CS4347/CS5647

Sound and Music Computing

L2a: Recap of DFT and Audio Representations

Wang Ye

www.comp.nus.edu.sg/~wangye

wangye@comp.nus.edu.sg

Topics to Cover (*selective approach*)

Part A: The Core

- Introduction
- Review of DFT, Audio Representation, and Machine Learning
- Music Representation, Analysis and Transcription
- Automatic Music Transcription (AMT)
- Automatic Speech Recognition (ASR)
- Generative Models for Text-to-Speech (TTS) & Singing Voice Synthesis (SVS)

Midterm break

Part B: The Breadth

- Singing voice processing
- Music production audio effects
- Automatic Music Generation
- Synthesis of sound & music – a DSP approach
- Project presentations/demo

Assessment (100% CA):

• Participation effort	**15%**
• Including lectures, tutorials, survey, *Canvas*, etc.	

• 3 Individual assignments	**40%**
• Week 2 (DFT)	10%
• Week 4 (AMT)	15%
• Week 6 (ASR)	15%

1st half of the semester

• 1 Group project	**45%**
• Week 10 (Mid-project assessment)	5%
• Week 13	
1) Presentation	10%
2) Code	10%
3) Final report	20%

2nd half of the semester

You can propose your own project and form your own project team.

The content of CS4347/CS5647 is carefully curated for the purpose of **education**

I never enjoyed the spoon-feeding approach as a student, and started to explore a better approach 10 years ago.

Students who firmly believe in the idea that the spoon-feeding approach is education, are advised not to take this course because my approach will be very different and you might not enjoy it! As NUS students, you have many choices to satisfy your educational needs.

CS4347/CS5647 is guided by my own educational model that is created to:

- ***Enable active and joyful learning***
- ***Cultivate growth mindset (as opposed to fixed mindset)***

Cognitive Neuroscience-informed Educational Model

(Operational)

F • T = P+

Multi-**I**ntelligences
Imagination
Initiative
Integrity

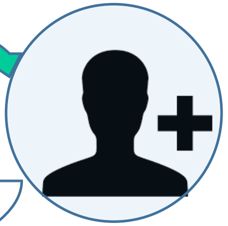
Excite
Energize
Engage
Enable
Collaboration
Connect the dots

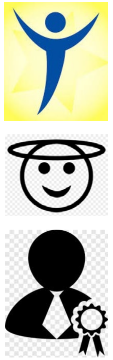
Person
*Project
*Paper
*Patent
*Product
*PhD

Transformation

Education is the kindling of a flame,
not the filling of a vessel (Socrates)

(philosophical)





Ability, curiosity & desire to learn

Inspire students to find their eigenfrequencies
Ignite a fire within!

Happy lifelong learner with a strong honor code!

Cognitive Neuroscience-informed Educational Model

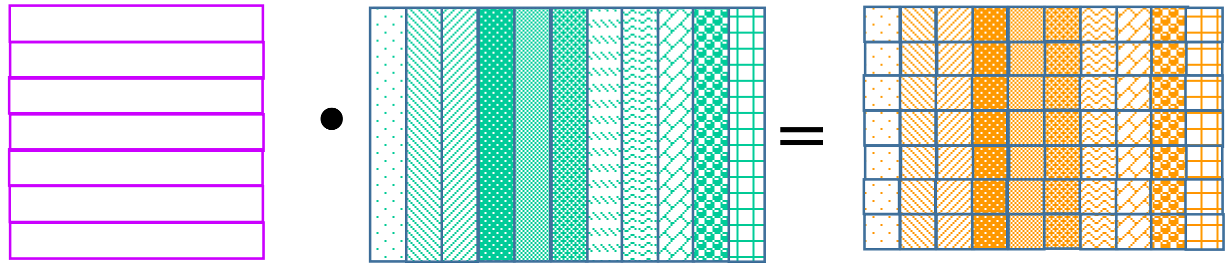
(Generalized for classroom teaching)

