SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Machine Listening

LG전자 고급 빅데이터 전문가 교육과정

서울대학교 융합과학부 음악오디오연구실
이교구

Lecture 1
- Introduction
- Nature of Sound
- Fundamentals of Digital Audio Signal Processing

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Today's Topics

- Introduction
- What is machine listening?
- Nature of sound
- Human hearing
- Fundamentals of digital audio signal processing

Introduction

Instructor

- 이교구
- 융합과학부 지능정보융합전공
- kglee@snu.ac.kr
- http://marg.snu.ac.kr

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Instructor

- **Education**
  - Ph.D., Computer Music and Acoustics, Stanford University (2008)
  - M.S., Electrical Engineering, Stanford University (2007)
  - M.M., Music Technology, New York University (2002)
  - B.S., Electrical Engineering, Seoul National University (1996)
- **Experience**
  - Professor, GSCST, Seoul National University (2009-present)
  - Associate Dean, GSCST, Seoul National University (2016-2018)
  - Senior Researcher, Gracenote Inc. (2007-2009)
- **Research Interests**
  - Machine Listening (a.k.a. Computer Audition)
  - Music Retrieval
  - Source Separation
  - Music Perception/Cognition
  - Recommender Systems

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Lab Instructors

- **김완수**
  - 서울대학교 융합과학부 음악오디오연구실 박사과정
  - 서울대학교 융합과학부 음악오디오연구실 석사
  - KAIST 공학사(기계공학, 기술경영)
  - Research interests: sound event detection/classification, digital audio watermarking

- **이주헌**
  - 서울대학교 융합과학부 음악오디오연구실 박사과정
  - 서울대학교 자유전공학부 공학사(전기정보공학), 인문학사(미학)
  - Research interests: singing voice synthesis, TTS (text-to-speech synthesis)

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Course Outline

- Day 1
  - Introduction
  - What is machine listening?
  - Nature of sound
  - Fundamentals of digital audio signal processing
  - Project guideline
- Day 2
  - Acoustic feature extraction
  - Speech/music analysis

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Course Outline

- Day 3
  - Neural networks
  - Improving machine learning
  - Deep learning applied to audio I (MLP, DNN, SPEECH MNIST)
- Day 4
  - Advanced neural networks
  - Deep Learning Applied to Audio II (CNN, RNN)
  - Real-world Audio Applications
- Day 5
  - Machine listening applications
  - Project presentations

Machine Listening

Universe

SNU Convergence

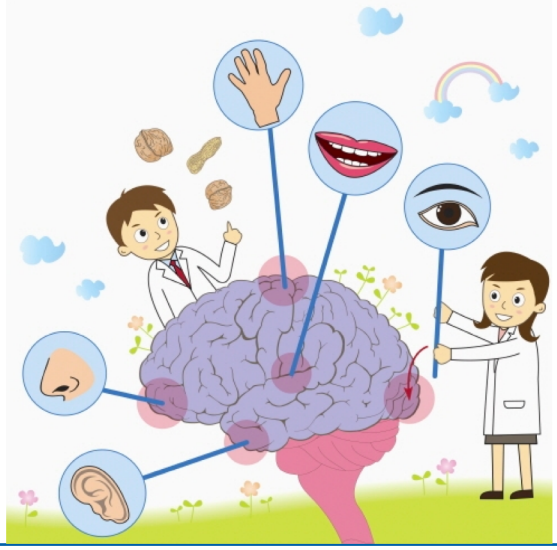
MARG MUSIC & AUDIO RESEARCH GROUP

Our 1.5kg-brain = universe?

SNU Convergence

MARG MUSIC & AUDIO RESEARCH GROUP

Sensation & Perception



Let there be light

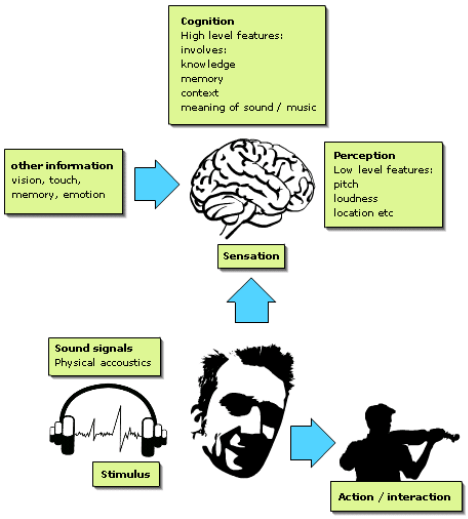
And God said, “Let there be light”.

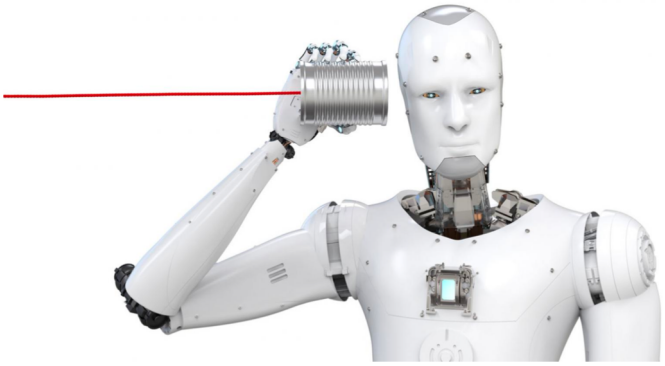
Hearing 청각 聴覚

- Helen Keller (June 27, 1880 – June 1, 1968)
- “*Blindness separates people from things;*

  deafness separates people from people.”

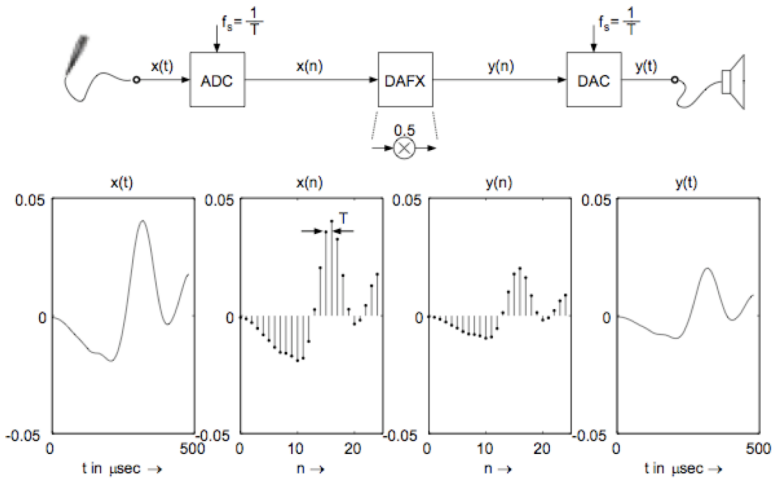
Machine Listening



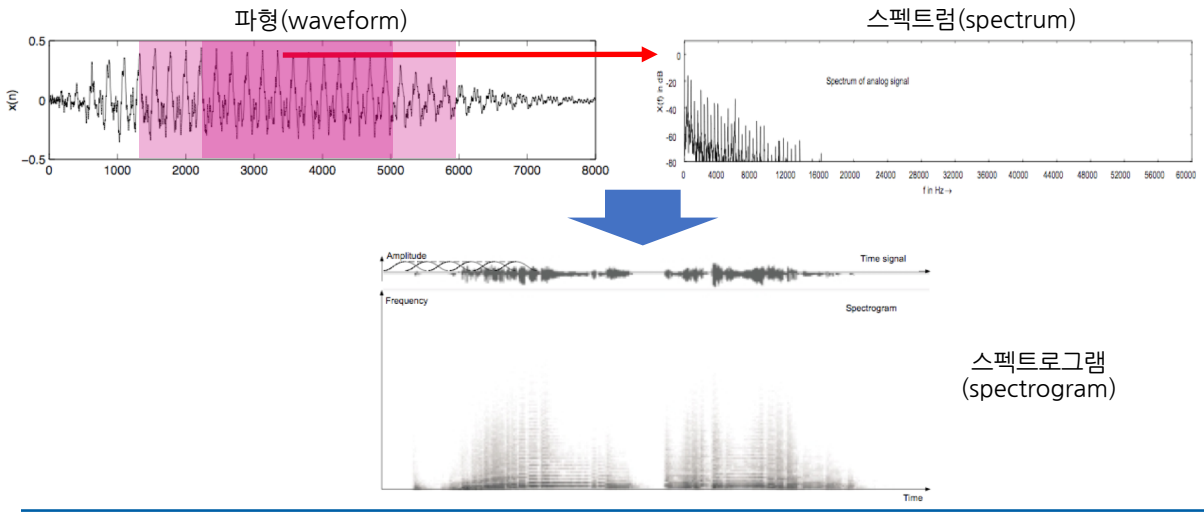


Machine listening (computer audition)

