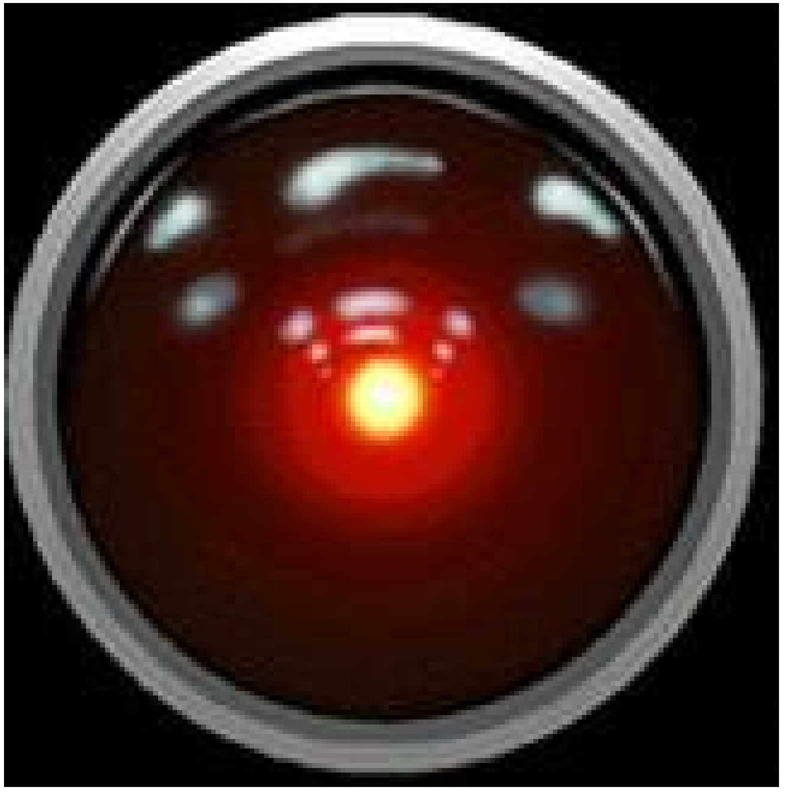


CS 378: Autonomous Intelligent Robotics

Instructor: Jivko Sinapov

http://www.cs.utexas.edu/~jsinapov/teaching/cs378/

Audio Processing
and Computational Perception
of Natural Sound

Why Sound?

Why Sound?

What actually happened:

The robot dropped a soda-can

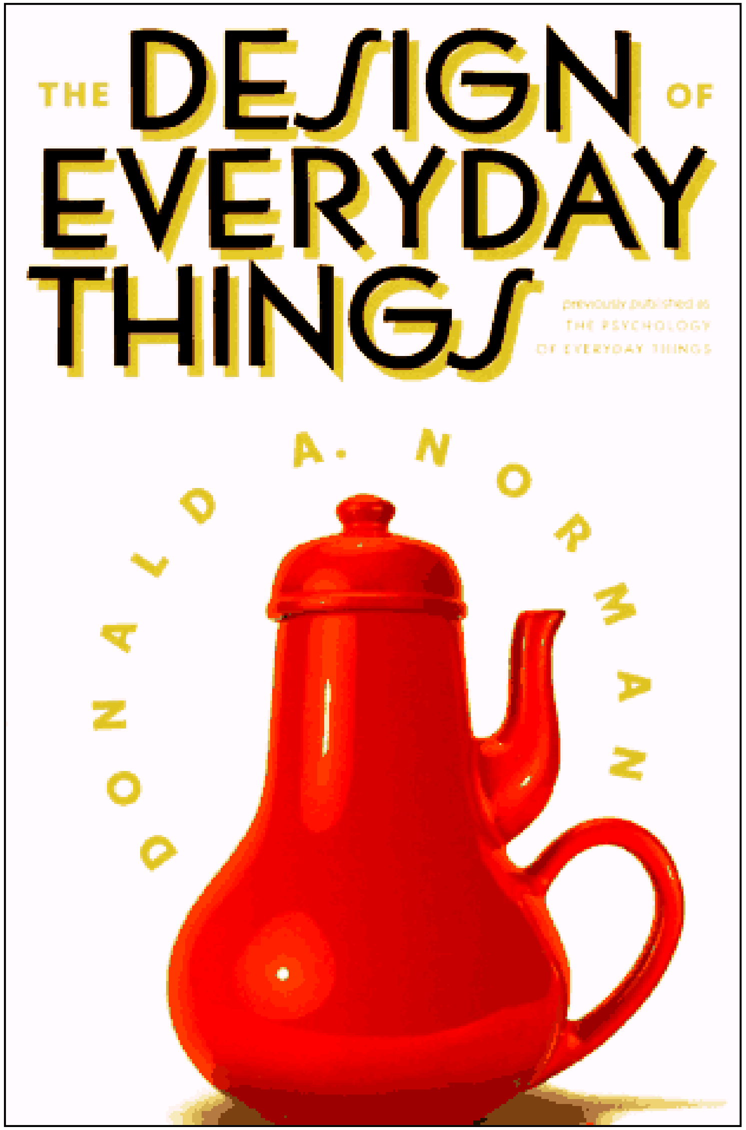
Why Natural Sound is Important

"…natural sound is as essential as visual information because **sound tells us about things that we can't see**, and it does so while our eyes are occupied elsewhere. "

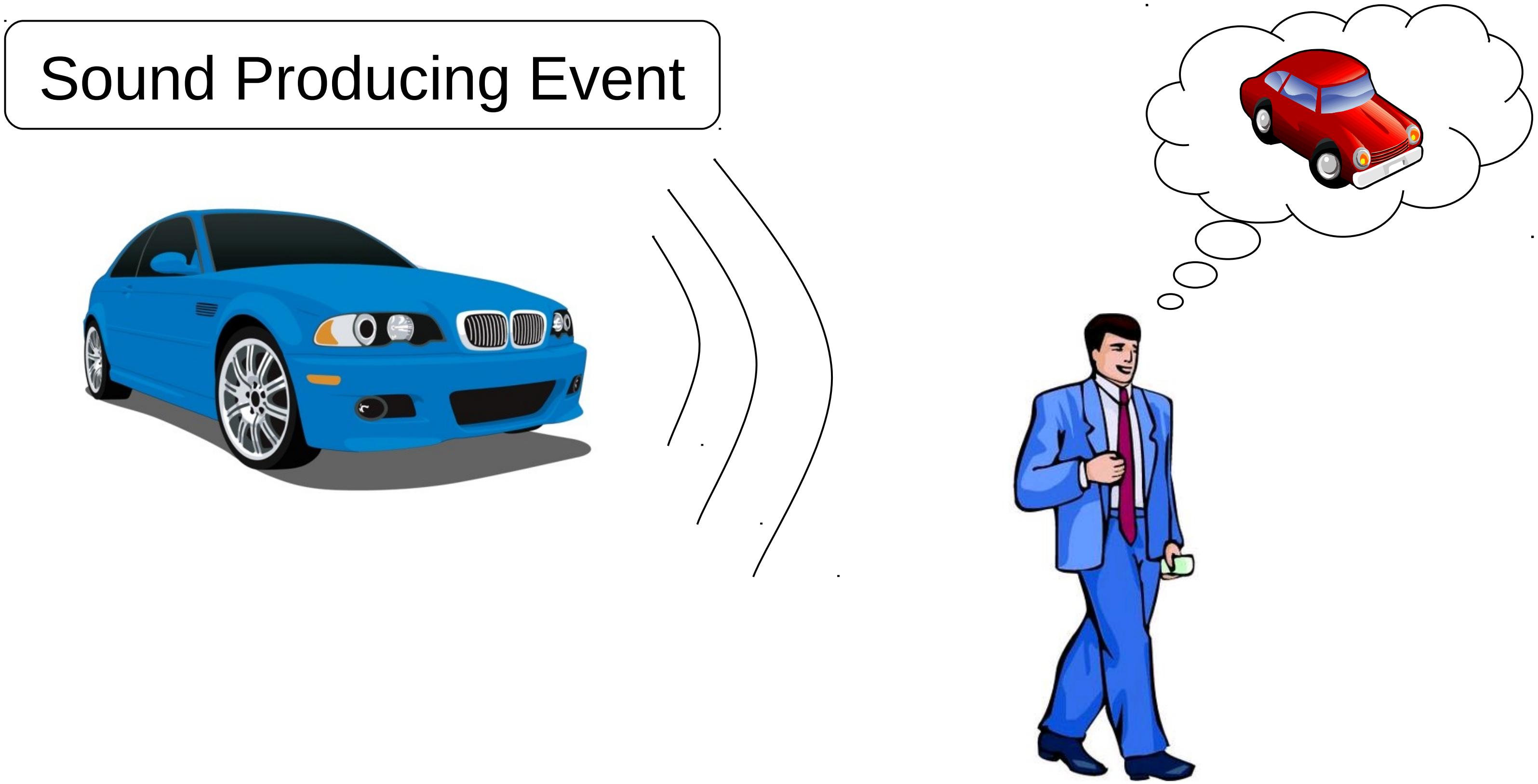
"Sounds are generated when materials interact, and the sounds tell us whether they are **hitting, sliding, breaking, tearing, crumbling, or bouncing**. "

"Moreover, **sounds differ according to the characteristics of the objects**, according to their size, solidity, mass, tension, and material. "

Don Norman, "*The Design of Everyday Things*", p.103



Why Natural Sound is Important



[Gaver, 1993]

Why should a robot use acoustic information?

Human environments are cluttered with objects that generate sounds

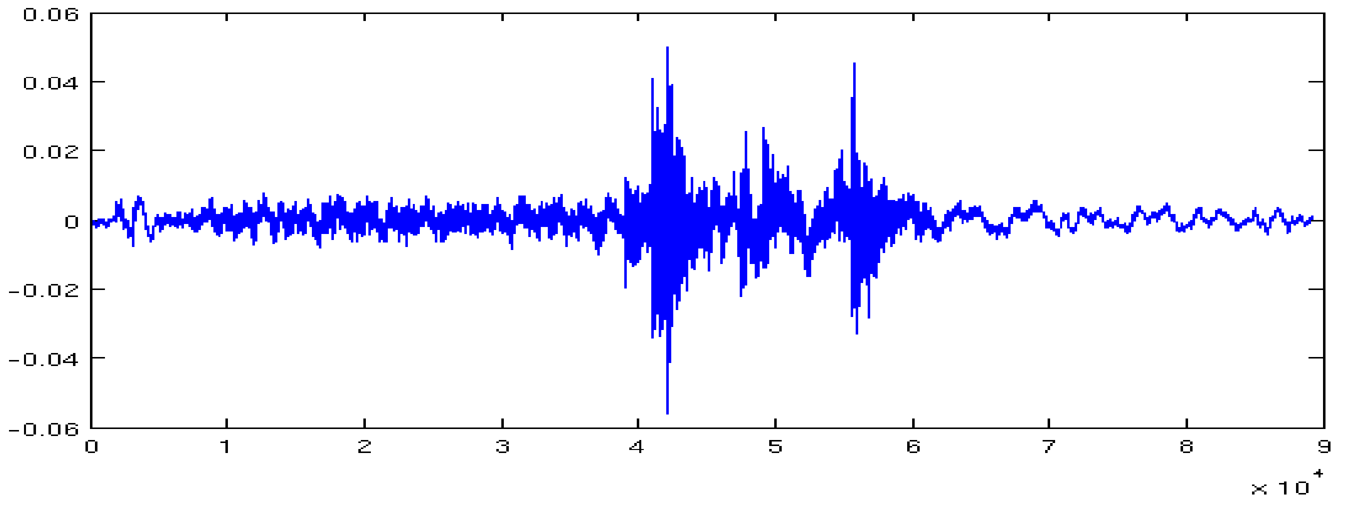
Help a robot perceive events and objects outside of field of view

Help a robot perceive material properties of objects, and form natural object categories

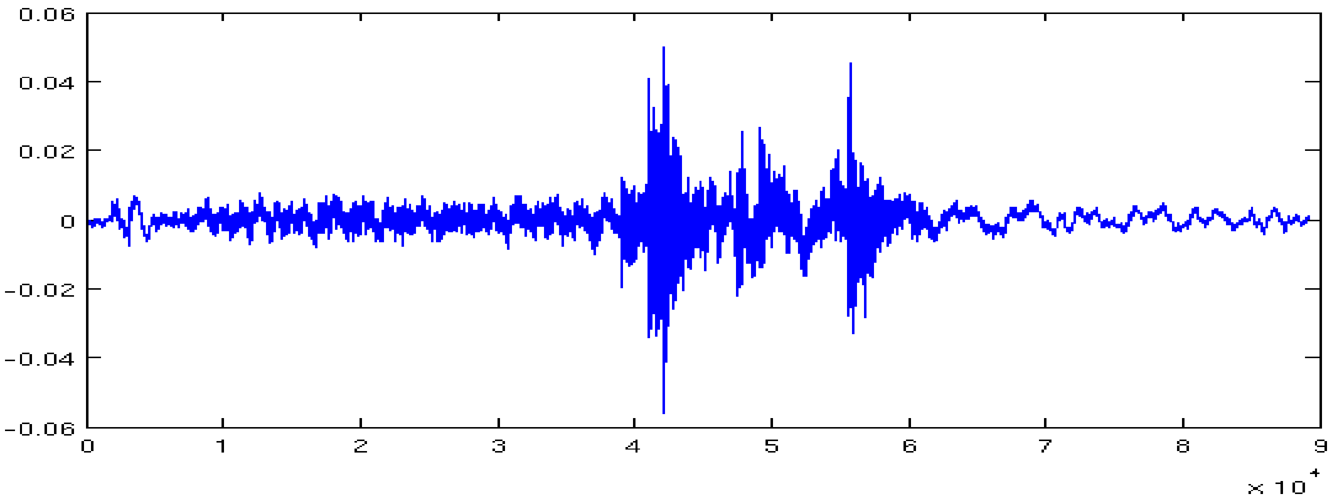
What is Sound?

What is Sound?

What is Sound?