F • T = P+

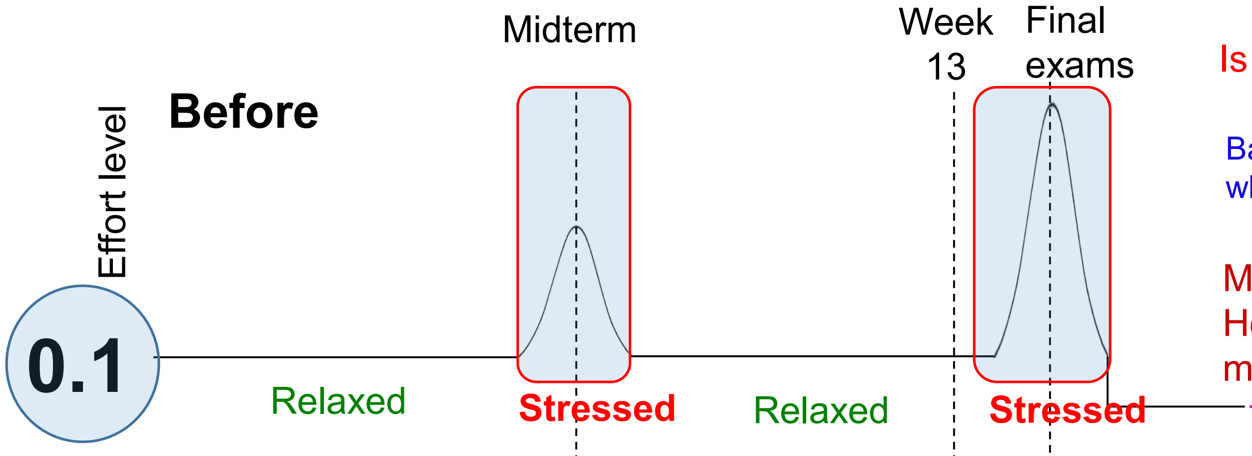
Multi-Intelligences
Imagination
Initiative
Integrity

Excite
Engage
Energize
Enable
Collaboration
Connect the dots

Person
*Project
*Paper
*Patent
*Product
*PhD



students   courses   resonance

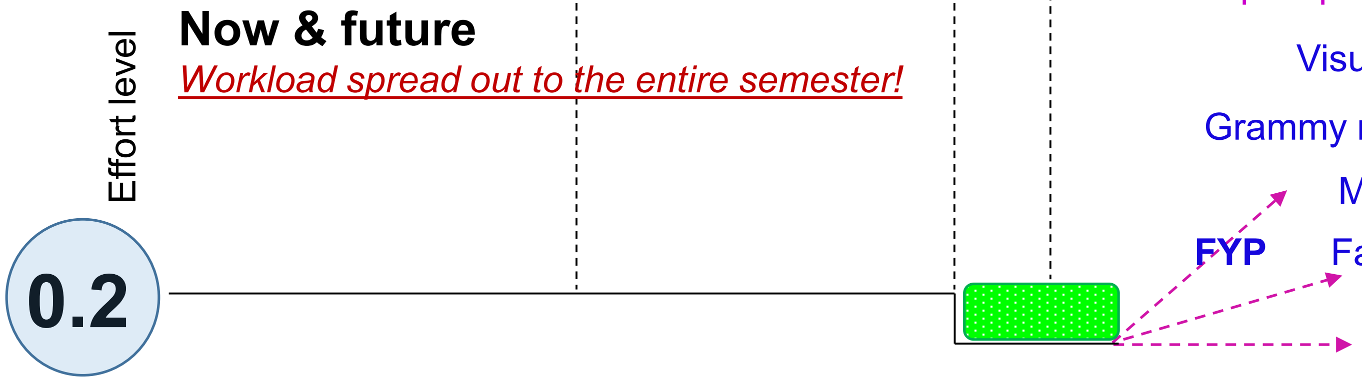


Is such an old educational approach effective?

Based on my own assessment, not really. That was why I decided to include DFT in the CS4347 syllabus.

My duty for student workload management
Helped me to find a significant problem with my initial assessment scheme.

My observation and reflection on our assessment schemes and student workload helped me to revamp the syllabus of the course completely, taking the knowledge from cognitive neuroscience into consideration.