Digital signals

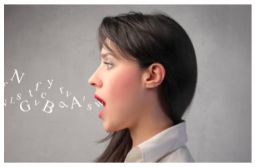


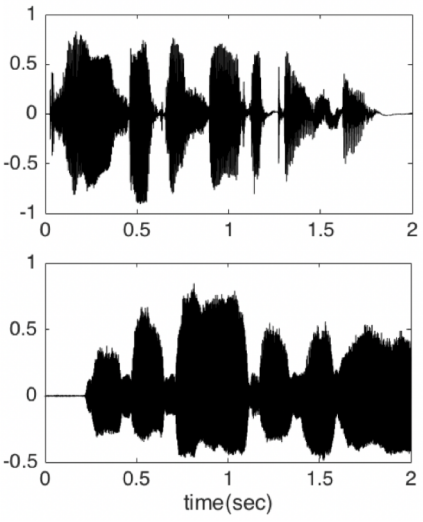
Representations of a sound

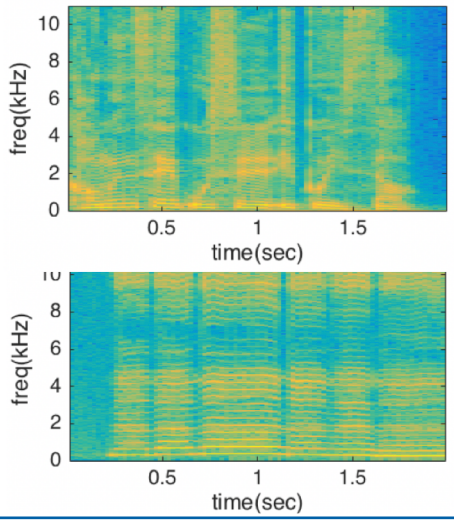


스펙트로그램
(spectrogram)