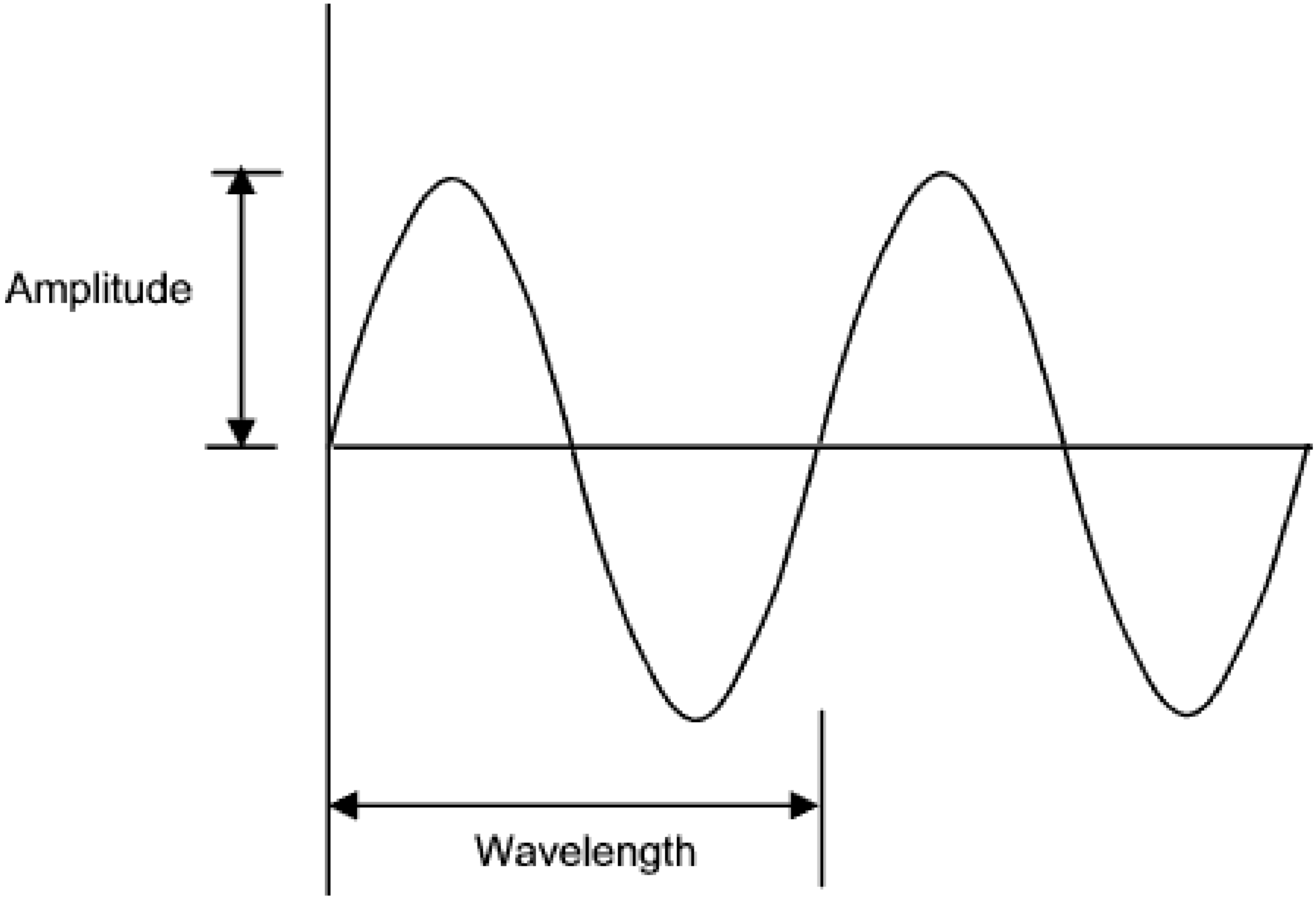


What is Sound?



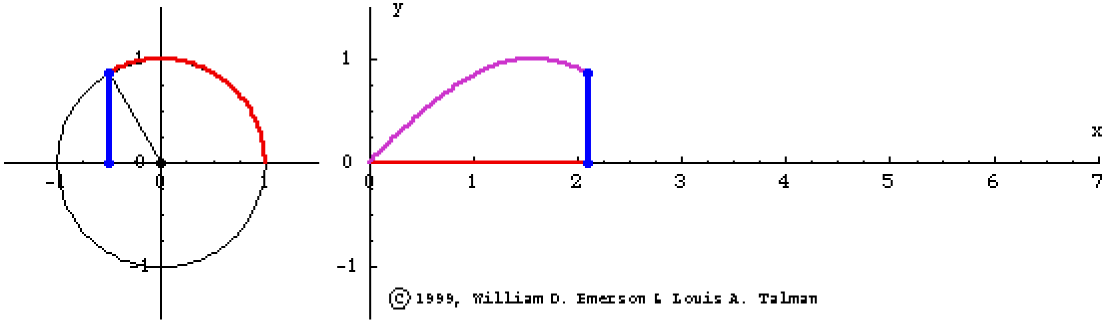
….from a computer's point of view, raw audio is a sequence of 44.1K floating point numbers arriving each second

Sine Wave



[http://www.audiophilejournal.com/what-is-a-hz-or-hertz-in-audio/]

Sine Curve



[http://clem.mscd.edu/~talmanl/HTML/SineCurve.html]

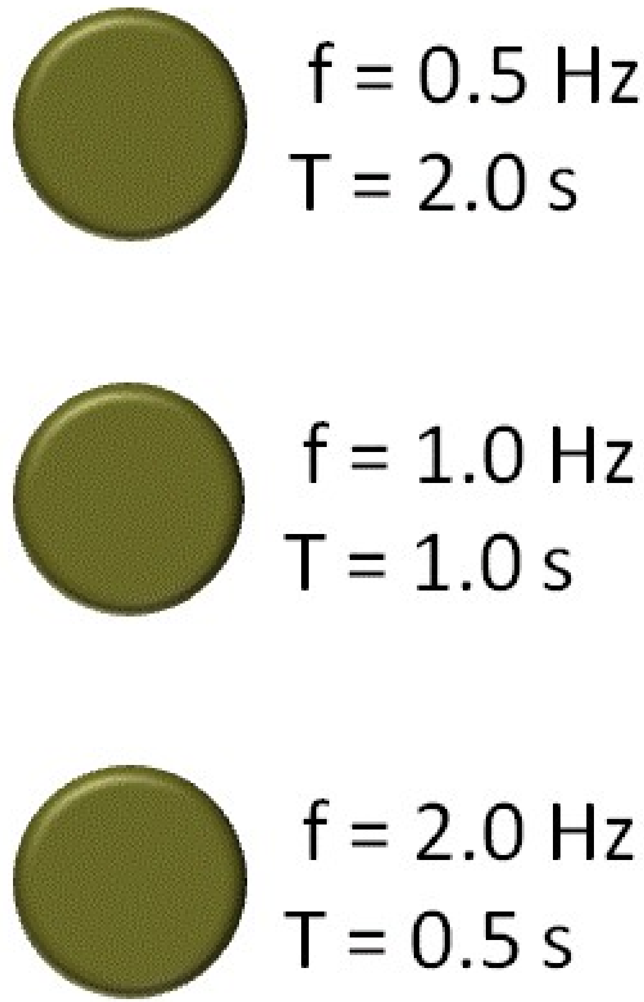
Frequency

- Measured in Hertz (Hz)
- Named after Heinrich Hertz
- 1 Hertz = 1 repetition per second
- Typically denoted with the letter f

Period

- How long does one cycle take?
- It is the reciprocal of the frequency
- Measured in seconds
- Typically denoted with the letter T

Frequency vs Period Animation



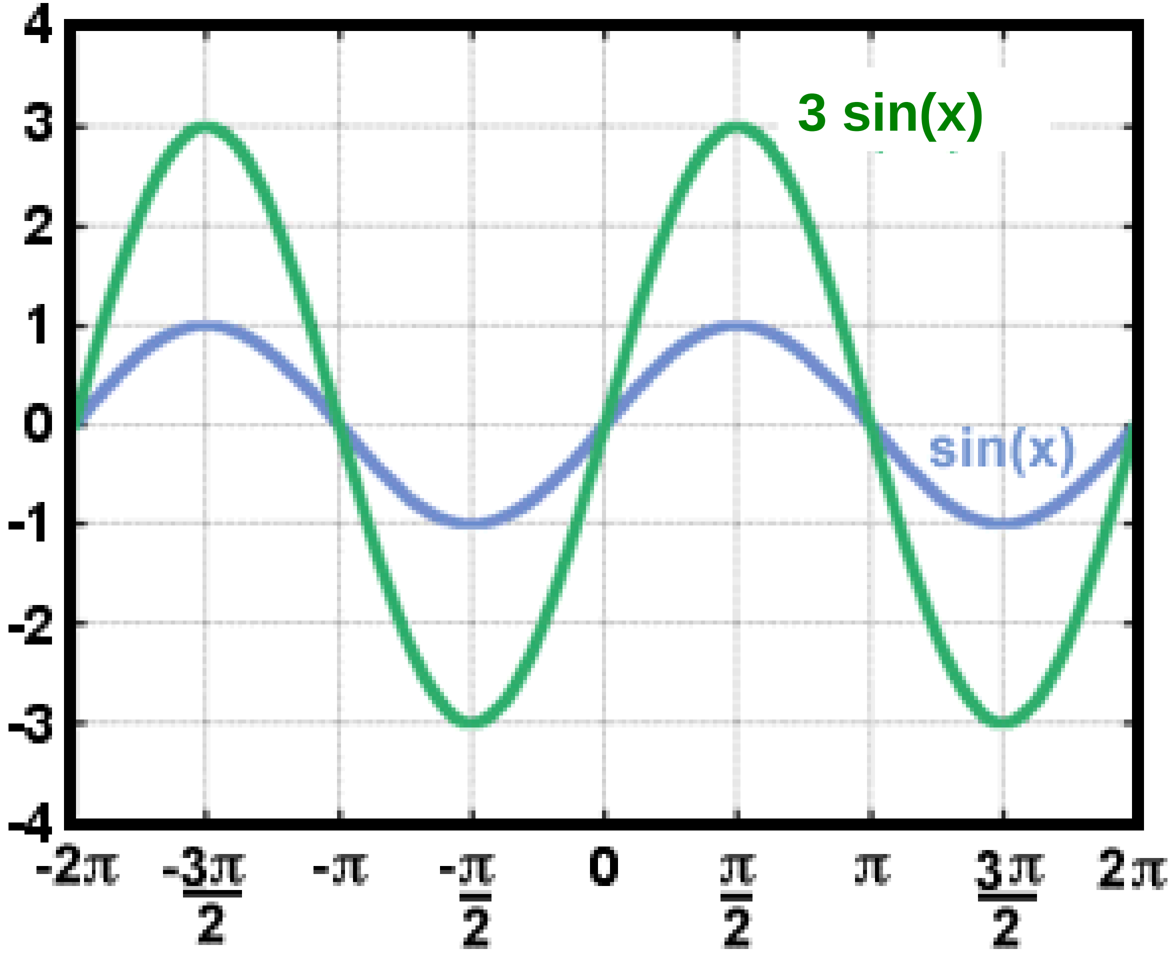
[http://en.wikipedia.org/wiki/Frequency]

Frequency vs Period

Frequency	1 mHz (10^{-3})	1 Hz (10^{0})	1 kHz (10^{3})	1 MHz (10^{6})	1 GHz (10^{9})	1 THz (10^{12})
Period (time)	1 ks (10^{3})	1 s (10^{0})	1 ms (10^{-3})	1 μs (10^{-6})	1 ns (10^{-9})	1 ps (10^{-12})

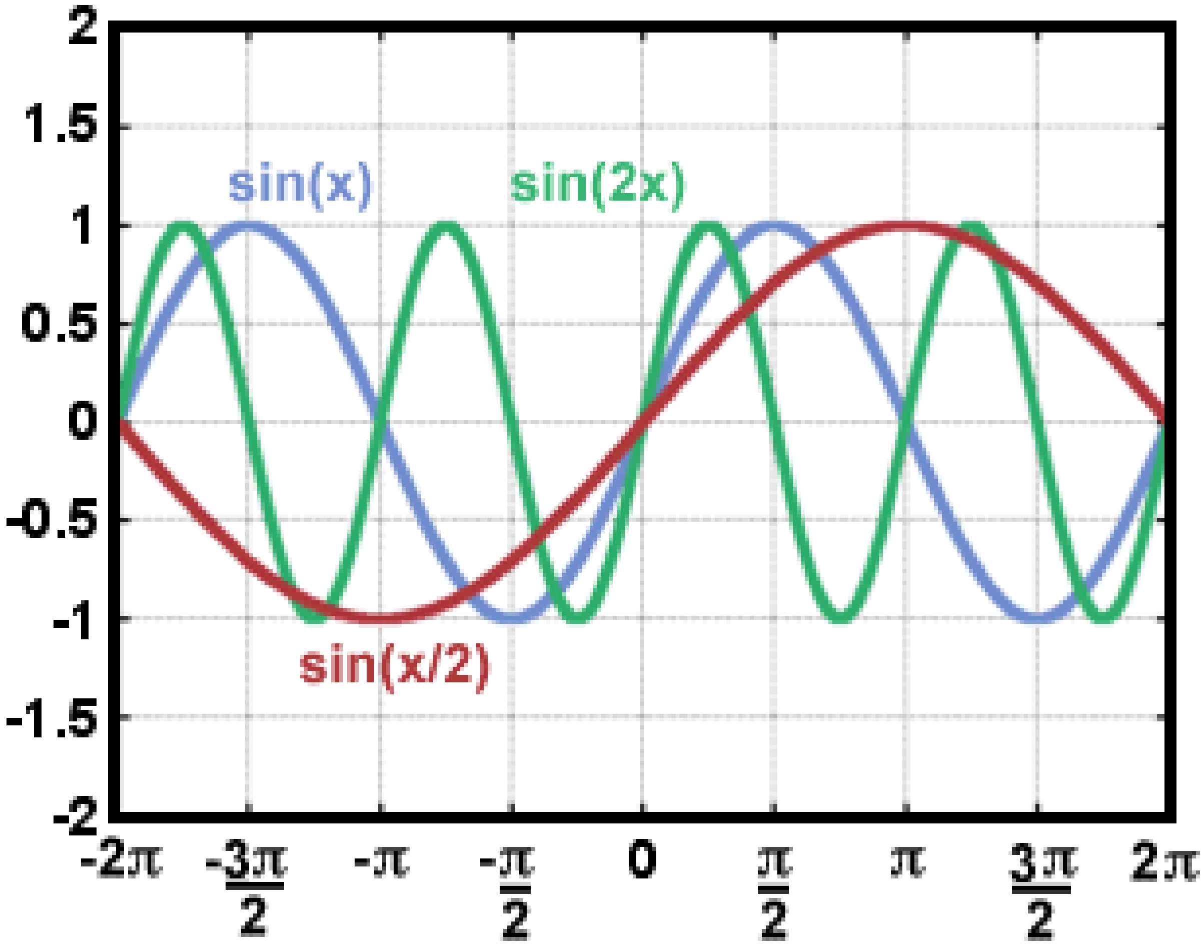
$$T = \frac{1}{f}$$

Amplitude
(vertical stretch)