The outcomes of the new educational approach are quite promising!

Visual artist Startups

Grammy nominated music producer

MIT, Stanford, CMU, …

Facebook, Huawei, Creative,…

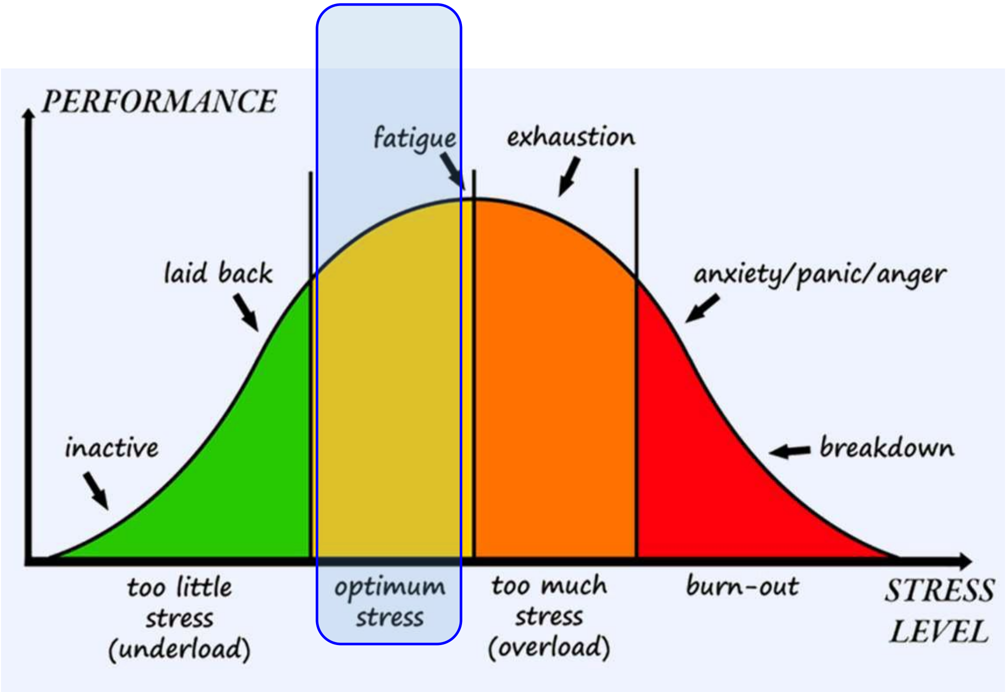
Lifelong learners!

Hope this course will offer you a different learning experience so that you could enjoy the learning journey and would continue exploring *after* graduation.

Types of Stress

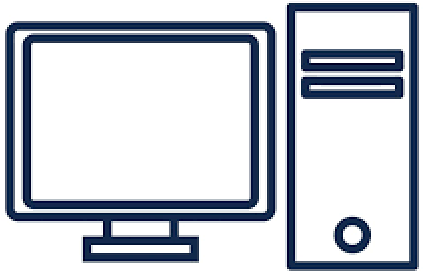
The rationale for the revamp is to keep our students in the optimum stress level!

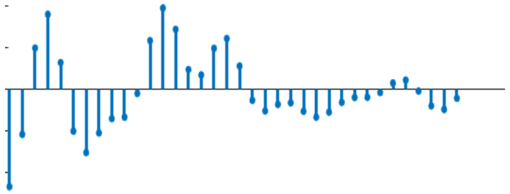
- Good stress
  - Within our control
  - Generates energy, drive & excitement
  - Meaningful stress
- Bad stress
  - Chronic Stress
  - Damages physical, mental & emotional health



This slide is borrowed from Dr Andrew Epaphroditus Tay & Jeanie Chu
Health & Wellbeing
Office of the President, NUS