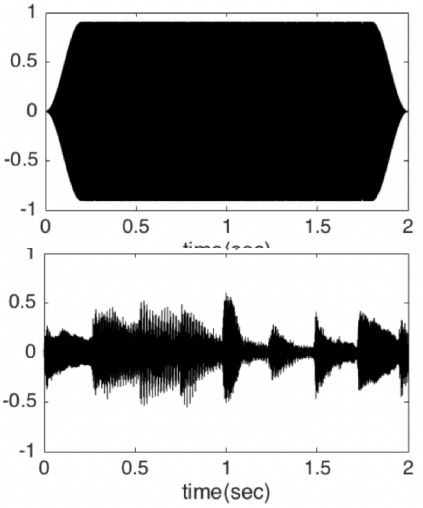
More spectrograms

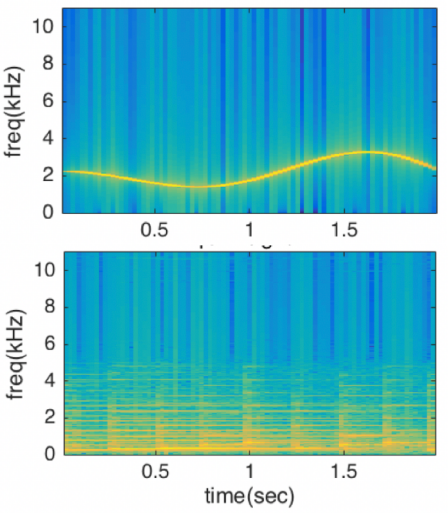




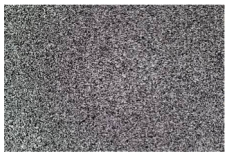


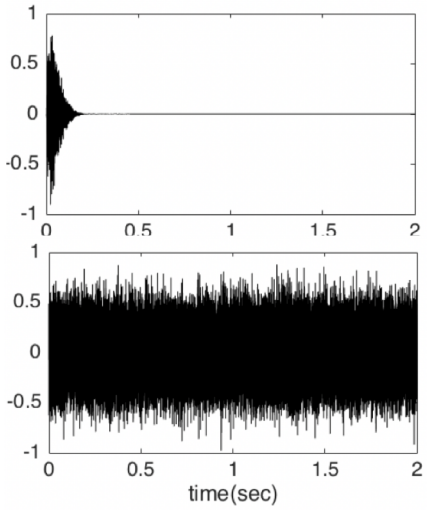
More spectrograms

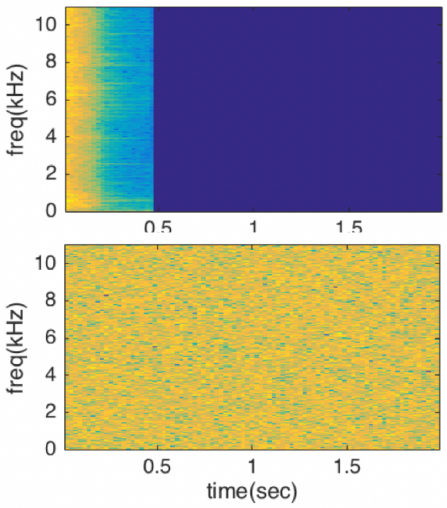




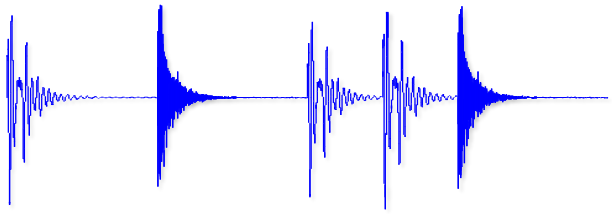
More spectrograms



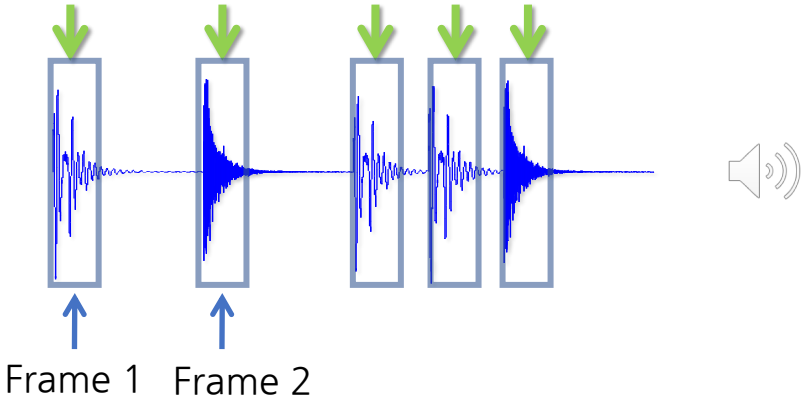




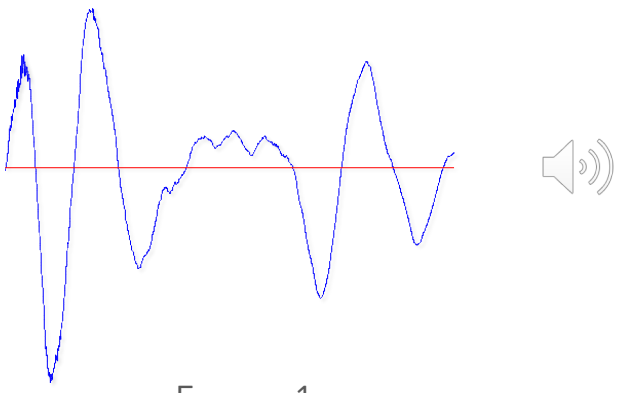
Sound classification/recognition



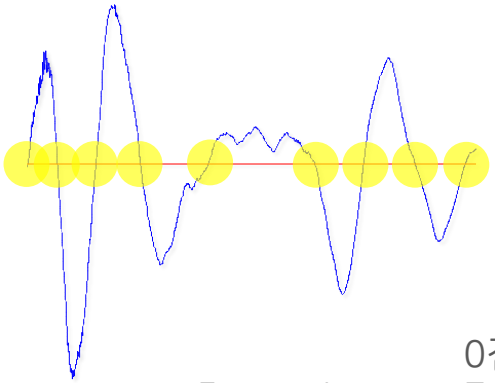
Sound classification/recognition



Sound classification/recognition



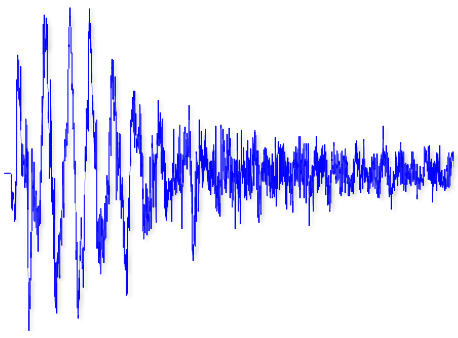
Sound classification/recognition



Frame 1

0점을 지나가는 횟수
Zero crossing rate = 9

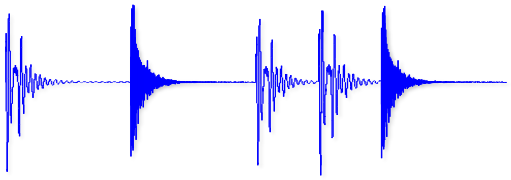
Sound classification/recognition



Frame 2

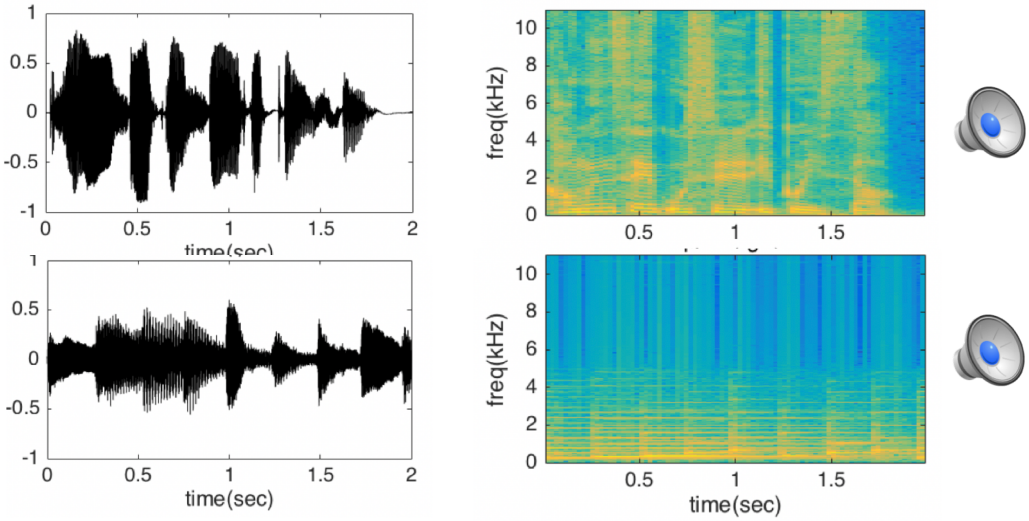
Zero crossing rate = 423

Sound classification/recognition



Frame	ZCR
1	9
2	423
3	22
4	28
5	390

Sound classification/recognition



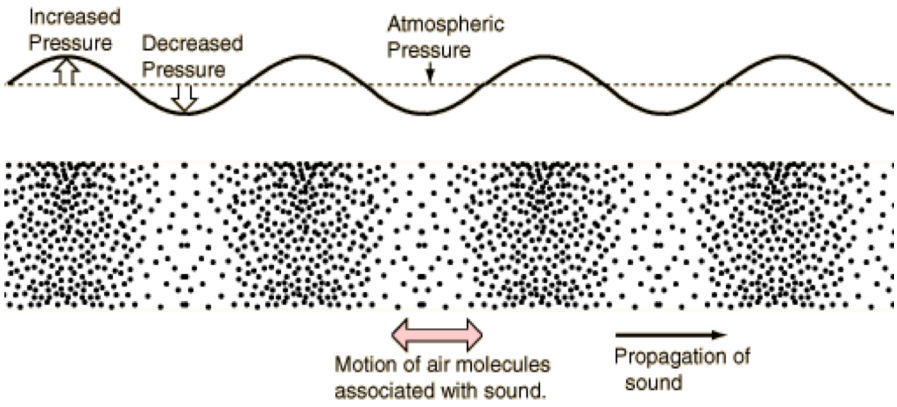
Machine listening applications

- Speech recognition or speech-to-text (STT)
- Sound event detection/classification
- Source separation
- Music identification
- Speech synthesis or text-to-speech (TTS)
- Acoustic biomarkers
- And more…

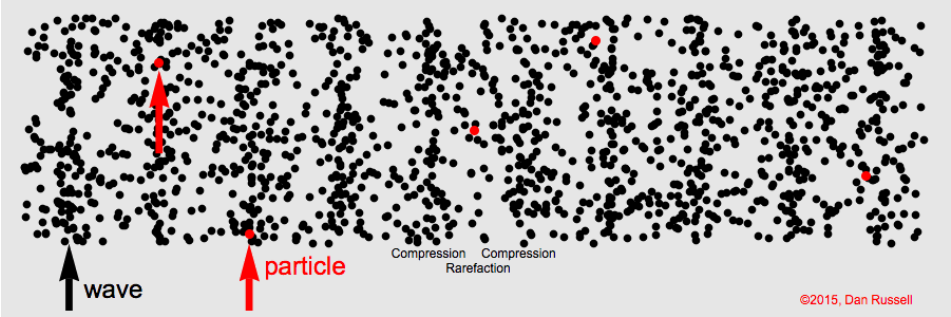
SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

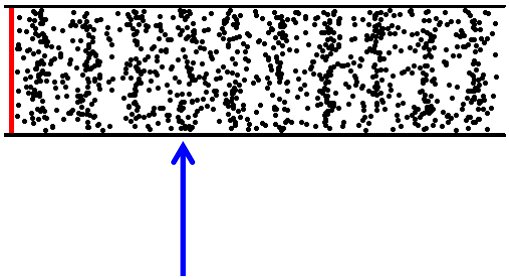
What is Sound?

Sound wave



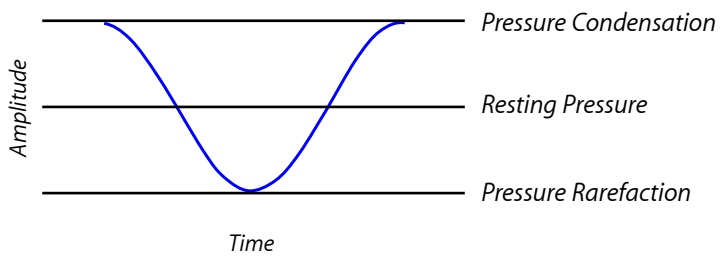
Sound wave





Consider a longitudinal sound pressure wave. Choose a point in space. How do we describe the pressure at that point as a function of time?

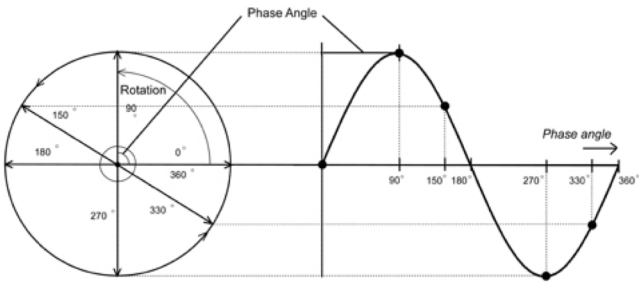
Consider a longitudinal pressure wave. Choose a point in space. How do we describe the pressure at that point as a function of time?