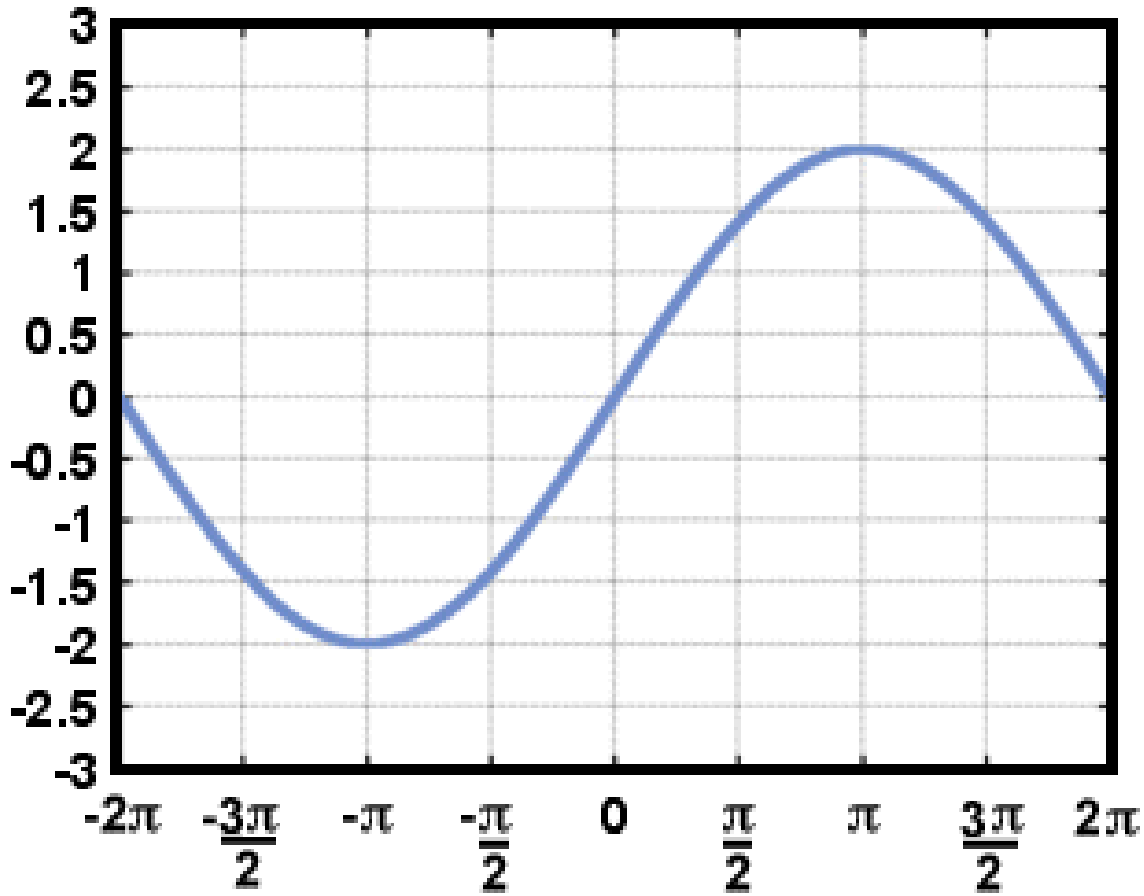
[http://www.sparknotes.com/math/trigonometry/graphs/section4.rhtml]

Frequency (horizontal stretch)



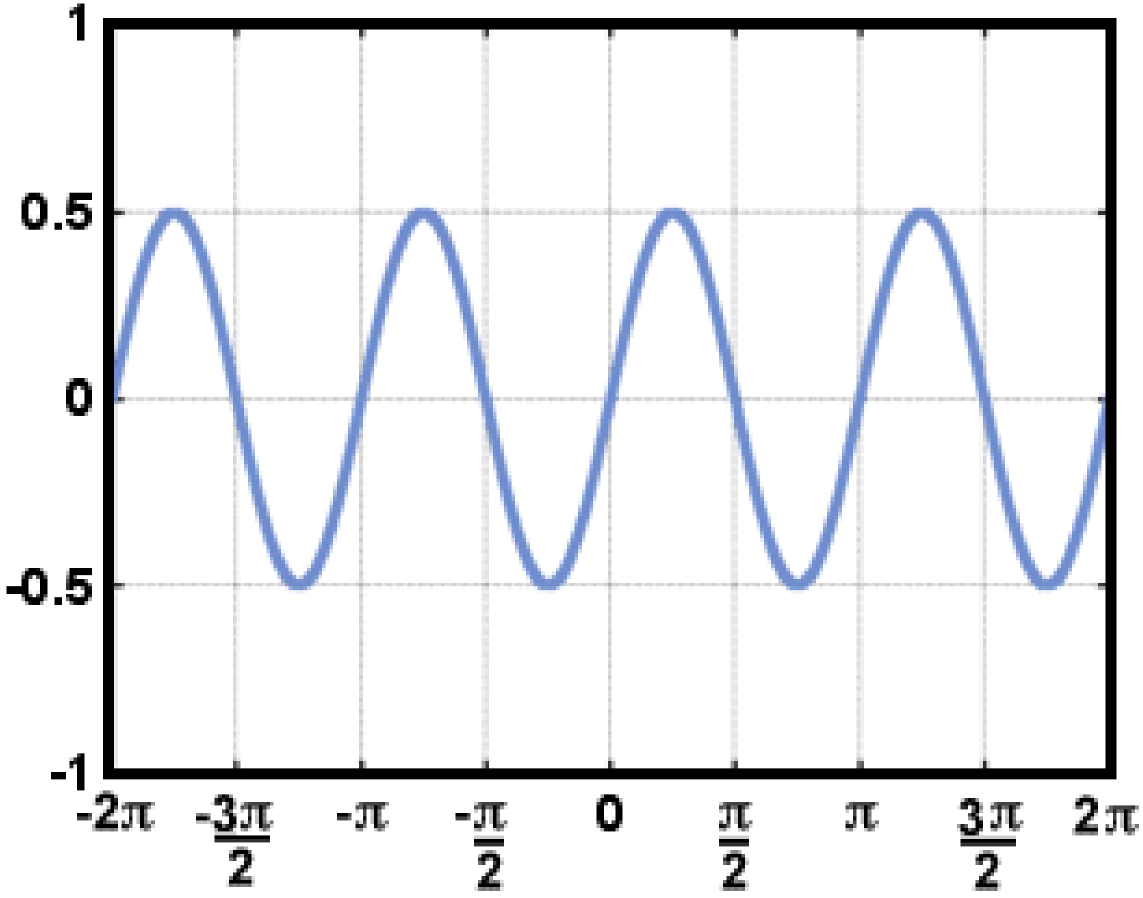
[http://www.sparknotes.com/math/trigonometry/graphs/section4.rhtml]

What is the Period and the Amplitude?



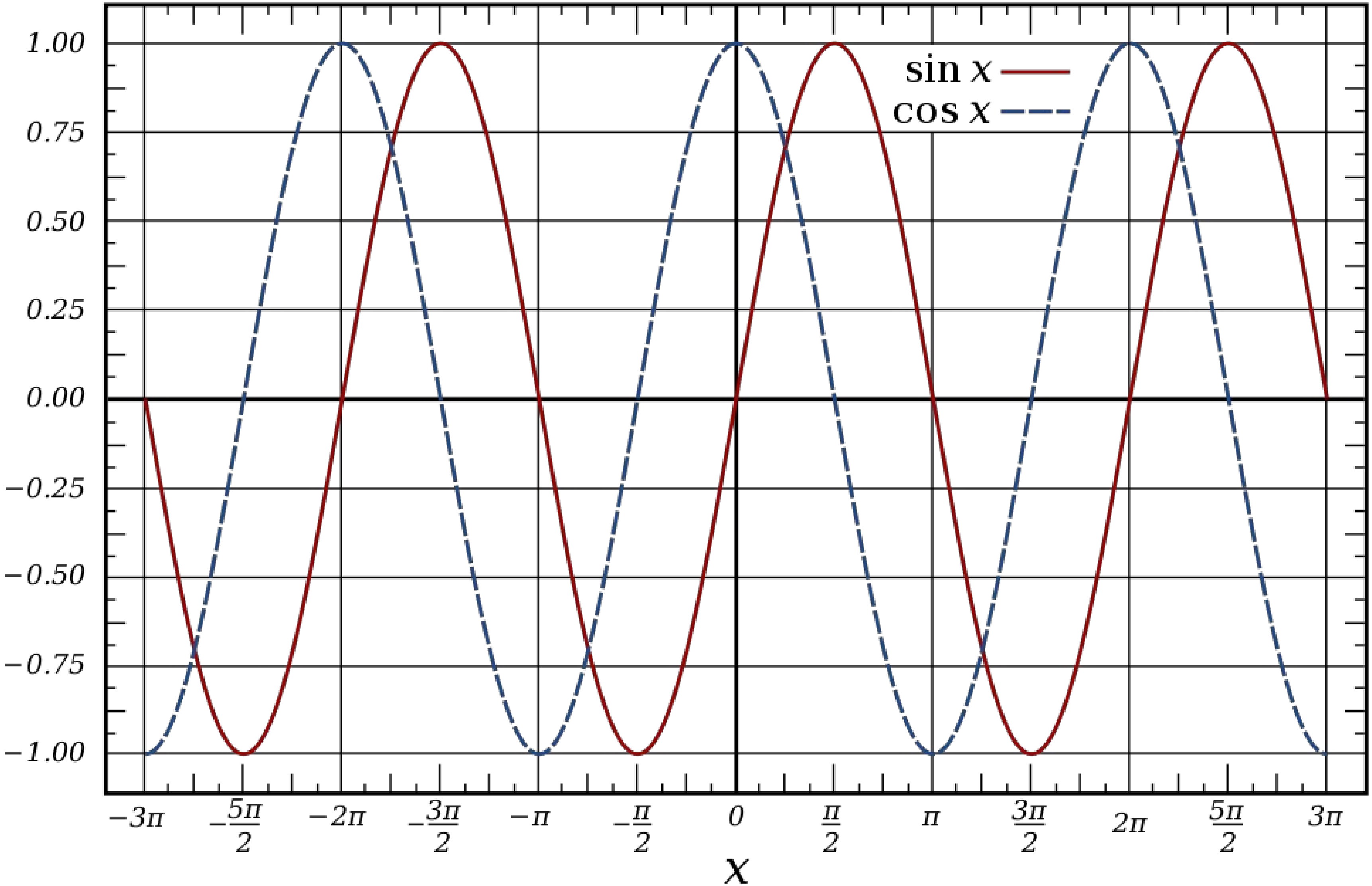
[http://www.sparknotes.com/math/trigonometry/graphs/problems_3.html]

What is the Period and the Amplitude?



[http://www.sparknotes.com/math/trigonometry/graphs/problems_3.html]

Sines vs Cosines



[http://en.wikipedia.org/wiki/Sine_wave]

Formula for the Sine Wave

$$y(t) = A \cdot \sin(\omega t + \phi)$$

Formula for the Sine Wave

$$y(t) = A \cdot \sin(\omega t + \phi)$$

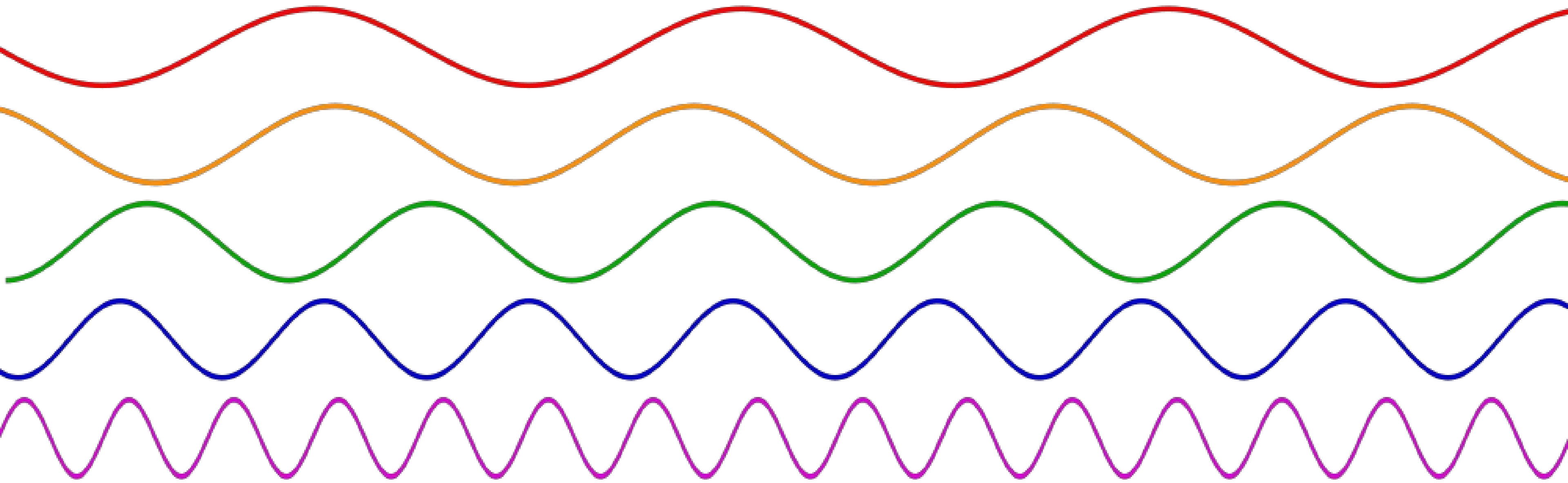
- A, the amplitude, is the peak deviation of the function from its center position.
- ω, the angular frequency, specifies how many oscillations occur in a unit time interval, in radians per second
- φ, the phase, specifies where in its cycle the oscillation begins at t = 0.

A function x(t) is periodic if we can find a T for which the following hold

$$x(t) = x(t+T) = x(t+2T) = x(t+3T) = \ldots$$

Sinusoidal waves of various frequencies

Low Frequency



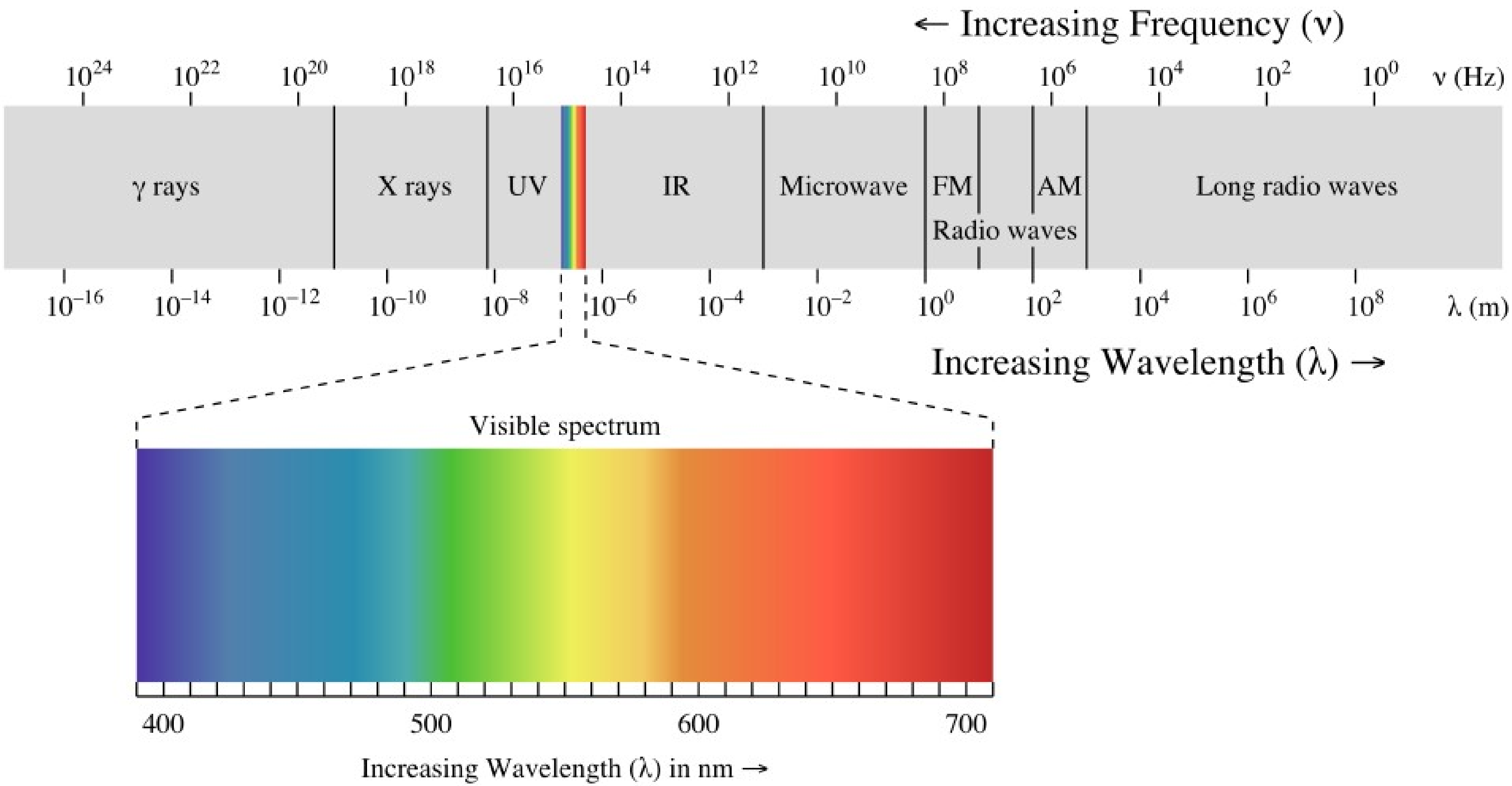
High Frequency

[http://en.wikipedia.org/wiki/Frequency]