CS4347/CS5647 is about Building Bridges





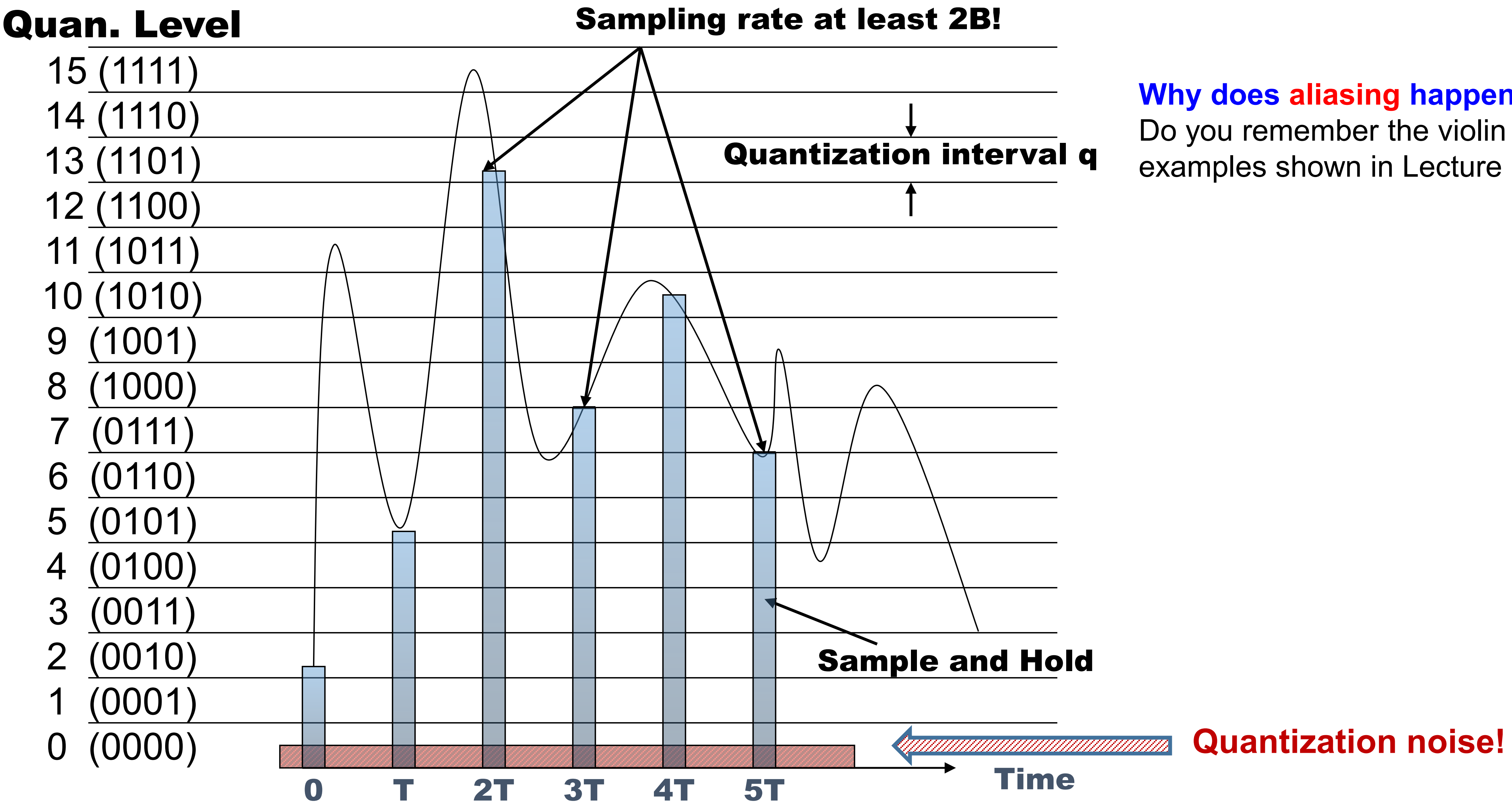
Nyquist Sampling theorem

Analog ⟷ Digital

Discrete Fourier Transform (DFT)