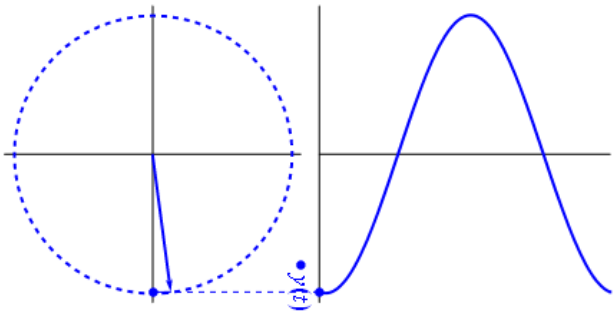


Sinusoids

The simplest sound: sine wave

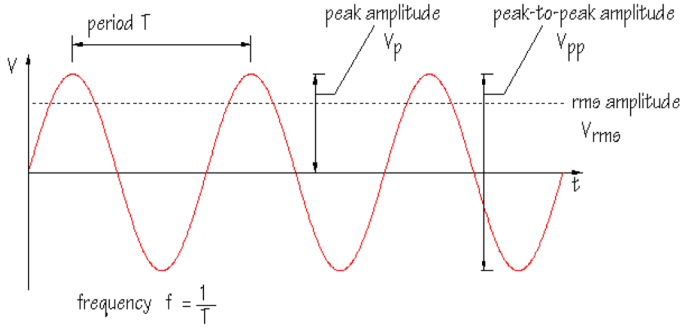
- The Y-coordinate as a function of angle, when a unit circle rotates.
- A cycle = one rotation.
- The period of a wave = the time taken to complete one cycle.





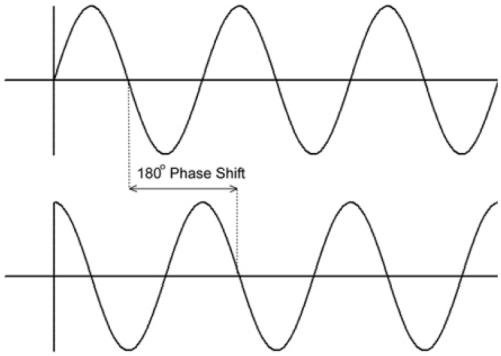
Sinusoid

- A sinusoid is completely determined by:
  - Frequency: 1/period (cycles/sec or Hz)
  - Amplitude: peak size over a period (pressure)
  - Phase: timing relative to t = 0 (radians or cycles)



Phase

- Two sine waves with identical amplitude and frequency that have a phase difference of 90 degrees.

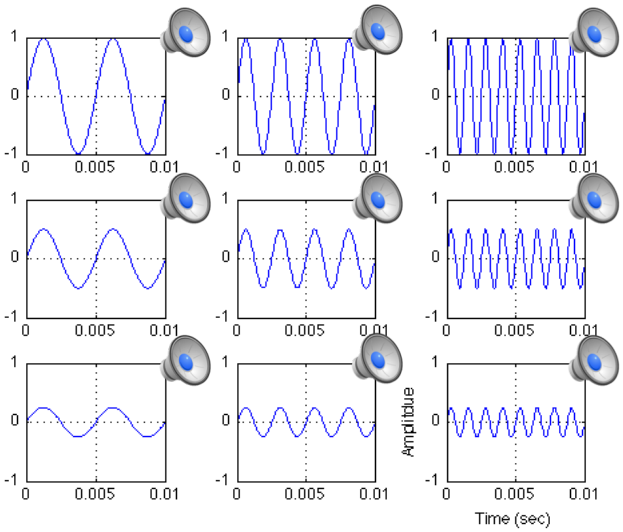


Can you "hear" the difference?

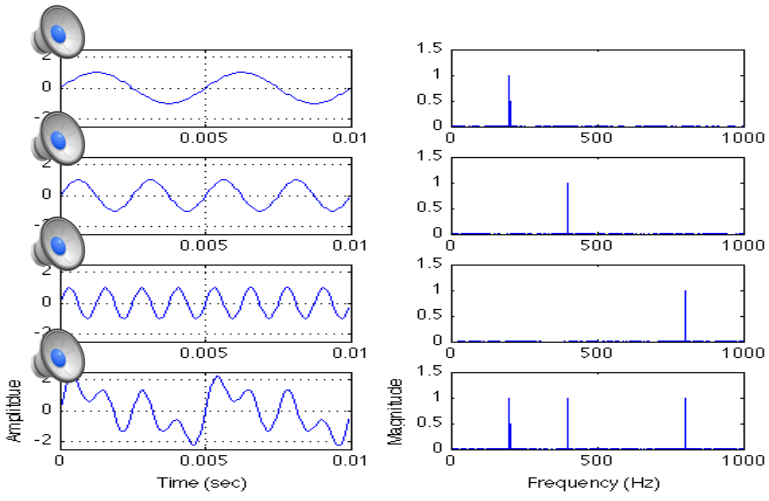
Amplitude, Intensity

- Amplitude: magnitude of pressure variations with respect to atmospheric pressure
  - Peak vs. RMS (root mean square) amplitudes
- Intensity (*I*): sound energy per unit area
  - $I = kP^2$ (W/m^2)
    - *P*: sound pressure
    - *k*: constant
- Sound intensity level (*L*): the level of the intensity of a sound relative to **a reference value (I_0)**
  - Logarithmic scale (decibel or dB)
  - $L = 10\log\left(\frac{I}{I_0}\right) = 10\log\left(\frac{P^2}{P_0^2}\right) = 20\log(\frac{P}{P_0})$ (dB)

Pure tones (sine waves)

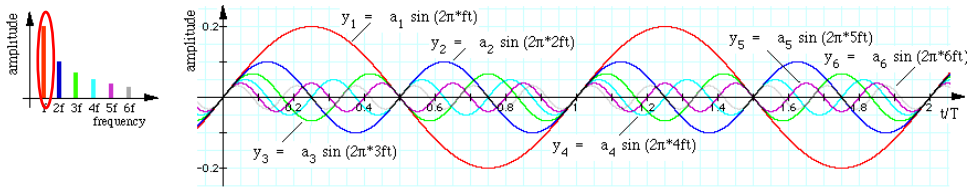


Complex tones

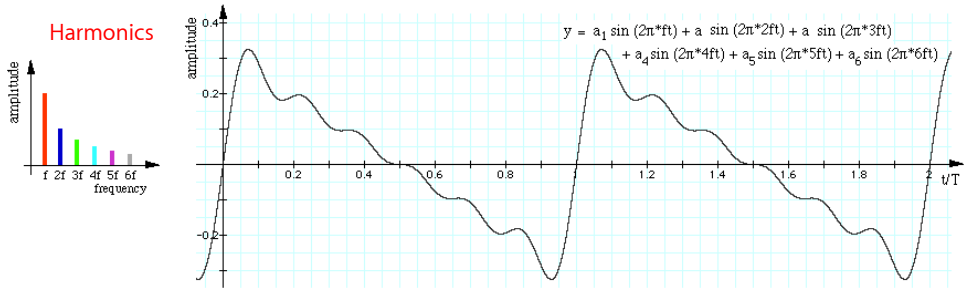


Pure tone vs. complex tone

Fundamental



Harmonics

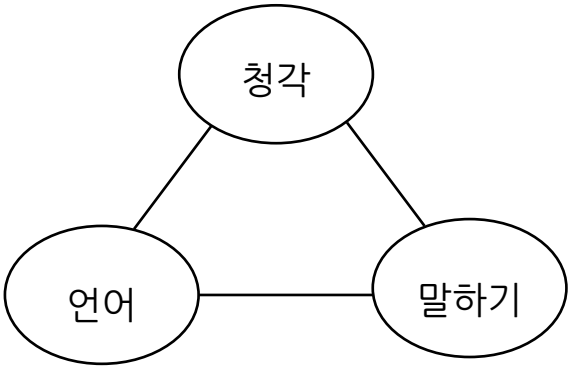


http://www.phys.unsw.edu.au/jw/musical-sounds-musical-instruments.html

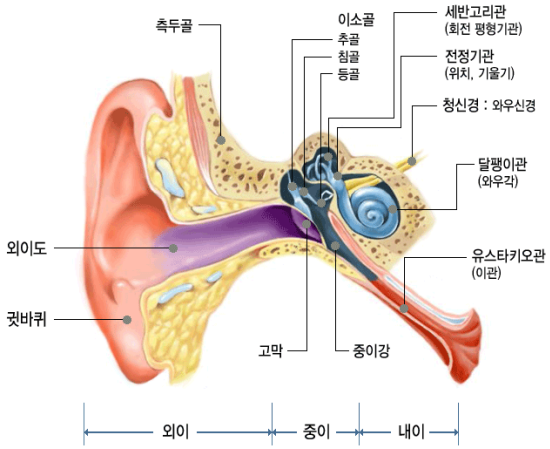
SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Human Hearing

Hearing (청각; 聽覺)