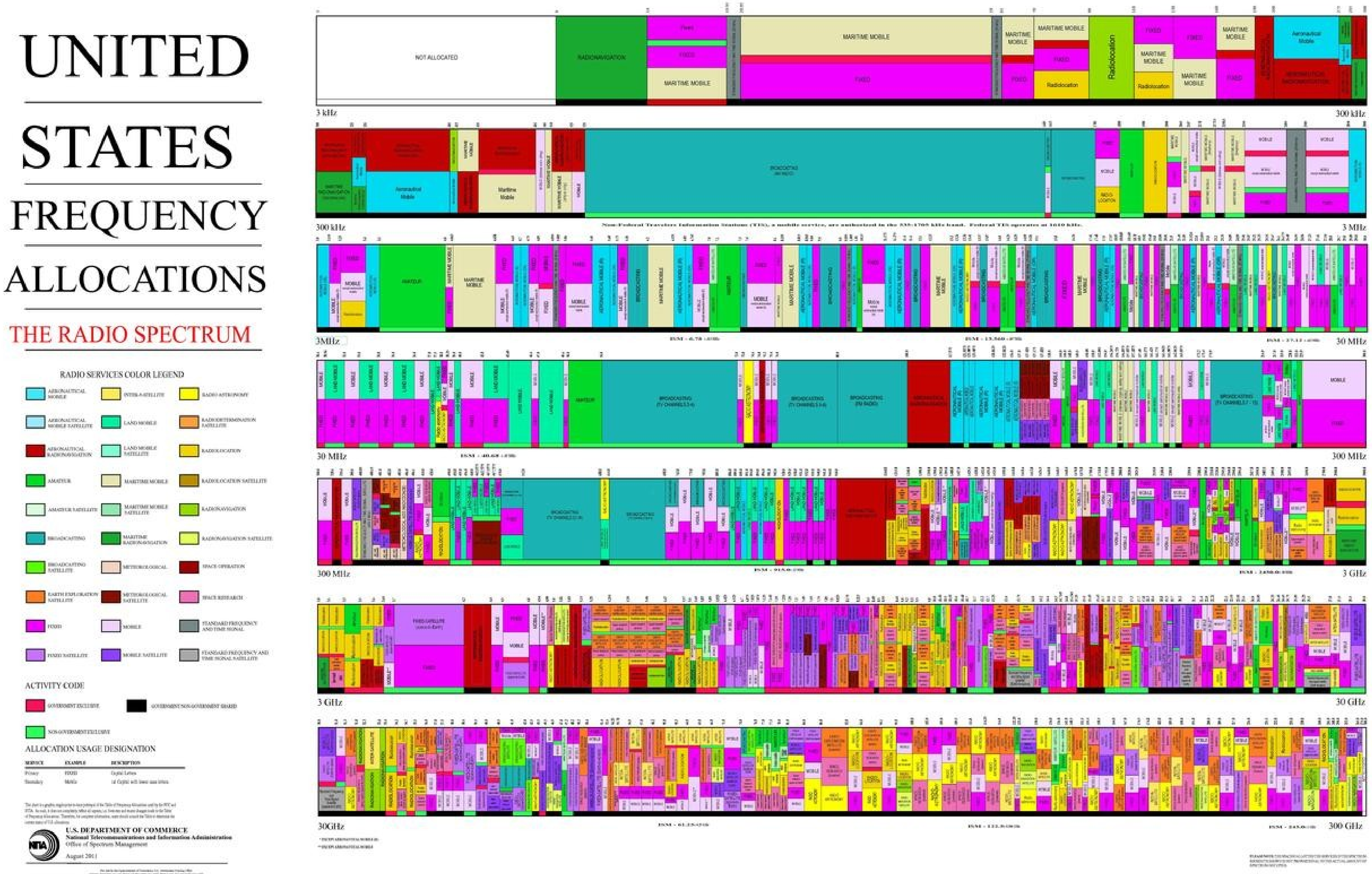
Spectrum

[http://en.wikipedia.org/wiki/Spectrum]

Light Spectrum



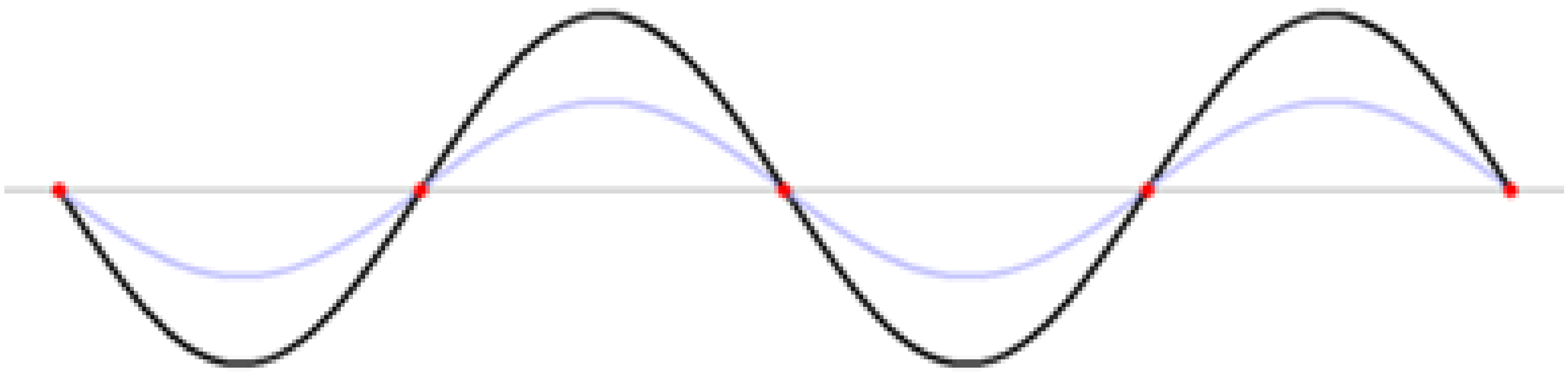
[http://en.wikipedia.org/wiki/Frequency]



[http://en.wikipedia.org/wiki/Spectrum_allocation]

Standing Wave

(shown in black, equal to the sum of the red and the blue waves traveling in opposite directions)

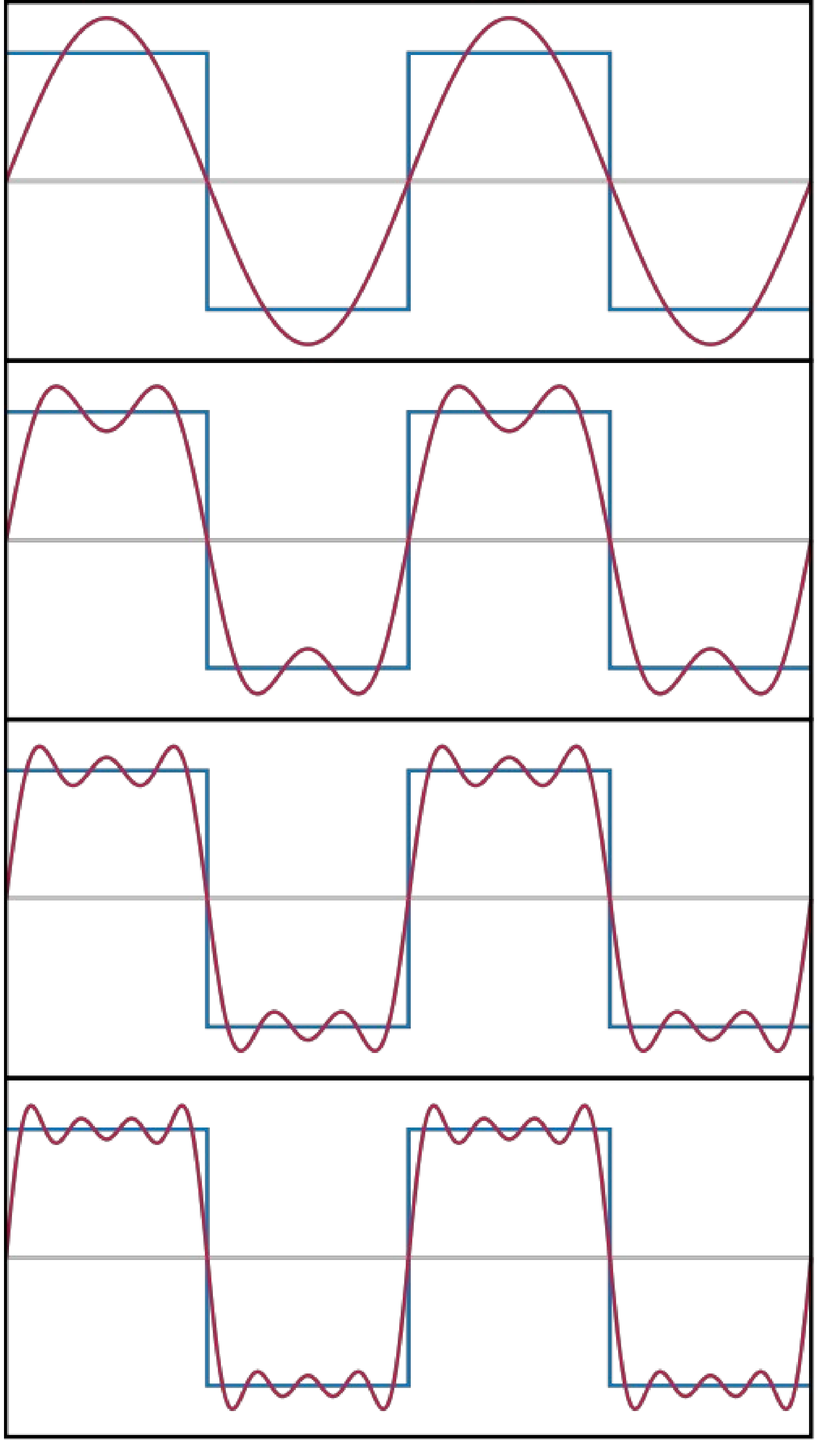


[http://en.wikipedia.org/wiki/Wavelength]

Fourier Series

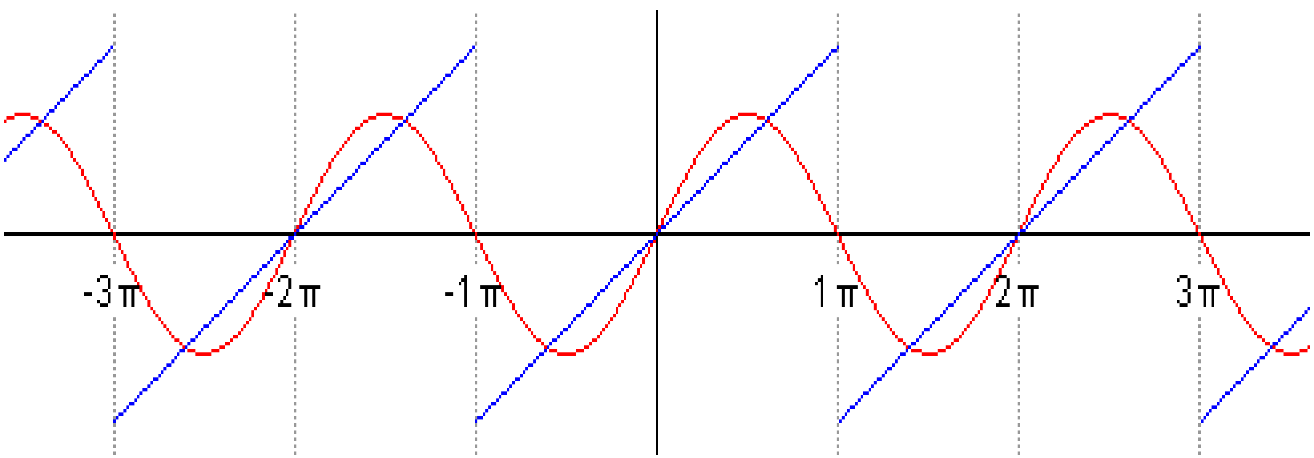
A Fourier series decomposes periodic functions or periodic signals into the sum of a (possibly infinite) set of simple oscillating functions, namely sines and cosines

Approximation

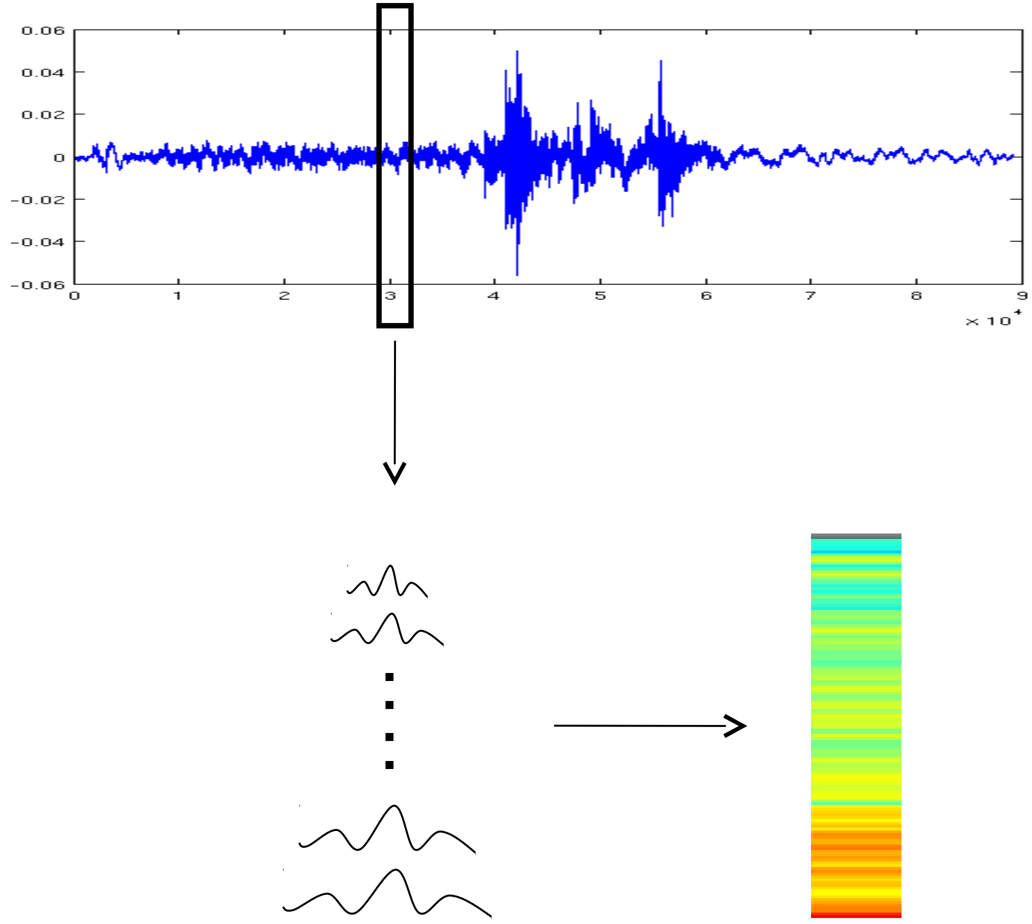


[http://en.wikipedia.org/wiki/Fourier_series]

