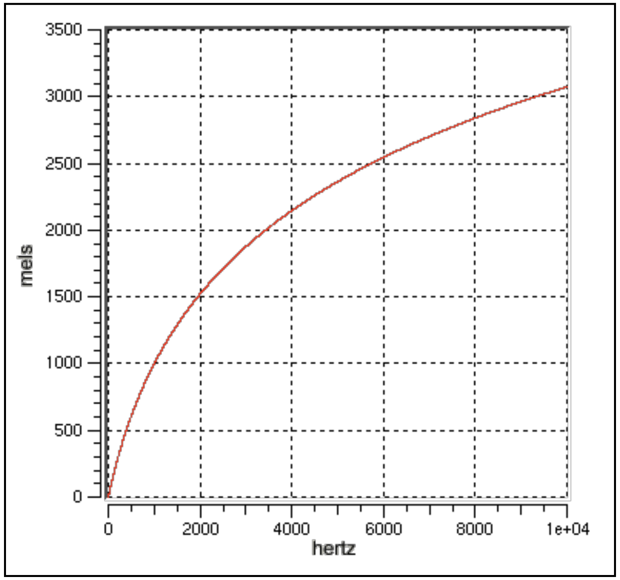
# Log spectrograms



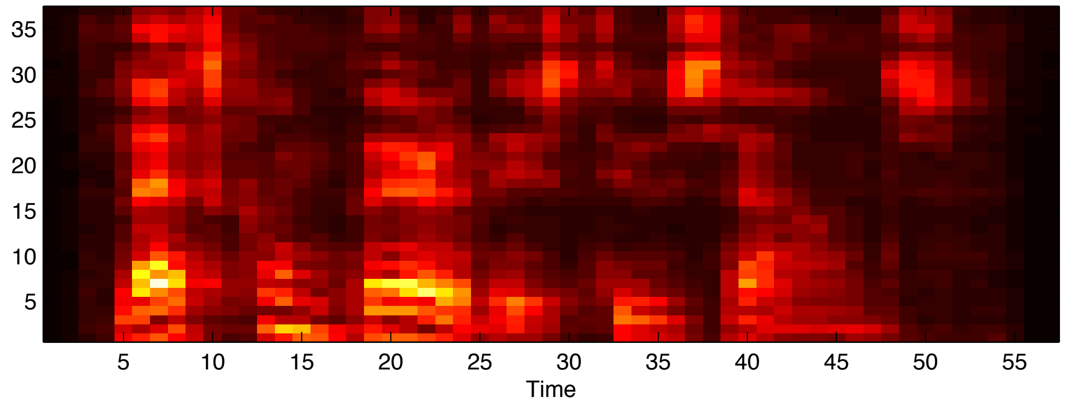
# A lesson from pitch perception

- Frequencies are not “linear”
  - Perceived scale is called mel
- Use that spacing instead
  - i.e. warp the frequency axis