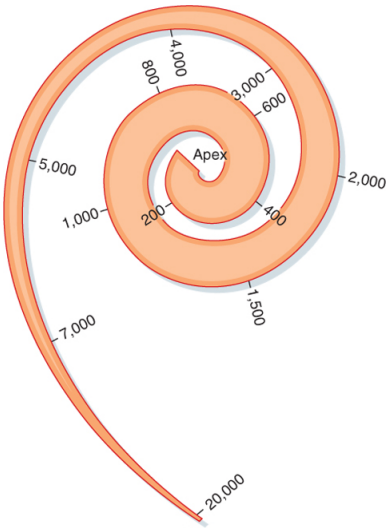


Peripheral auditory system



Localization Transmission Transduction

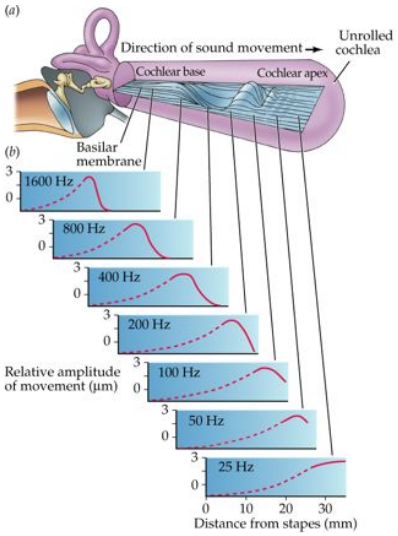
Cochlea



기저막의 폭 :
- 첨부 : 약 0.50mm
- 기저부 : 약 0.04mm

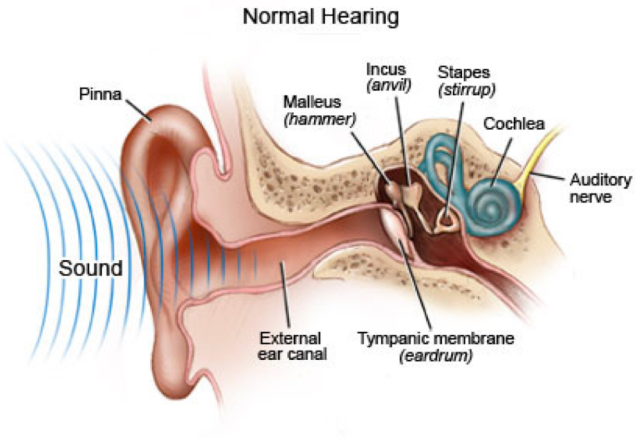
평균 폭 :
- 기저회전 : 약 0.21mm
- 중간회전 : 약 0.34mm
- 첨부회전 : 약 0.36mm

길이 :
- 약 3.2mm



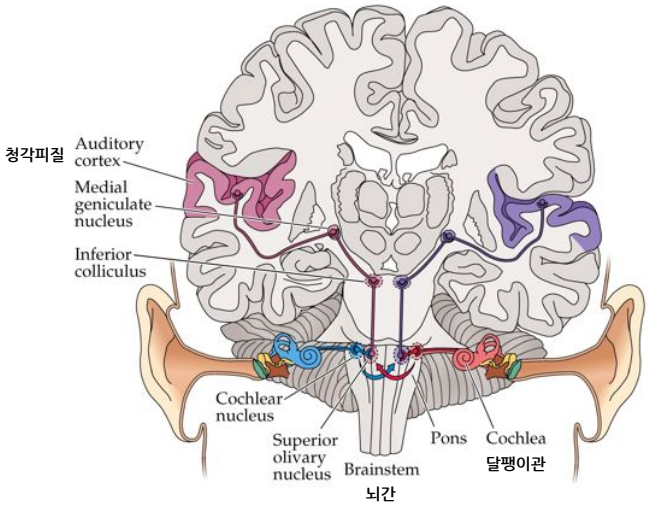
Cochlear mechanics

From external ear to auditory nerve

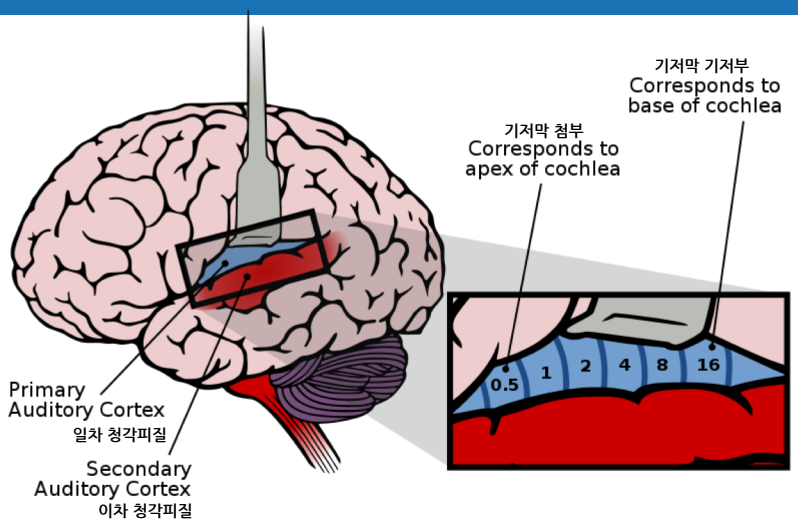


https://www.youtube.com/watch?v=PeTriGTENoc

Central auditory pathways



Auditory Cortex



Frequency range of hearing

- 인간: 20-20,000 Hz
- 개: 40-60,000 Hz
- 쥐: 1,000-70,000 Hz
- 박쥐: 20-150,000 Hz
- 돌고래: 75-150,000 Hz

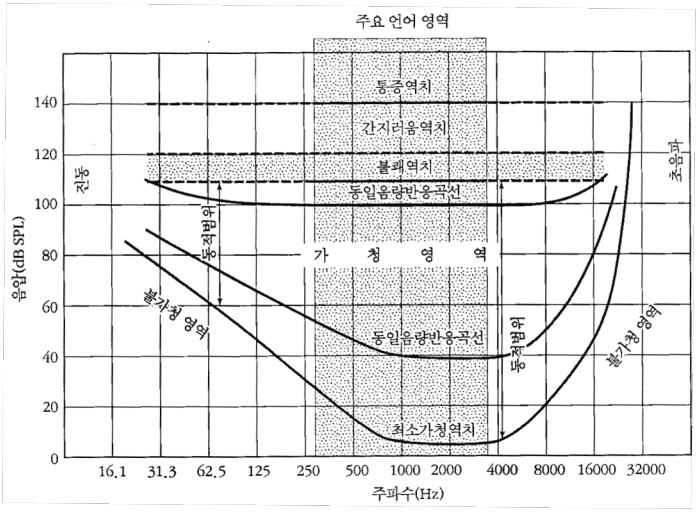
Loudness

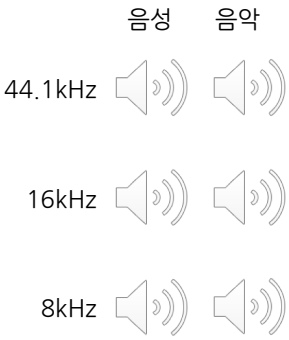
- dB SPL (Sound Pressure Level)
  - 상대적인 소리 압력의 크기를 상용 로그 배율로 표시한 것:

    $dB\ SPL = 20\log(P/P_0)$
  - 기준 P_0 = 0 dB SPL = 0.0002 dyne/cm2 = 20 μPa
- dB HL (Hearing Level)
  - 정상 젊은 성인 연령층(18-30세, 남/여)의 가청 역치의 평균(0 dB HL)을 기준으로 음 강도를 표시
  - 0 dB HL ≒ 7 dB SPL in 1 kHz