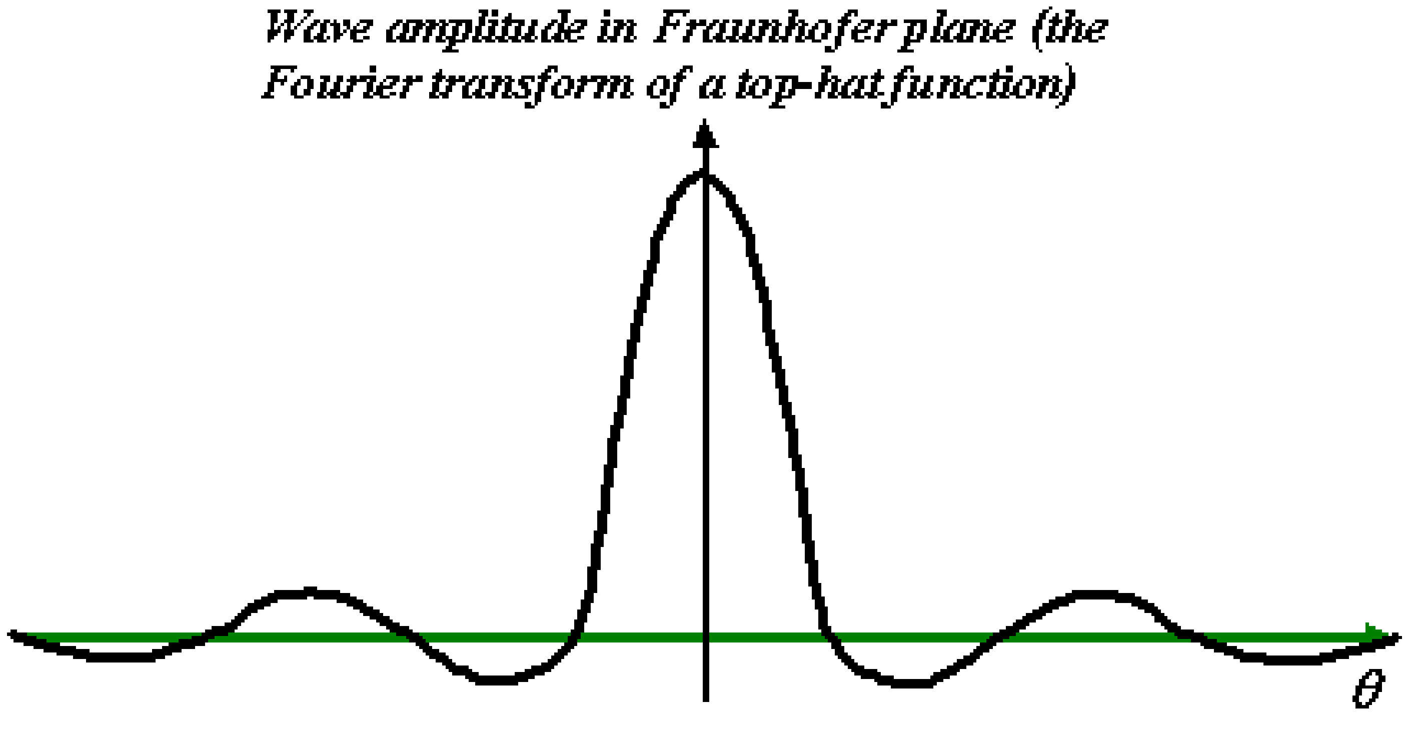
Approximation



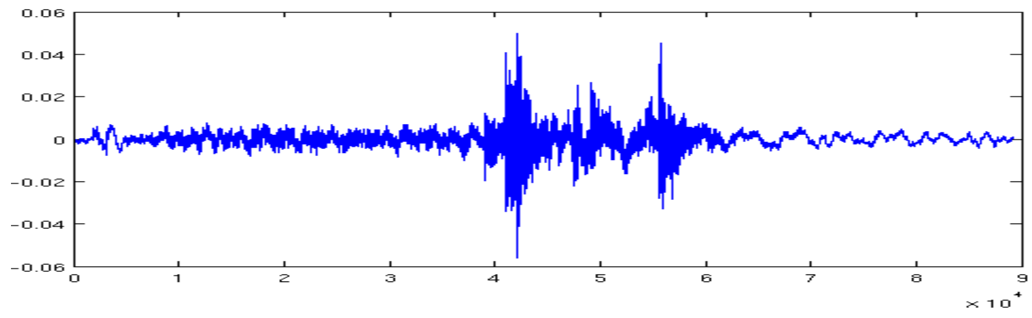
Discrete Fourier Transform

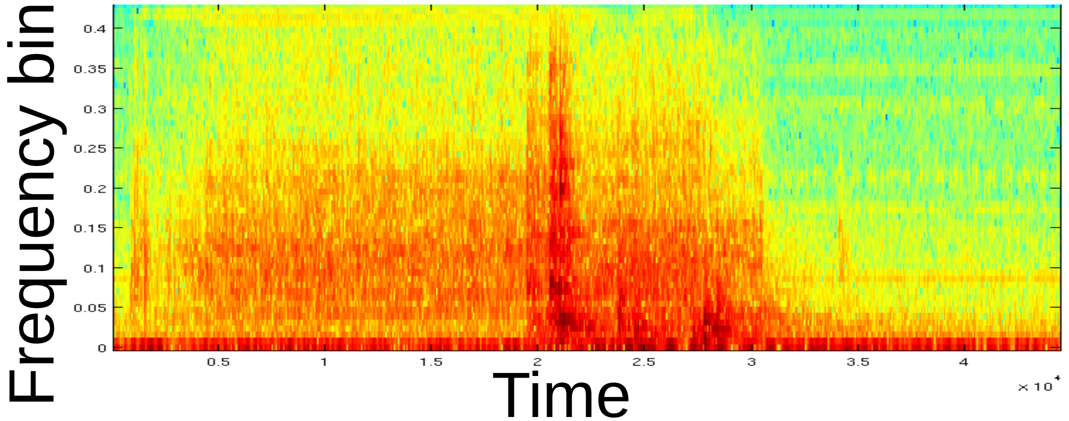


Discrete Fourier Transform



Discrete Fourier Transform





Research Question

Can the DFT be used by a robot to perceive objects and their properties using sound?

Research Question

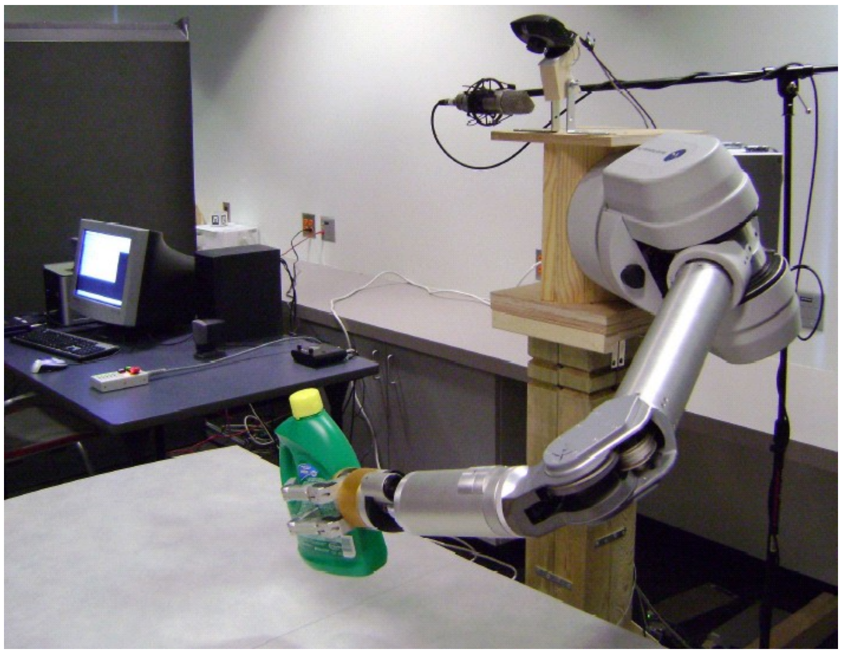
Can the DFT be used by a robot to perceive objects and their properties using sound?

How should the robot associate a particular sound with an object?

Object Exploration by a Robot

Object Exploration by a Robot

Objects



[Sinapov, Weimer, and Stoytchev, ICRA 2009]

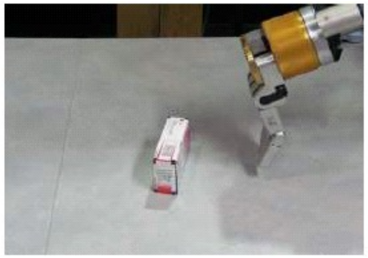
Behaviors

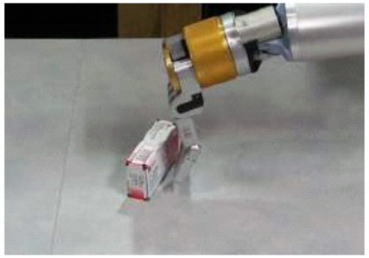
Grasp:

Push:

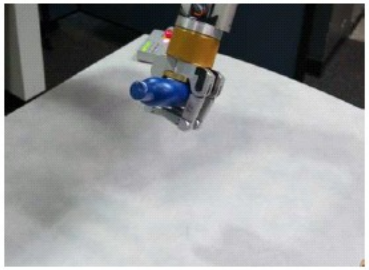
Shake:

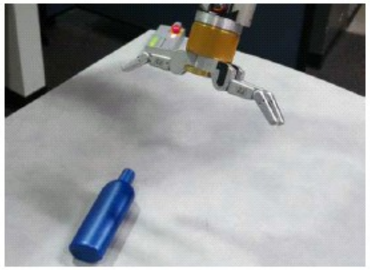
Tap:



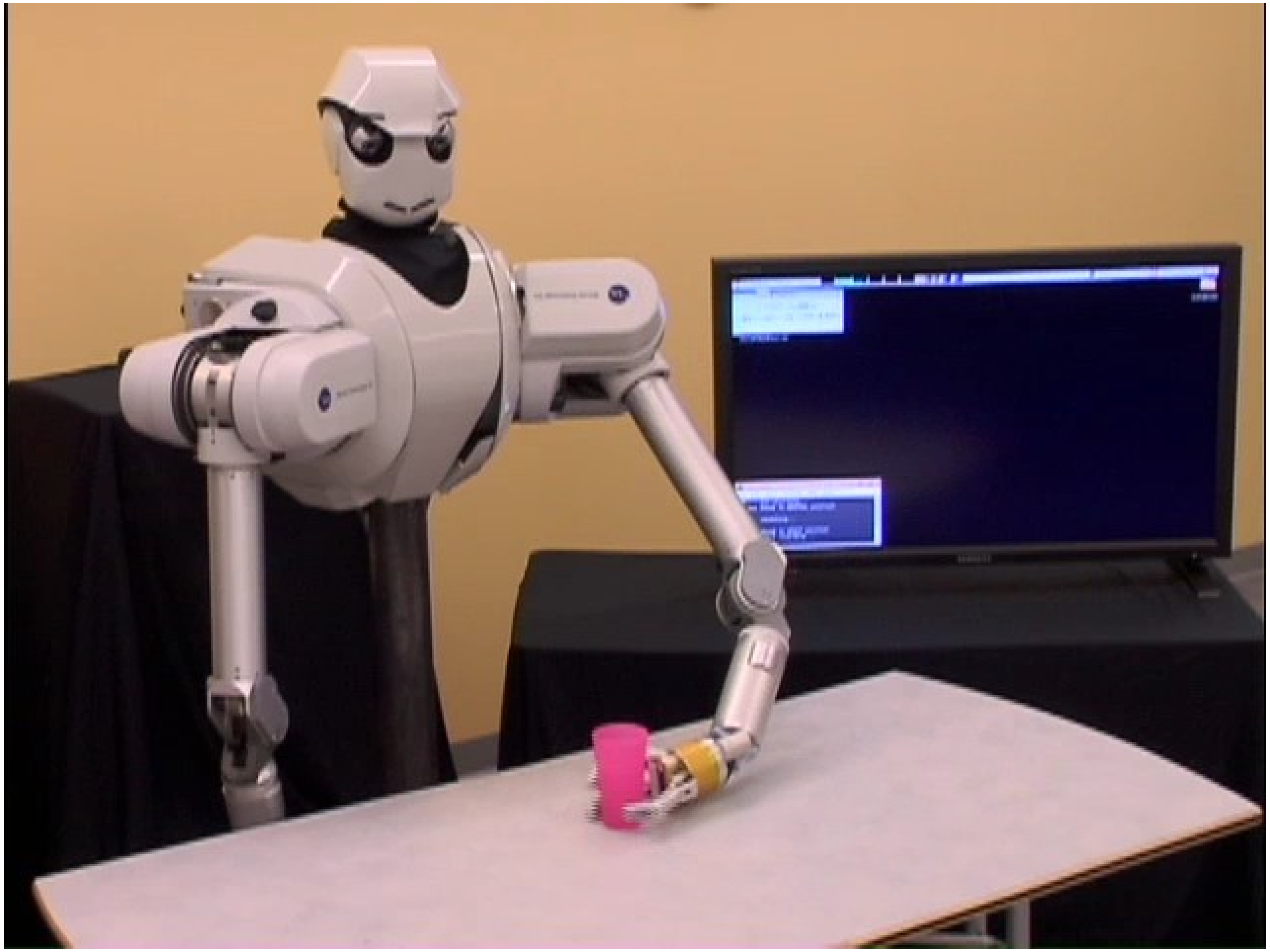


Drop:





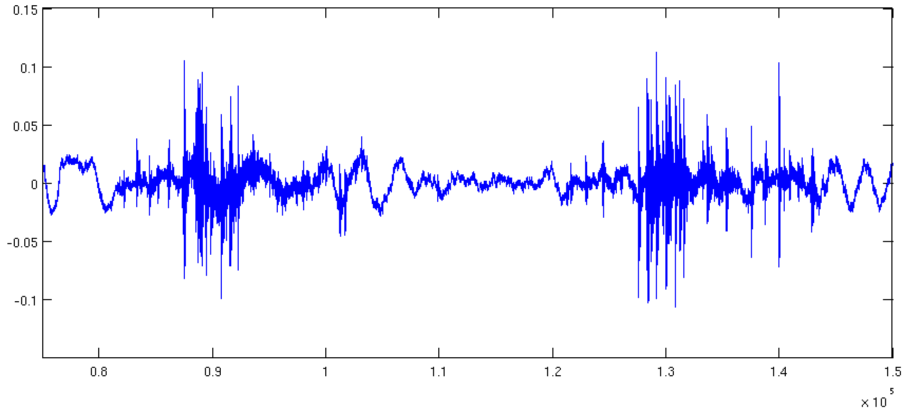
Recognition Video



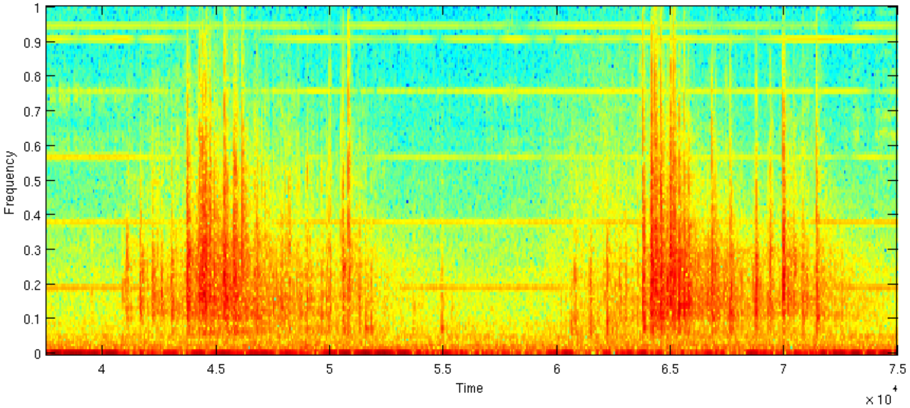
Audio Feature Extraction

Behavior Execution:

WAV file recorded:

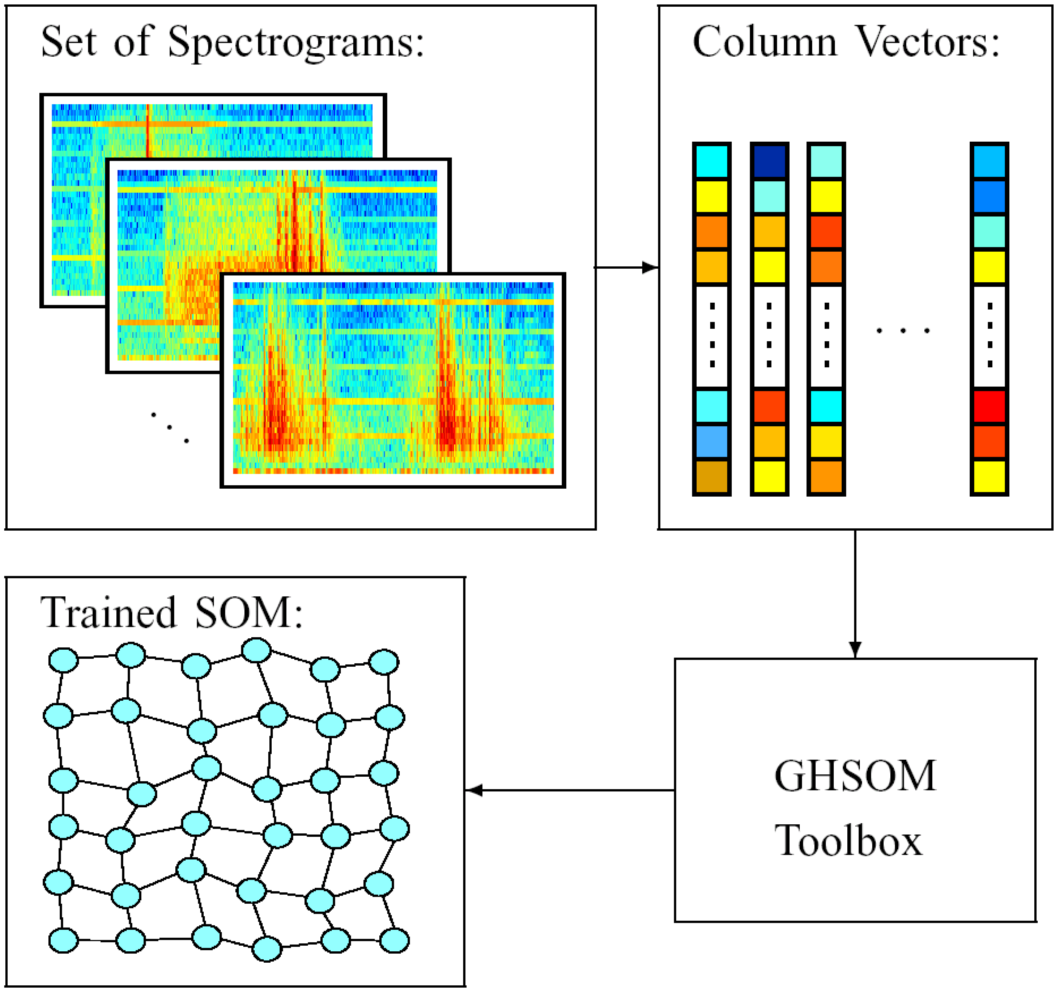


Discrete Fourier Transform:



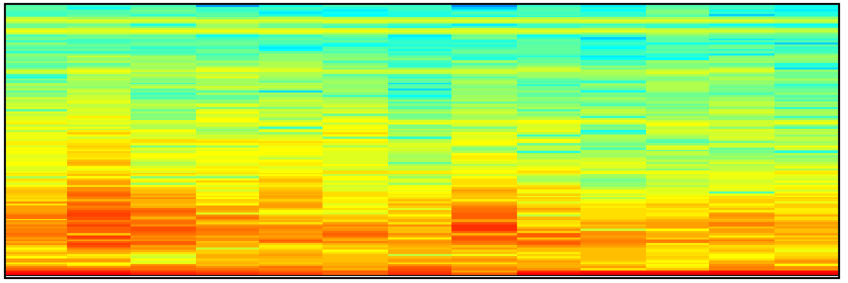
Audio Feature Extraction

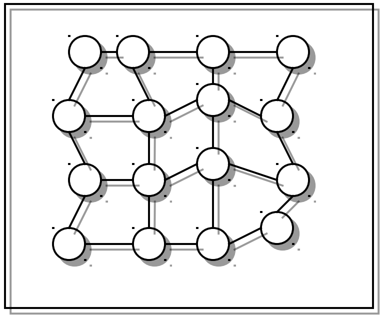
1. Training a self-organizing map (SOM) using DFT column vectors:



Audio Feature Extraction

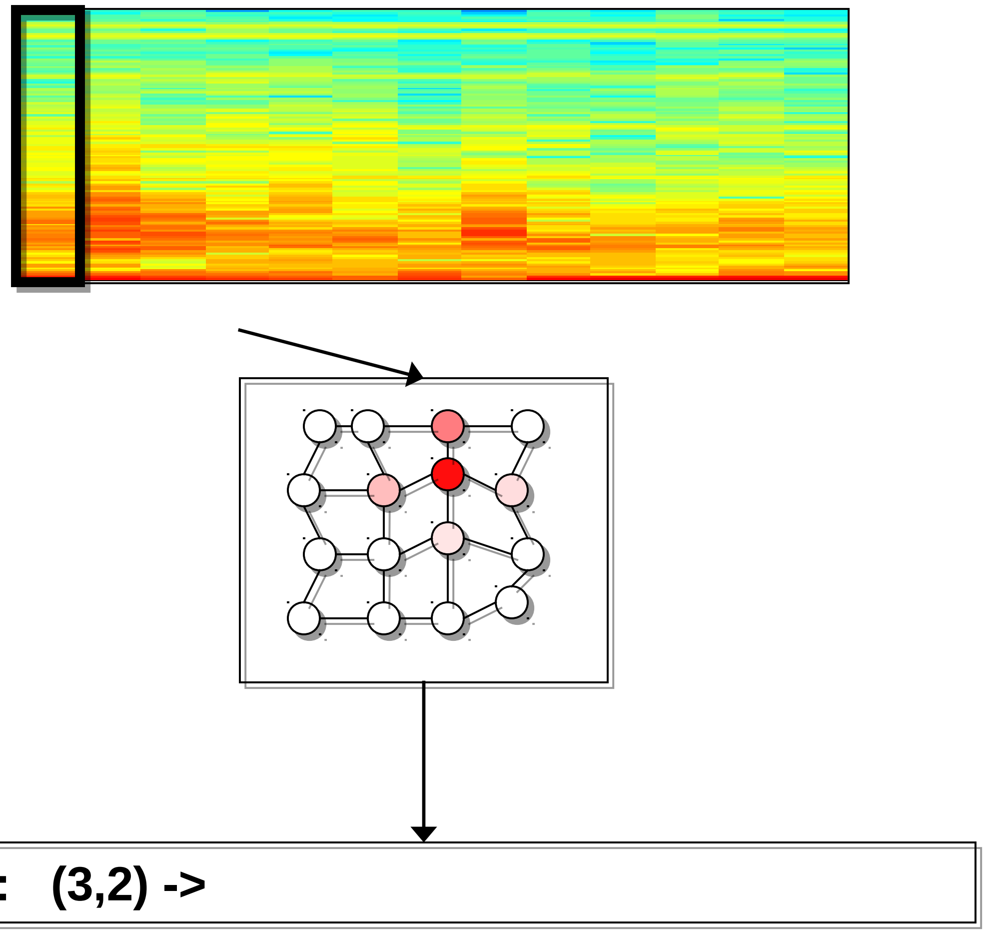
2. Use SOM to convert DFT spectrogram to a sequence:





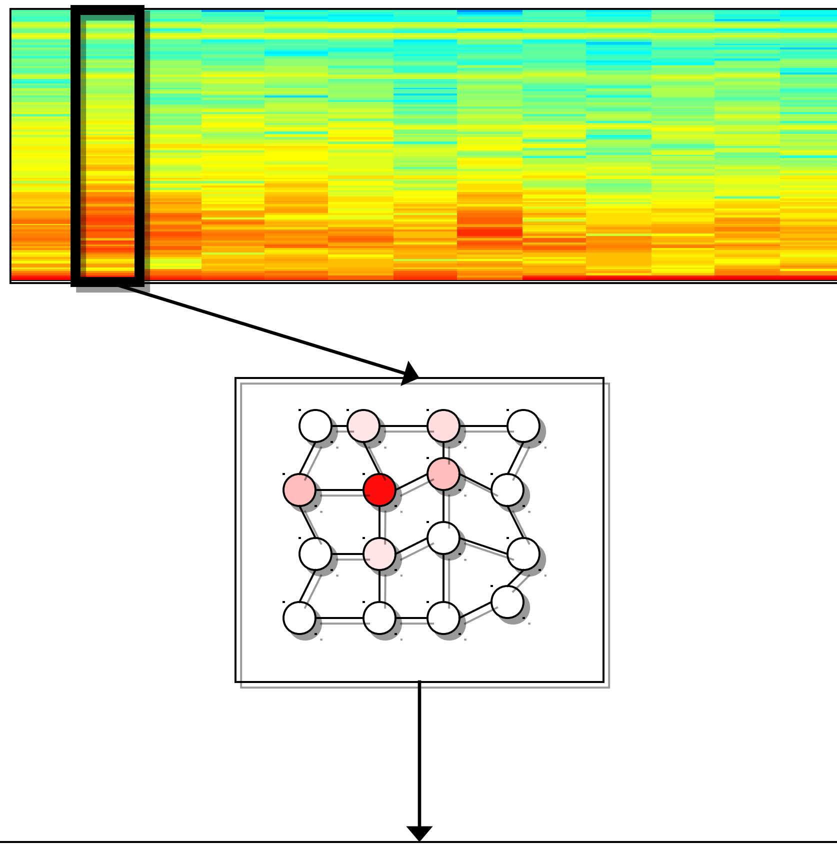
Audio Feature Extraction

2. Use SOM to convert DFT spectrogram to a sequence:



Audio Feature Extraction

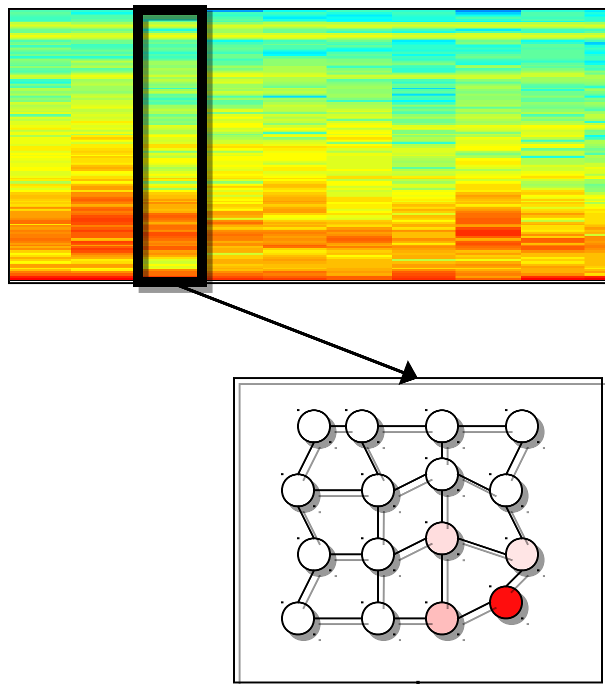
2. Use SOM to convert DFT spectrogram to a sequence:



S_i: (3,2) -> (2,2) ->

Audio Feature Extraction

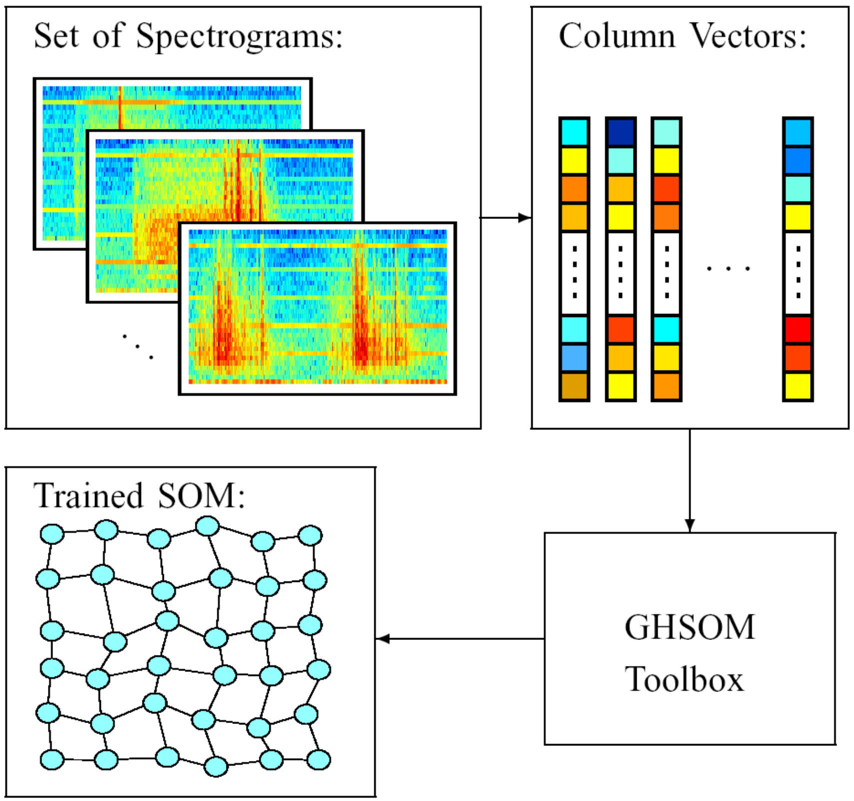
2. Use SOM to convert DFT spectrogram to a sequence:



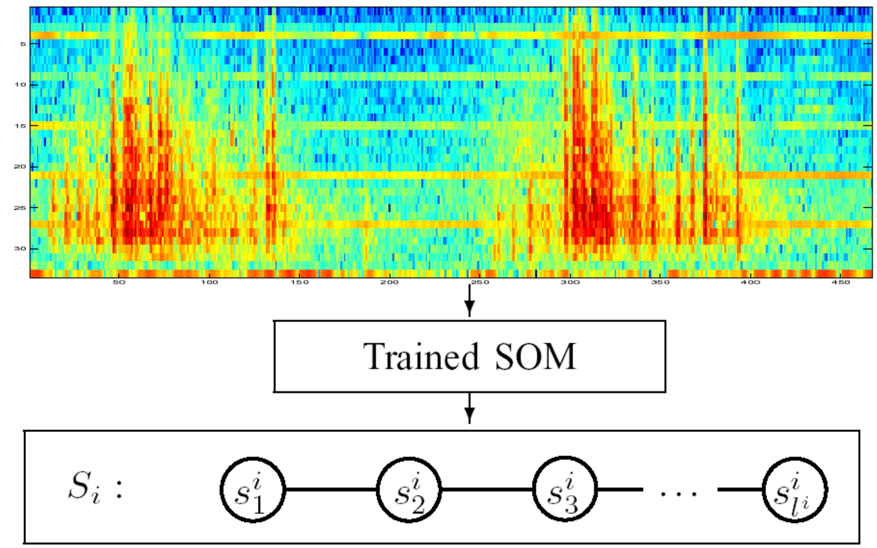
***S_i*: (3,2) -> (2,2) -> (4,4) -> ….**

Audio Feature Extraction

1. Training a self-organizing map (SOM) using column vectors:

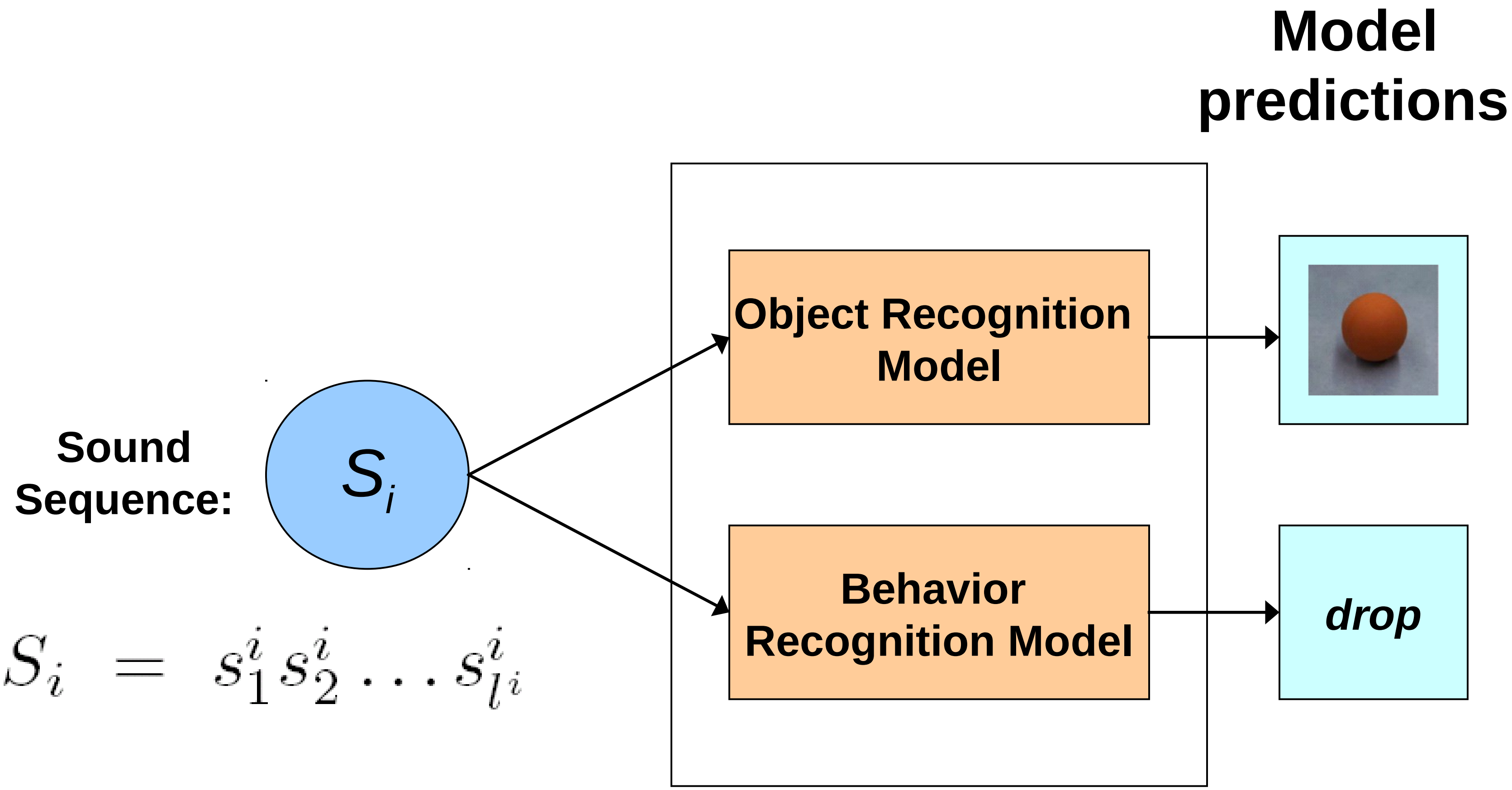


2. Discretization of a DFT of a sound using a trained SOM



$S_i = s_1^i s_2^i \cdots s_{l^i}^i$ is the sequence of activated SOM nodes over the duration of the sound

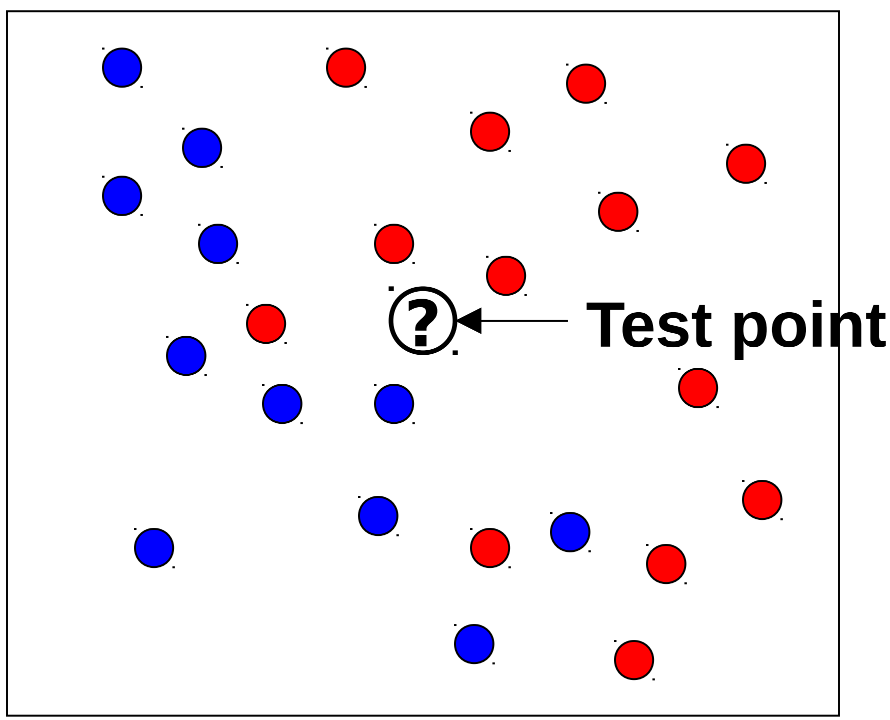
Problem Formulation



$$S_i = s_1^i s_2^i \dots s_{l^i}^i$$

Recognition Model

- k-NN: ***memory-based*** learning algorithm



With k = 3:

2 neighbors

1 neighbors

Therefore,

Pr(red) = 0.66

Pr(blue) = 0.33

Off-Line Evaluation

- 10 trials performed with each of the 36 objects with each of the 5 behaviors
- A total of 1800 interactions, about 12 hours
- 10 fold cross-validation
- Performance Measure for object and behavior recognition:t

$$\%\,Accuracy = \frac{\#\,correct\;predictions}{\#\,total\;predictions} \times 100$$

Evaluation Results

Behavior	k-Nearest Neighbor	Support Vector Machine
Grasp	67.89 %	75.27 %
Shake	49.47 %	50.56 %
Drop	85.79 %	80.56 %
Push	82.89 %	84.44 %
Tap	78.15 %	75.84 %
Average	72.84 %	73.33 %

Chance accuracy = 2.7 %

Evaluation Results

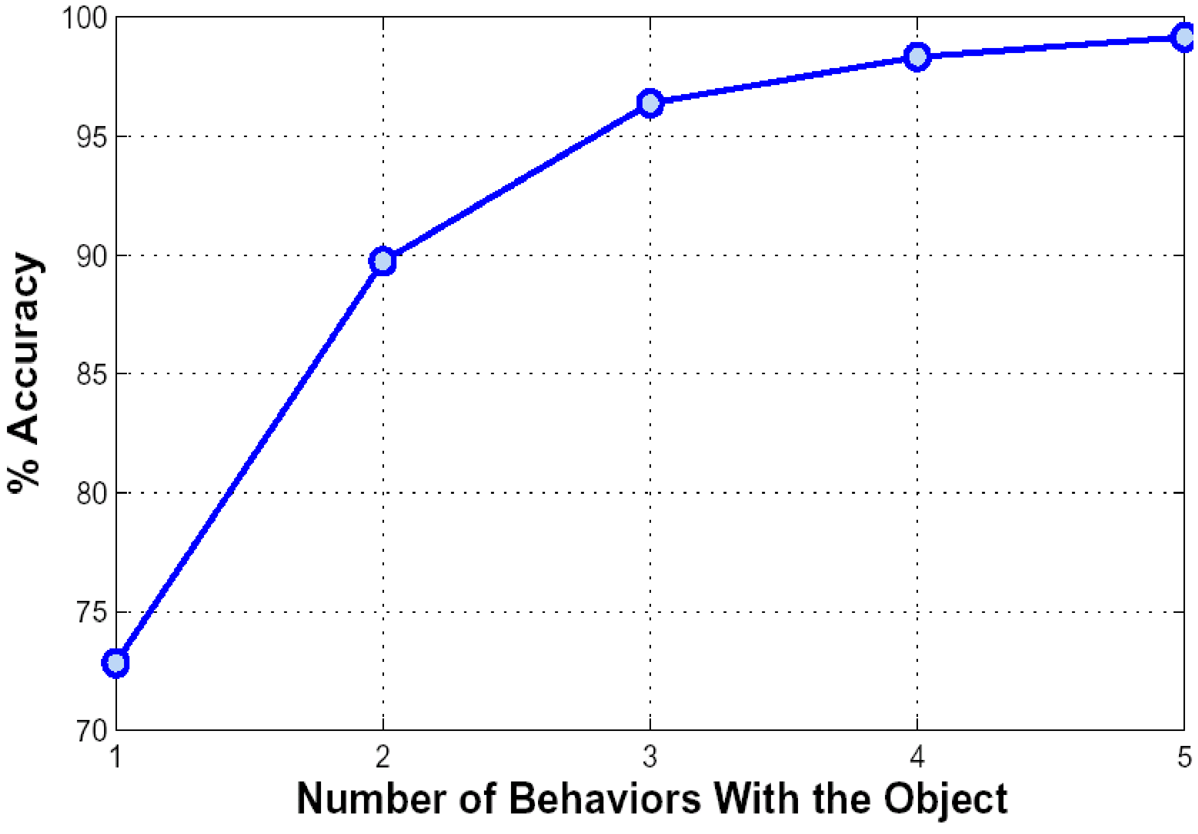
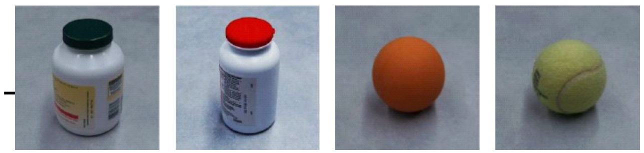


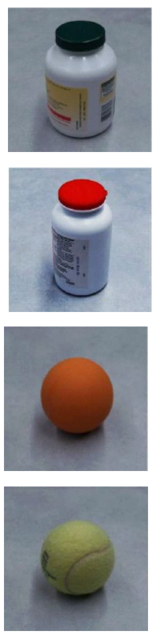
Fig. 6. Object recognition performance with k-Nearest Neighbor as the number of interactions with the object is varied from 1 (the default, used to generate Table I) to 5 (applying all five behaviors on the object).