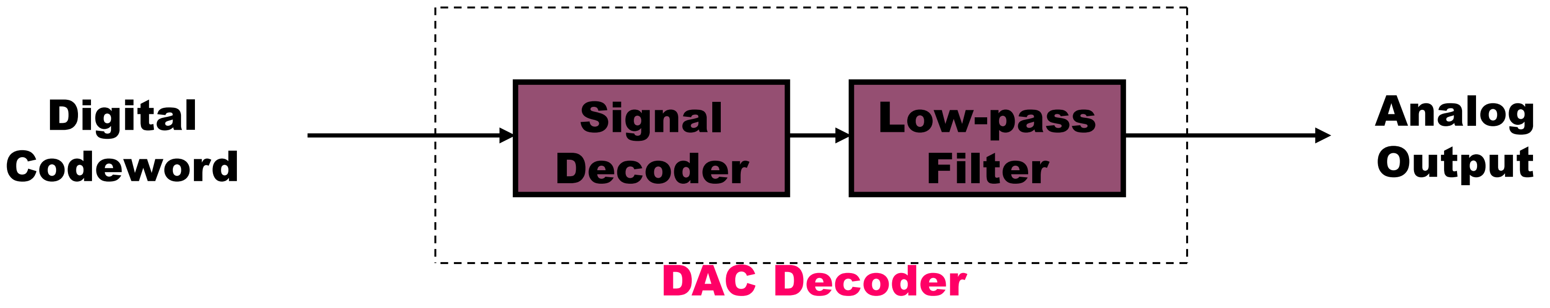
Time ⟷ Frequency

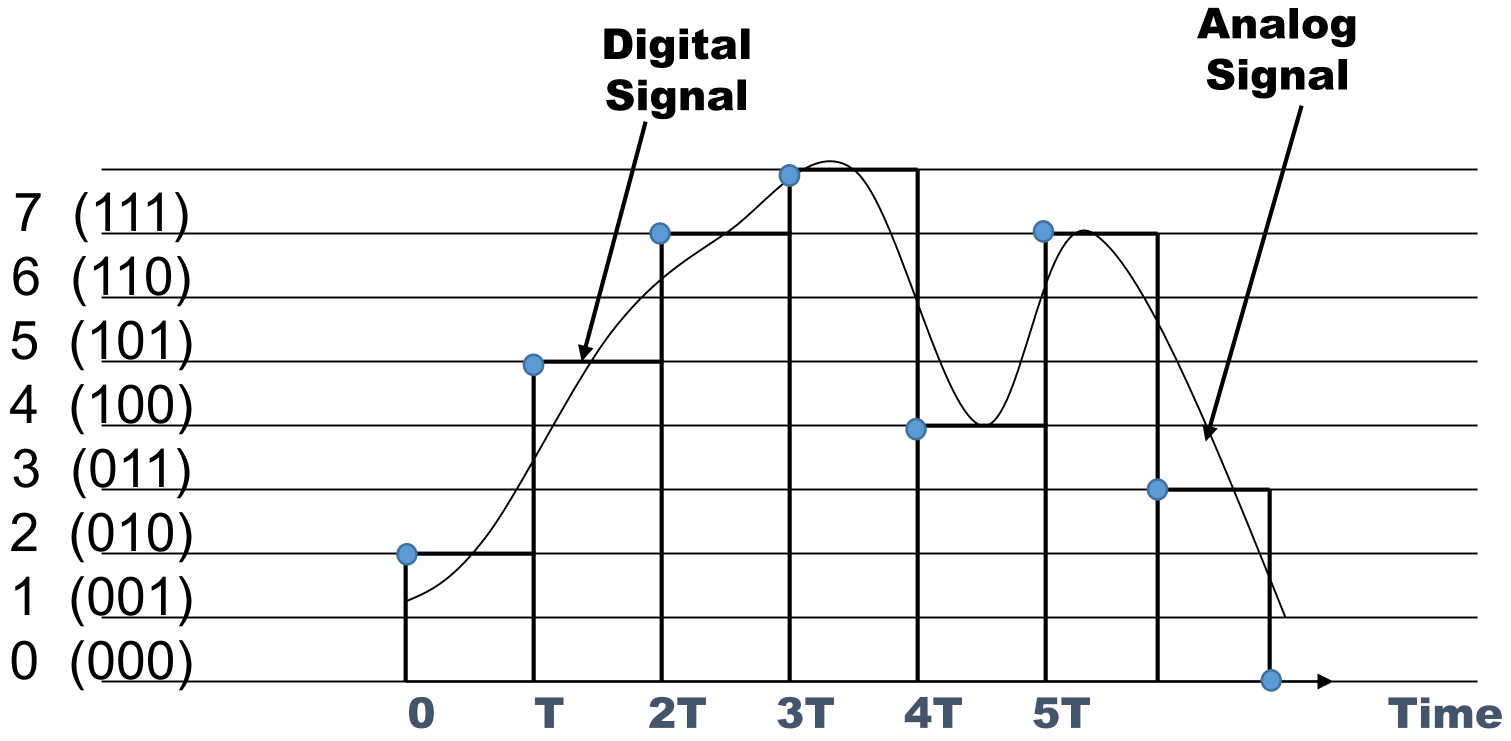
Sampler, Quantizer and Encoder
(ADC: Analog-to-Digital Converter)