Hearing threshold & audible range

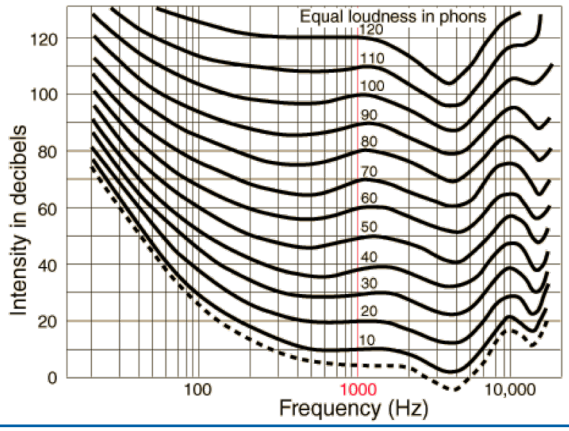
- 역치
- 불쾌역치
- 동적범위
- 가청범위
- 언어영역



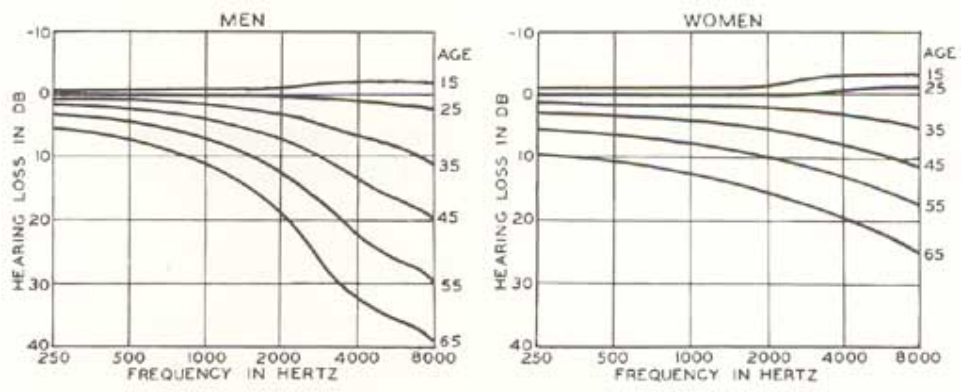


Equal loudness curve

- 물리적 음압이 같아도 주파수가 다르면 음의 강약을 다르게 느낌
- Phon: 1 kHz를 기준으로 같은 크기로 들리는 점들을 연결



Hearing loss due to age



Spectral Analysis

Fourier theorem

- Fourier theorem says: any *periodic* signal can be decomposed into a sum of sinusoids

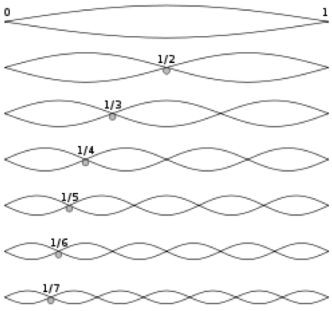
$$x(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(2\pi kft + \phi_k)$$

- f is called *fundamental* and $2f$, $3f$, … are *harmonics* of fundamental
- Sequence of sinusoids with harmonic frequency is *harmonic series*

Jean-Baptiste Joseph Fourier (1768-1830)

Fundamental and harmonics

- Most of the objects in the real world generate a sound with multiple frequency components.
- When the multiple components have a harmonic relationship, like a vibrating string of a guitar, we call the lowest frequency component as "fundamental" and other components as "harmonics."

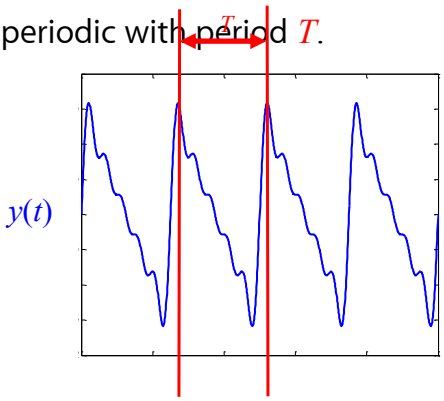


Fourier analysis

- The collage of sine waves to create more complex waves can be expressed through a mathematical system called the Fourier Transform. This is based on two principles:
  1. Any function of time can be represented as a sum of sine waves differing in frequency.
  2. Each component has a distinct frequency, amplitude and phase.
- Using these principles it is then possible to think of the complex wave of a piano string as the sum of many individual sine waves.

Fourier Series

- The function $y(t)$ is periodic with period T.

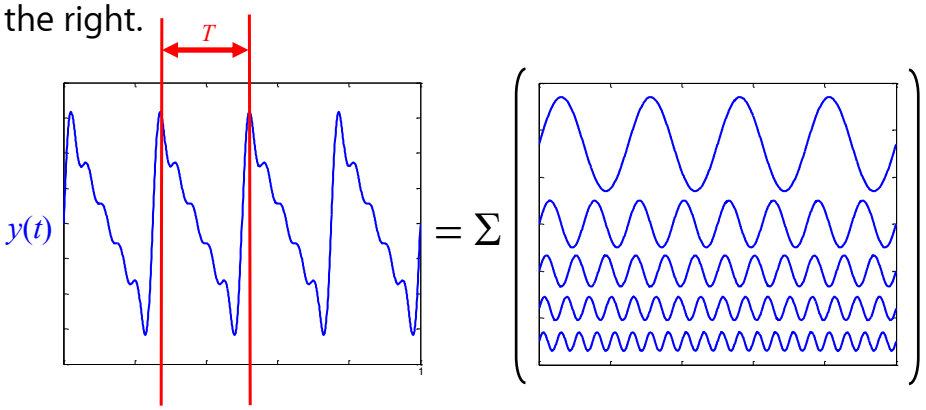


- By Fourier's theorem, $y(t)$ can be expressed as a sum of sine components.

SNU Convergence
MARG MUSIC & AUDIO RESEARCH GROUP

Fourier Series

- The individual sinusoids that were added to produce $y(t)$ are shown to the right.

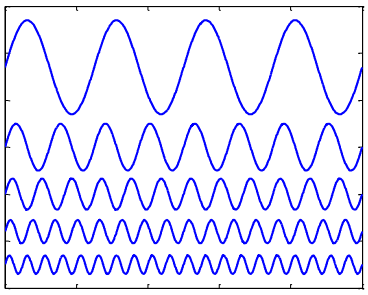


Fourier Series - Frequency

Fundamental → $f_1 = 1/T$

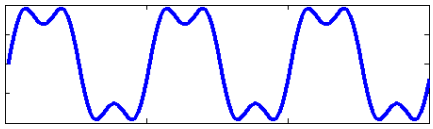
The remaining sinusoids are integer multiples of the fundamental frequency, thus forming a ***harmonic series.***

- $f_2 = 2f_1$
- $f_3 = 3f_1$
- $f_4 = 4f_1$
- $f_5 = 5f_1$

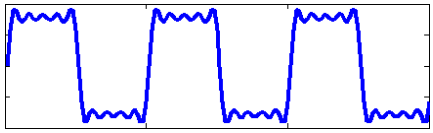


Fourier Series - Frequency

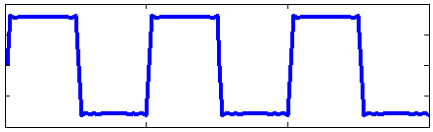
- In practice (on a computer), the series extends to as many terms as needed to make a close approximation.