Estimating Acoustic Object Similarity using Confusion Matrix

Predicted

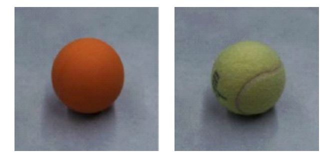


Actual

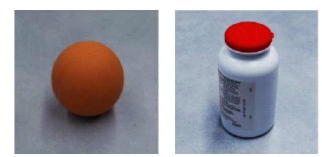


40	**4**	0	0
6	**42**	0	0
0	0	**21**	**6**
0	0	**8**	**35**

: similar



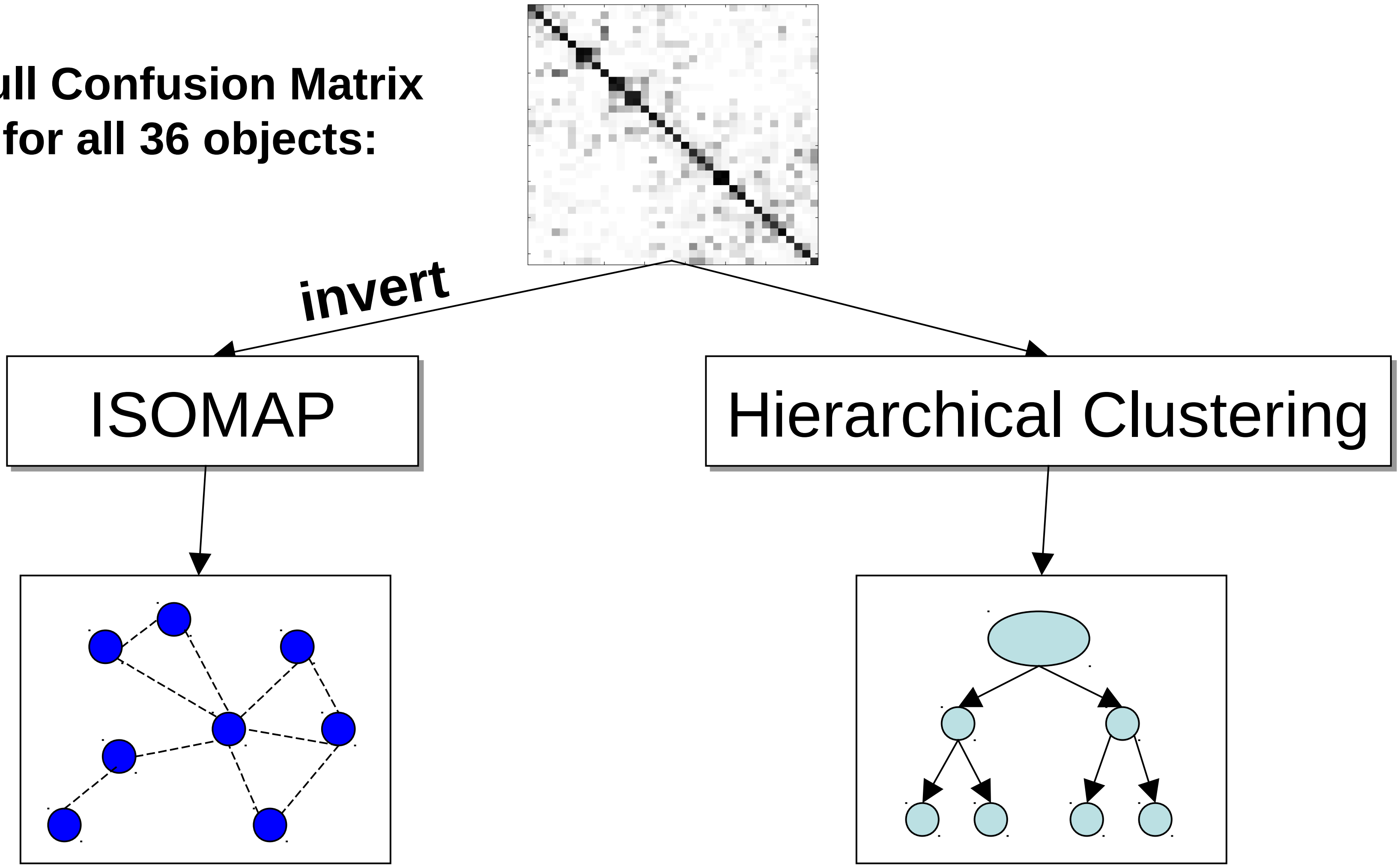
: similar

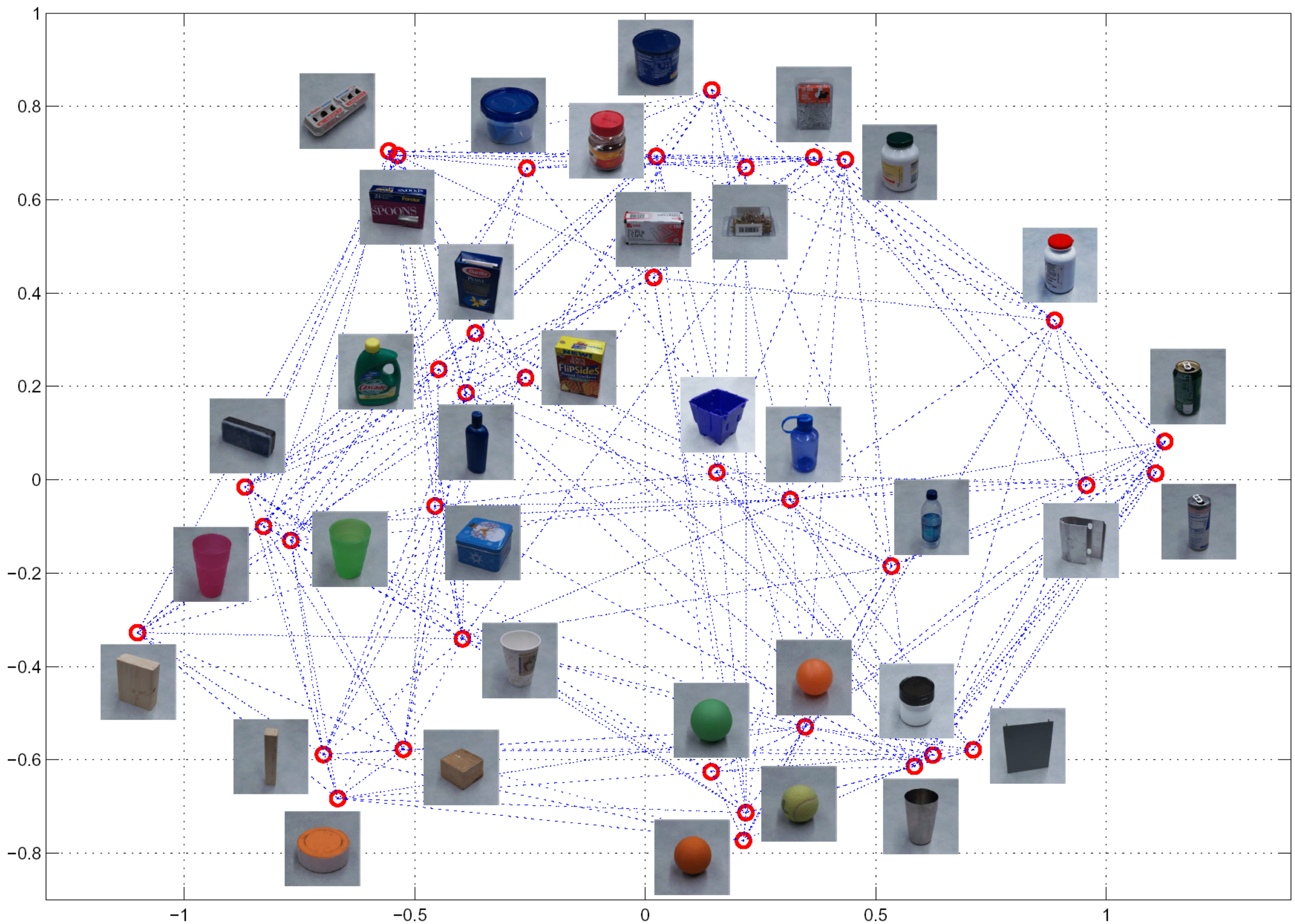


: different

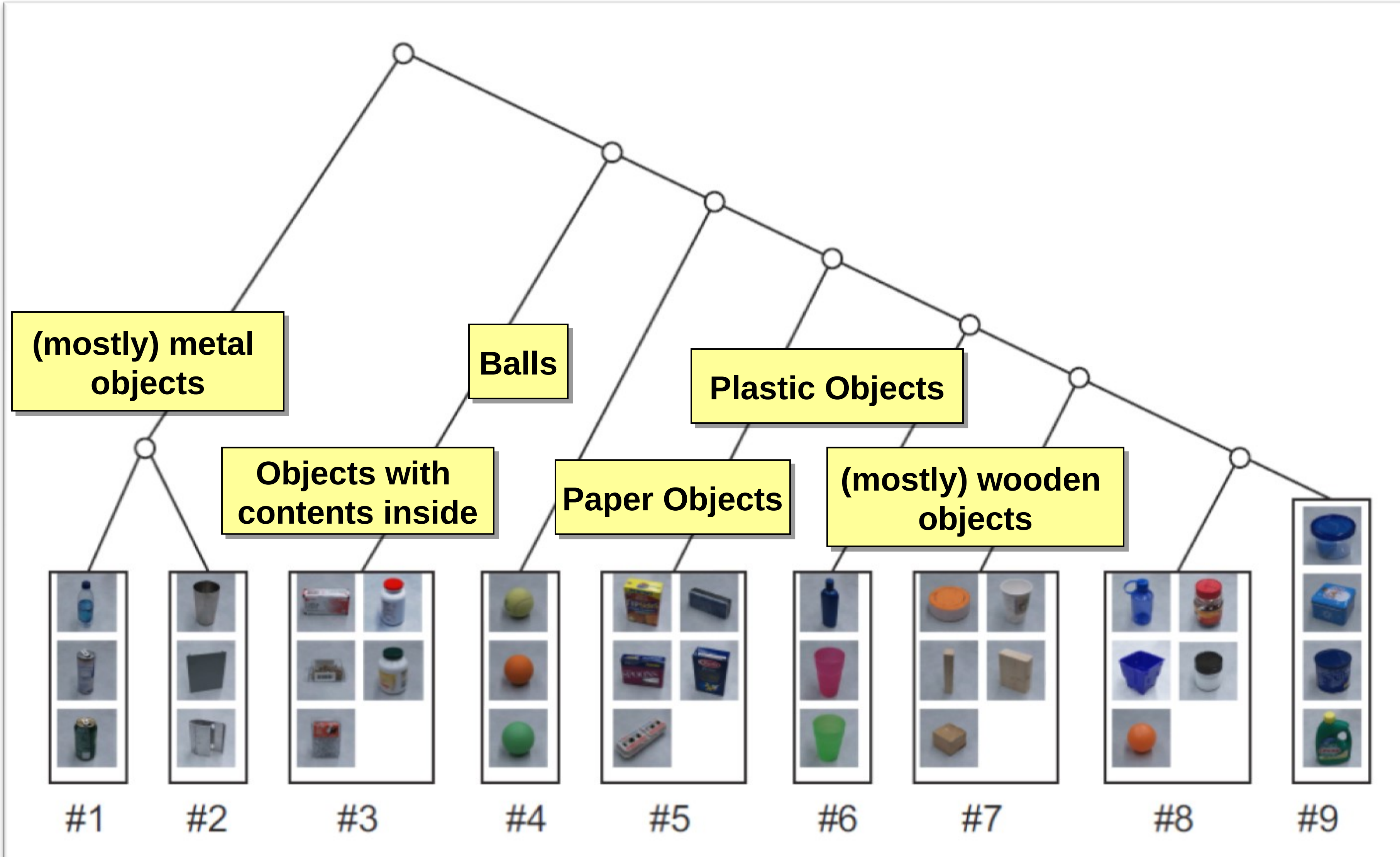
: different

Full Confusion Matrix for all 36 objects:



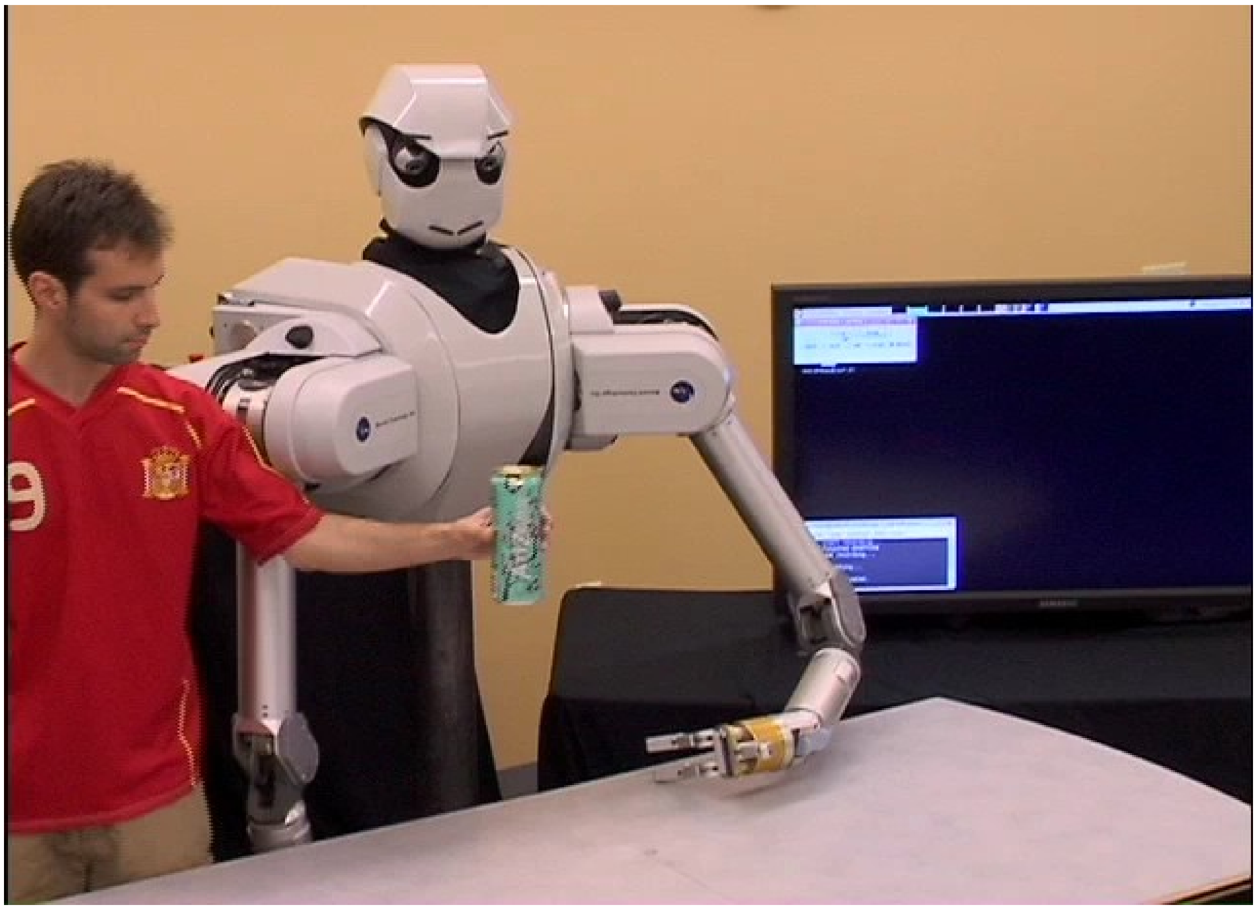


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−0.5
0
0.5
1
SPOONS
FliPsides


(mostly) metal objects
Balls
Objects with contents inside
Plastic Objects
Paper Objects
(mostly) wooden objects
#1
#2
#3
#4
#5
#6
#7
#8
#9

Recognizing the sounds of objects manipulated by other agents

Recognizing the sounds of objects manipulated by other agents



Further Reading

- Sinapov, J., Wiemer, M., and Stoytchev, A. (2008). **Interactive Learning of the Acoustic Properties of Objects by a Robot.** *In proceedings of the "Robot Manipulation: Intelligence in Human Environments" workshop held at the Robotics Science and System Conference*, 2008.
- Sinapov, J., Wiemer, M., and Stoytchev, A. (2009). **Interactive Learning of the Acoustic Properties of Household Objects.** *In proceedings of the 2009 IEEE International Conference on Robotics and Automation* (ICRA).

Discussion

- What kind of sounds should our mobile robots pay attention to?

- What would auditory perception allow them to do that they currently cannot?

THE END