Why does aliasing happen?
Do you remember the violin examples shown in Lecture 1?

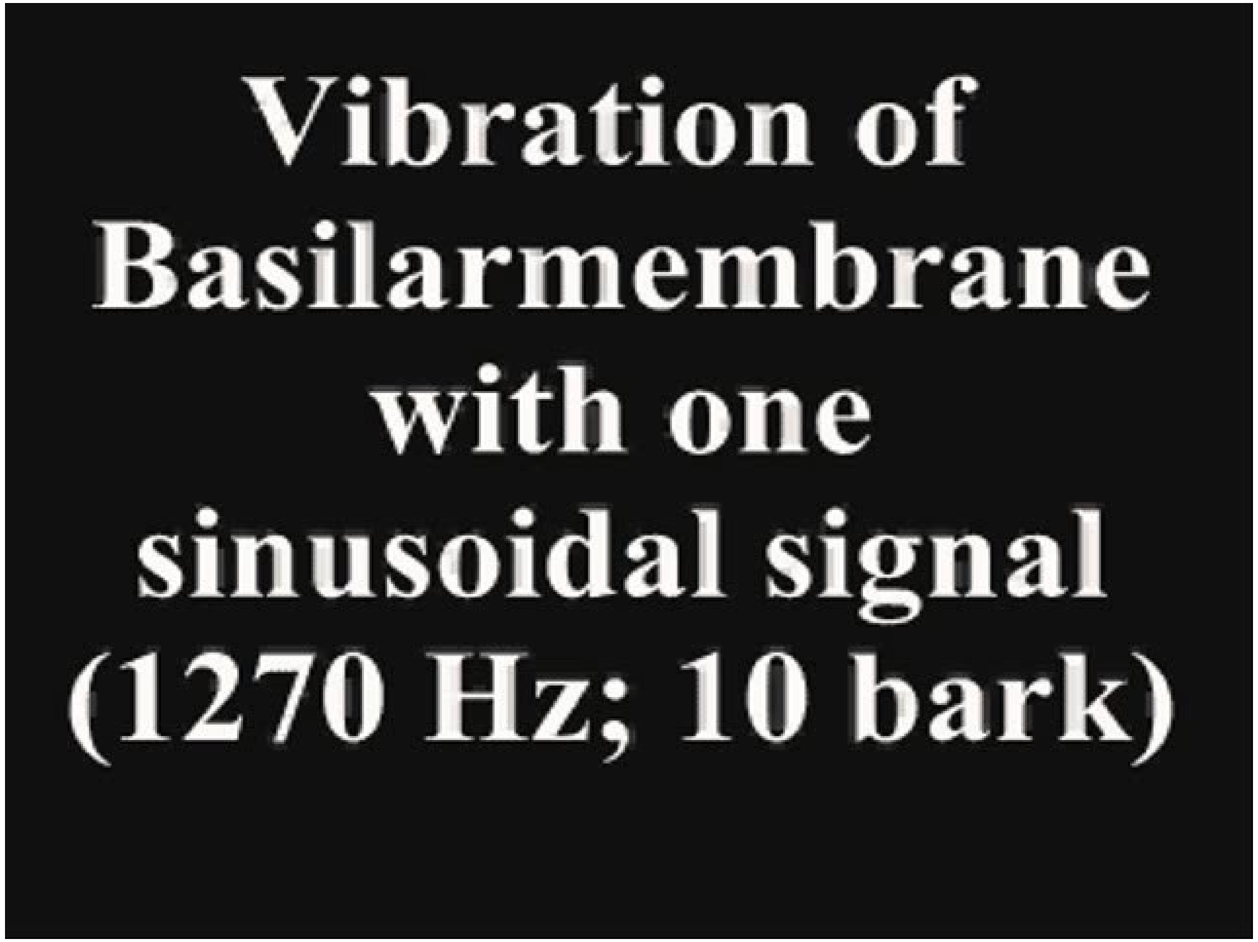
DAC: Digital-to-Analog Converter (**interpolation**)