First 3 harmonics

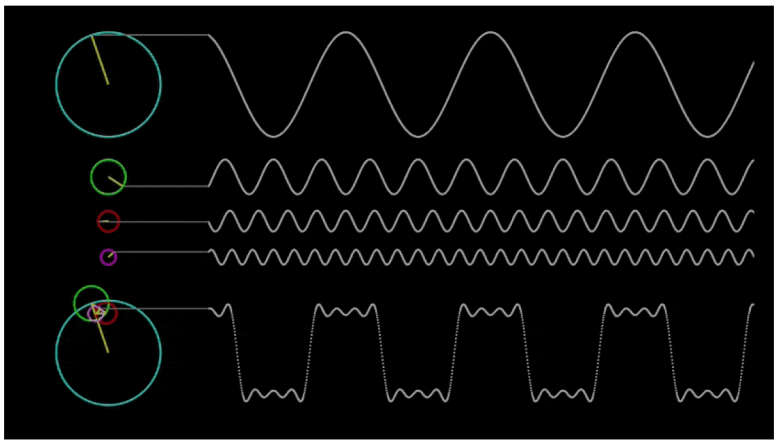


First 9 harmonics

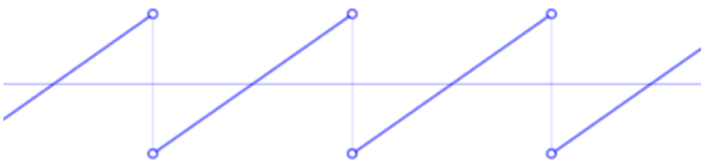


First 81 harmonics

Fourier Series – Square wave



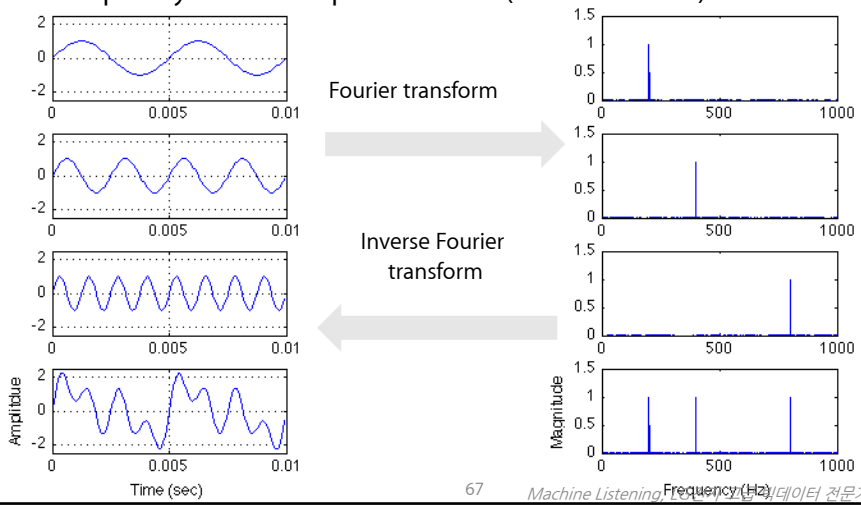
Fourier Series – Sawtooth wave



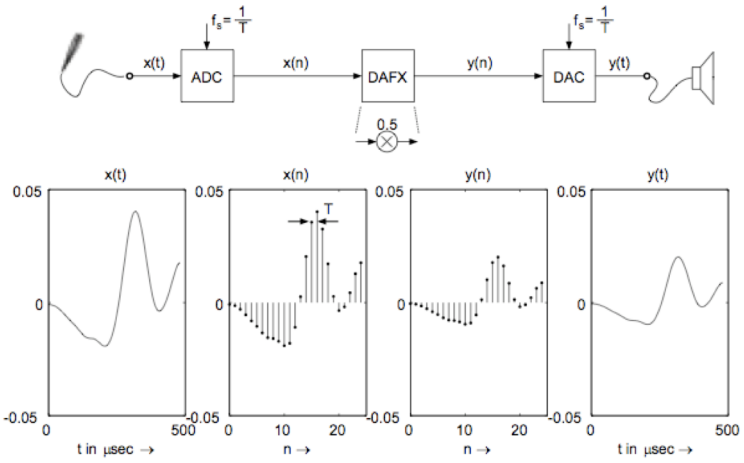
N = 0

Fourier transform

- Mathematical formula that converts a time-domain waveform into a frequency-domain representation (and vice versa)

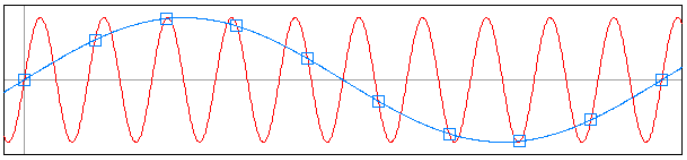


Digital Signals



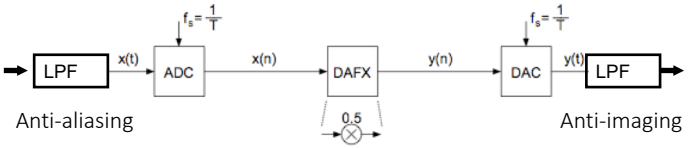
Sampling

- According to the sampling theorem: $f_s > 2f_{max}$ (Nyquist limit)
- Otherwise there is another, lower-frequency, signal that share samples with the original signal (aliasing)



- Related to the wagon-wheel effect: www.michaelbach.de/ot/mot_wagonWheel/index.html

Anti-aliasing[imaging]

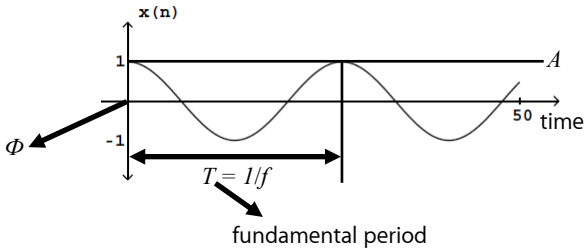


Sinusoid (continuous-time)

amplitude　time (seconds)　frequency (Hz)

$$x(t) = A\cos(\omega t + \phi) = A\cos(2\pi f t + \phi)$$

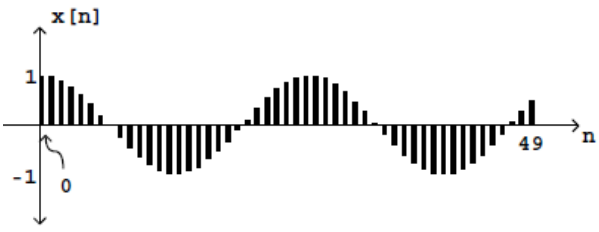
angular frequency　initial phase



Sinusoid (discrete-time)

- Since $t = \frac{n}{R}$, $x(n) = A\cos(\frac{2\pi f n}{R} + \phi)$

where R = sampling rate



Revisiting Fourier Theorem

- Fourier theorem says: any *periodic* signal can be decomposed into a sum of sinusoids

$$x(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(2\pi k f t + \phi_k)$$

- f is called *fundamental* and $2f$, $3f$, … are *harmonics* of fundamental
- Sequence of sinusoids with harmonic frequency is *harmonic series*

Fourier Transform (FT)

- The Fourier transform of a continuous-time signal $x(t)$ may be defined as

$$X(\omega) = FT[x(t)] = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt, \quad \omega \in (-\infty, \infty)$$

Discrete Fourier Transform (DFT)

- The Discrete Fourier Transform of a signal $x(n)$ may be defined as

$$X(k) = DFT[x(n)] = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N}, \quad k = 0,1,\ldots,N-1$$

- The resulting N samples $X(k)$ are complex-valued:

$$X(k) = X_R(k) + jX_I(k)$$

$$|X(k)| = \sqrt{X_R^2(k) + X_I^2(k)} \quad \text{: magnitude}$$

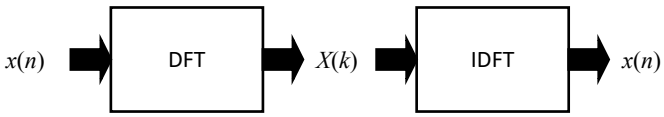
$$\varphi(k) = \arctan \frac{X_I(k)}{X_R(k)} \quad \text{: phase}$$

$$k = 0,1,\ldots,N-1$$

Inverse DFT (IDFT)

- The DFT allows perfect reconstruction of a signal $x(n)$ from its DFT $X(k)$ via inverse DFT defined as:

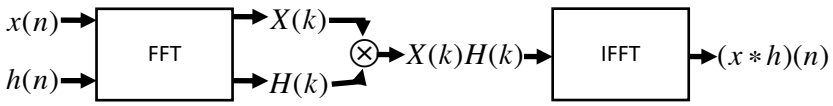
$$x(n) = IDFT[X(k)] = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N}, \quad n = 0,1,\ldots,N-1$$



Fast Fourier Transform (FFT)

- An algorithm to efficiently compute the DFT is known as the Fast Fourier Transform (FFT) and its inverse as the IFFT
- Computational complexity
  - N-point DFT: $O(N^2)$
  - N-point FFT: $O(N \log N)$
- The FFT is so fast that even time-domain operations, like convolution, can be performed faster using FFT and IFFT instead:

$$(x * h)(n) = \sum_{m=0}^{N-1} x(m)h(n-m)$$



Frequency Resolution

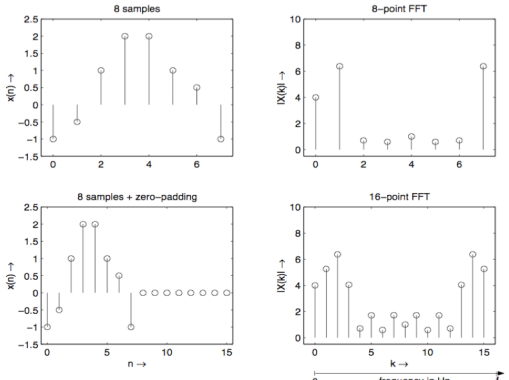
- Determined by how many sinusoids are used to describe a spectrum
- In N-point DFT or FFT, the frequency resolution is given by

$$\Delta f = f_s / N$$

- Intuitively, we can have finer frequency resolution if we increase N
- However, this results in poorer temporal resolution => tradeoff between frequency[spectral] and time[temporal] resolution