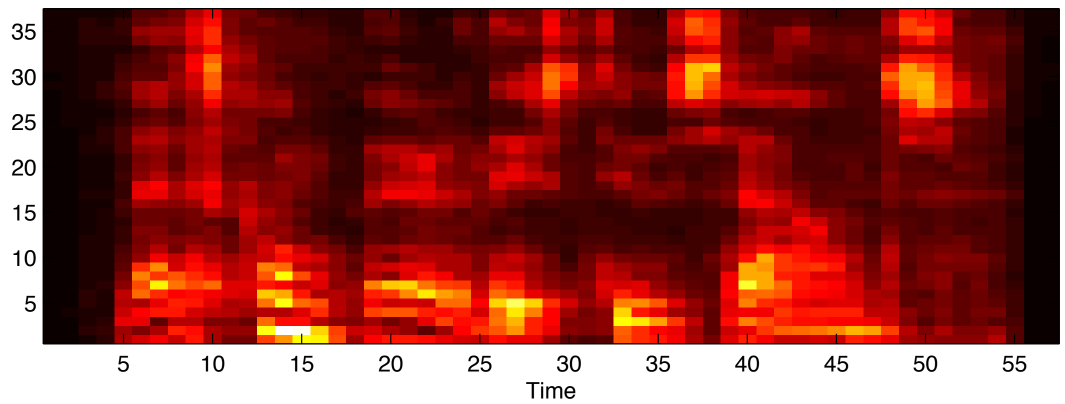


# "Mel spectra"

Log mel spectrogram

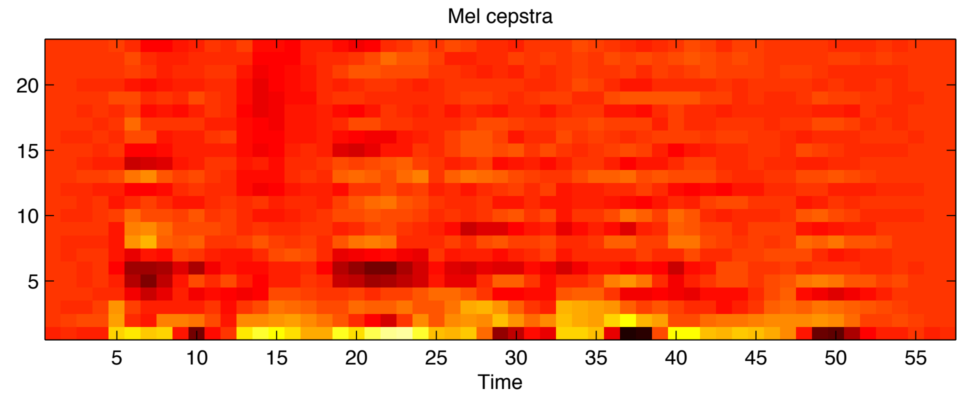


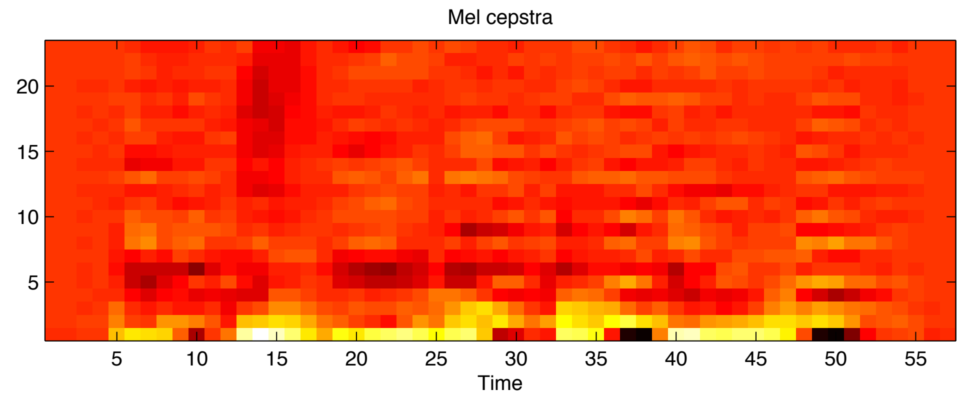

Log mel spectrogram



# One more trick

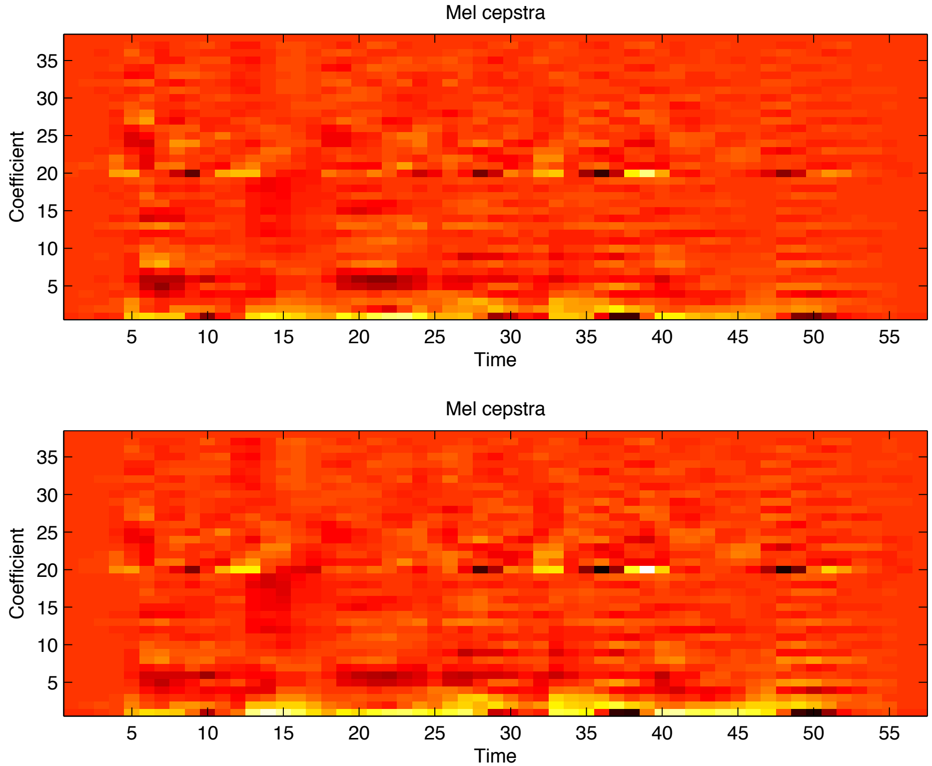
- Mel cepstra
  - Smooth the log mel spectra using one more frequency transform (the DCT)



# Adding some temporal info

- Deltas and delta-deltas
  - In sounds order is important
  - Using “delta features” we pay attention to change



# What more is there?

- Tons!
  - Spectral features
  - Waveform features
  - Higher level features
  - Perceptual parameter features
  - …

# Sound recap

- Go to time/frequency domain
  - We do so in the cochlea
- Frequencies are not linear
  - We perceive them in another scale
- Amplitude is not linear either
  - Use log scale instead
- Resulting features are used a lot
  - Further minor tweaks exist (more later)

# Next lecture

- Principal Component Analysis
- How to find features automatically
- How to “compress” data without info loss