Audio/Video Coder-Decoder often referred to as Audio-Video CODEC

Fourier Transform in Our Ear!



It is a physiological justification for the short-time Fourier transform (STFT)!

The Discrete Fourier Transform (DFT) is one of the most important and powerful tools in Digital Signal Processing (DSP).

For X[k], what are we computing here?
Or what is the physical interpretation of X[k]?

DFT Formula:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}$$

DFT Basis functions

Euler's formula:

$$e^{j\theta} = \cos(\theta) + j\sin(\theta)$$

Let's Demystify DFT together!

DFT Basis functions in the complex plane

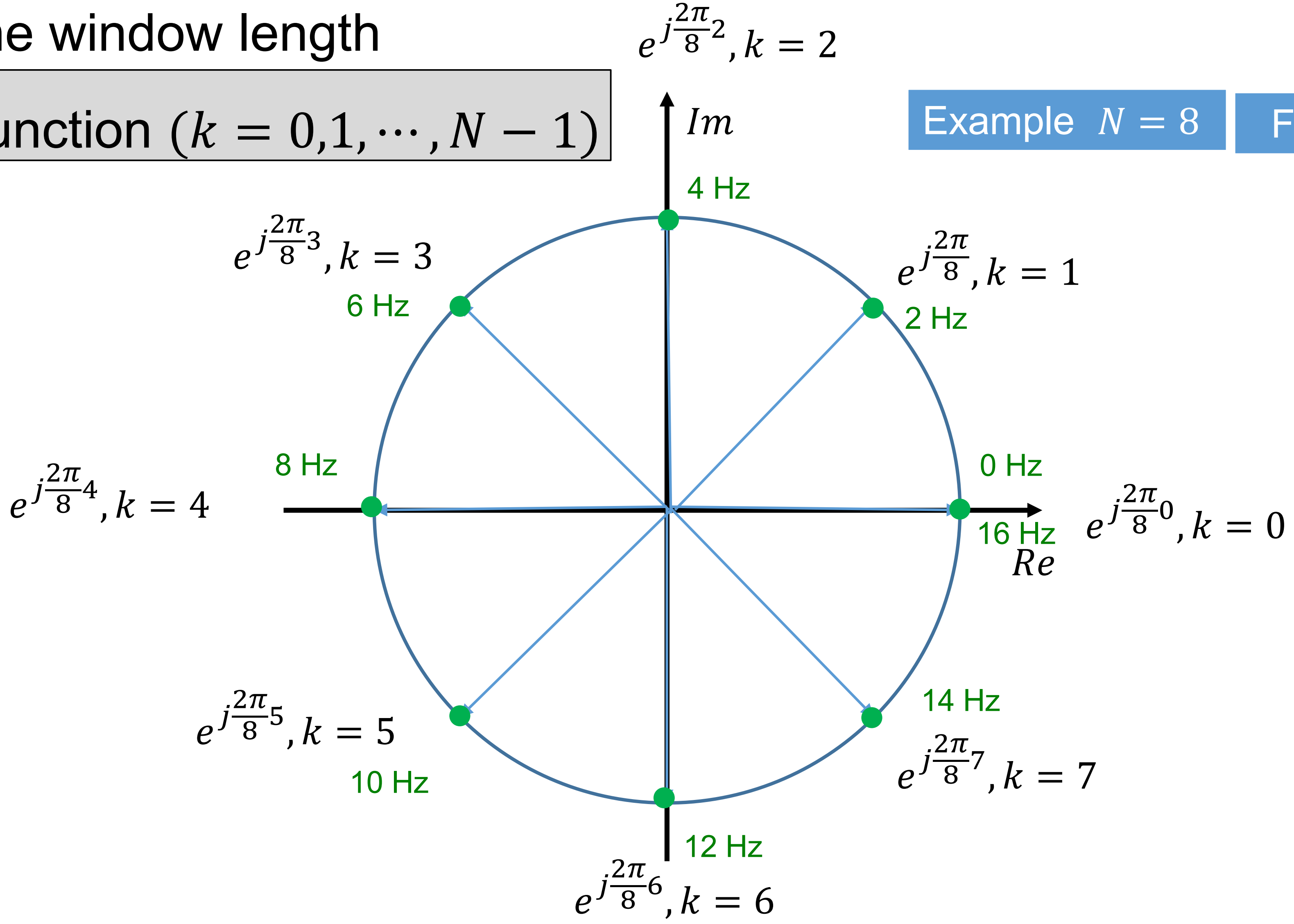
$e^{j\theta} = \cos(\theta) + j\sin(\theta)$