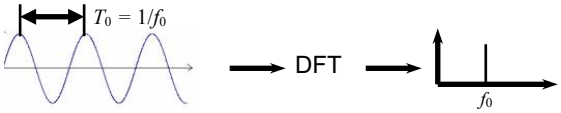
Zero-Padding

- A possible solution to increase frequency resolution without increasing N - *i.e.*, without losing time resolution
- Add zero-valued samples - thus doesn't change spectrum itself - to yield better spectral resolution



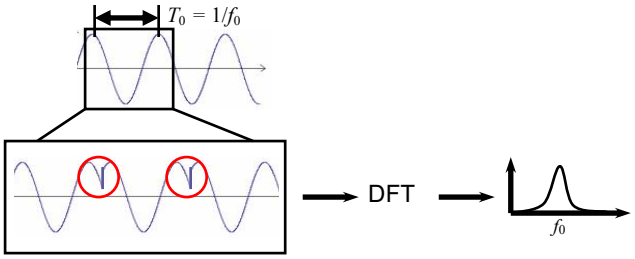
DFT of a Sinusoid

- In theory the DFT of a pure sinusoid results in a single sharp line at the frequency of the sinusoid



Spectral Leaking

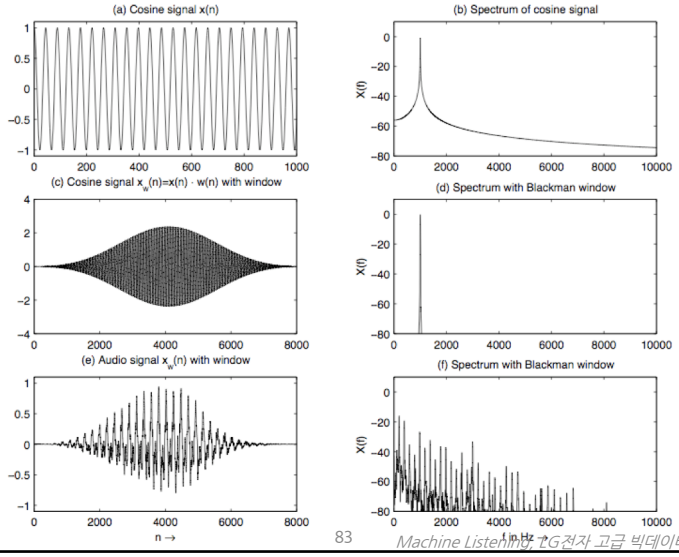
- In practice, unless we perform f_0-synchronous analysis, there are discontinuities (sharp changes) at the segment boundaries that introduce some noise. Thus the spectral line around f_0 is smeared.
- This is known as *spectral leaking*



Windowing

- In order to reduce spectral leaking, we need to avoid abrupt changes between the segment boundaries
- This is done by multiplying a *window* whose amplitude gradually reaches zero at both ends, thus guaranteeing the continuity of a segmented signal when repeated
- Possible windows are: rectangular, hamming, hann, Blackman, Gaussian, and so on.

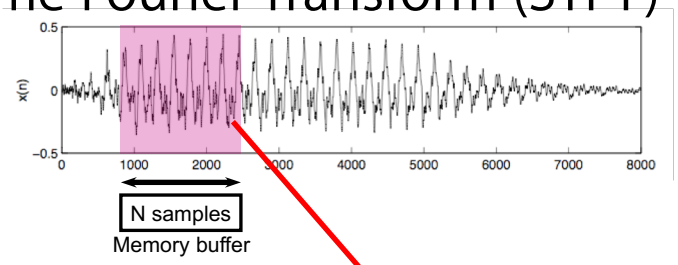
Windowing Example



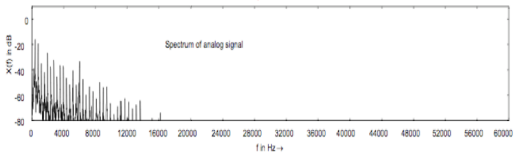
Time-frequency analysis

- Most sounds have changing frequency content over time.
- Using the STFT, independent DFTs are calculated on windowed segments.
- The segments usually overlap to compensate for the loss of temporal resolution.
- Produces a 2-D *spectrogram*.

Short-time Fourier Transform (STFT)

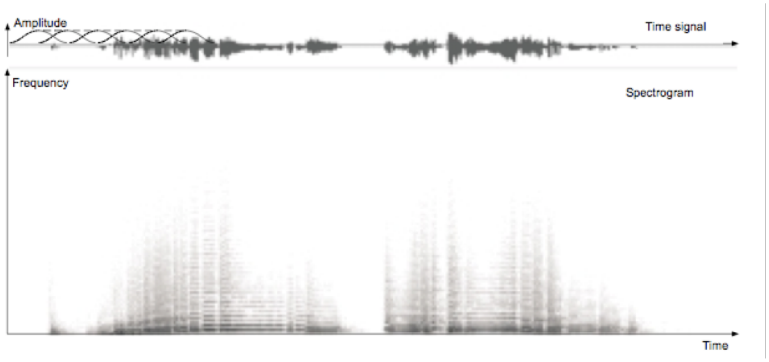


- For block processing, a short segment is sent to a buffer and processed as a block
- The DFT done in this way is called the short-time Fourier Transform or STFT

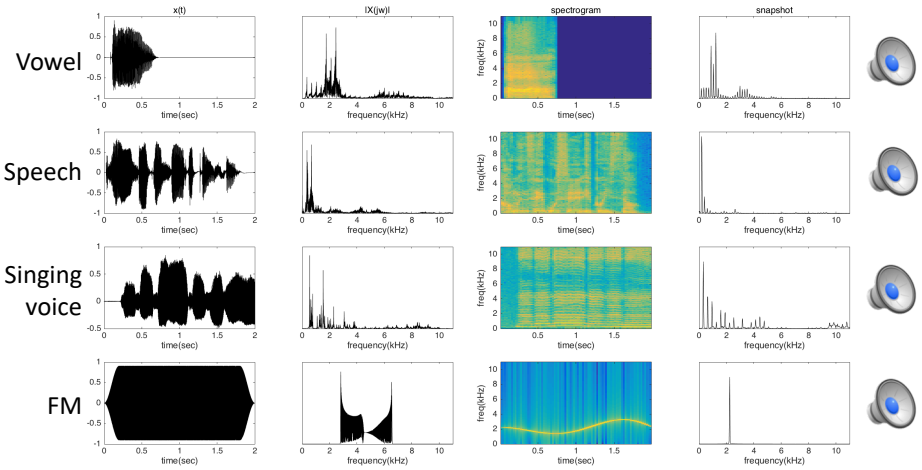


2-D Time-Frequency Representation

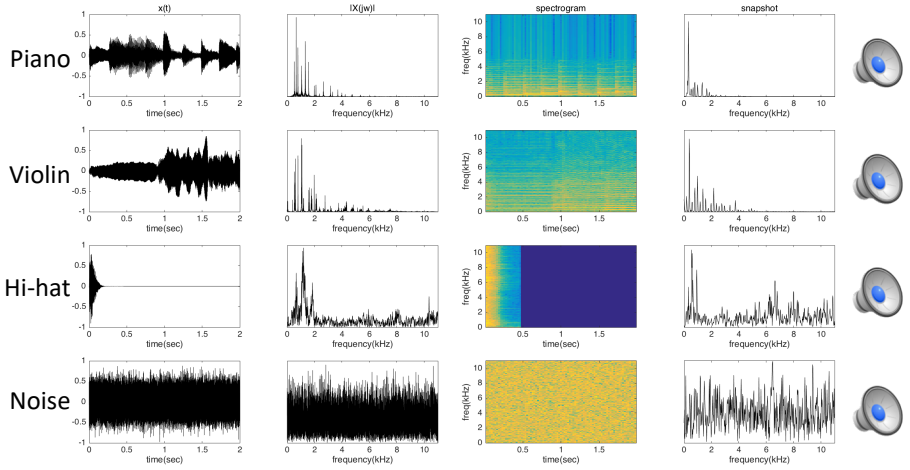
- Using the STFT, independent DFTs are calculated on windowed segments
- The segments usually overlap to compensate for the loss of temporal resolution
- Produces a 2-D *spectrogram*



Spectrogram examples



More examples



Questions?