Assume N is the window length

$e^{\frac{j2\pi}{N}k}$ as basis function $(k = 0,1,\cdots,N-1)$

Example $N = 8$

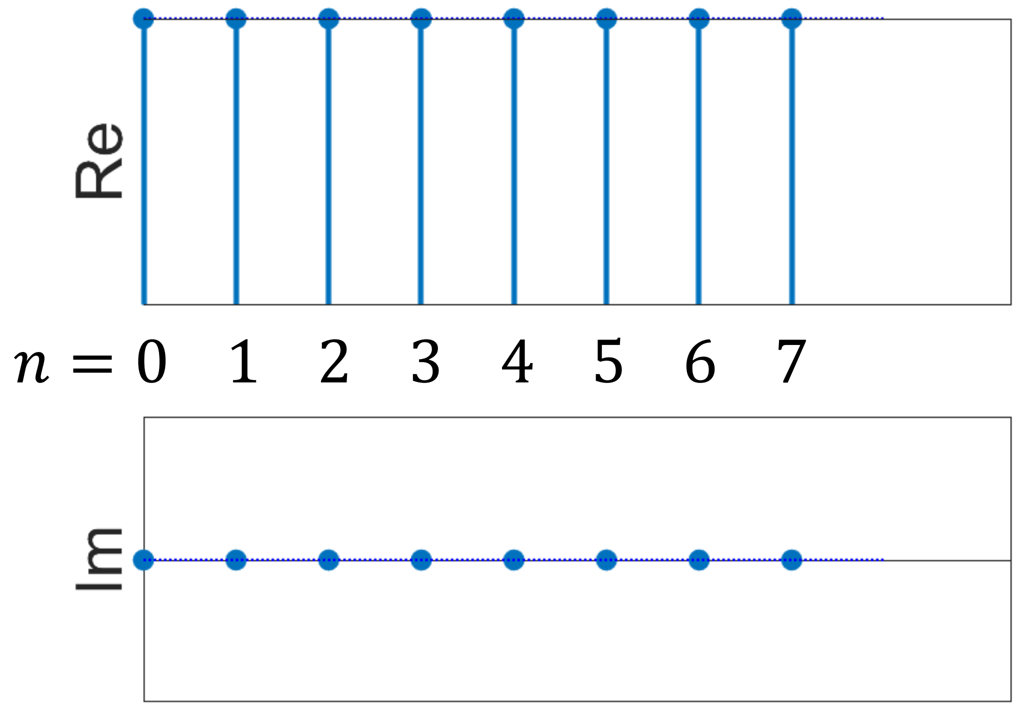
Fs = $16Hz$



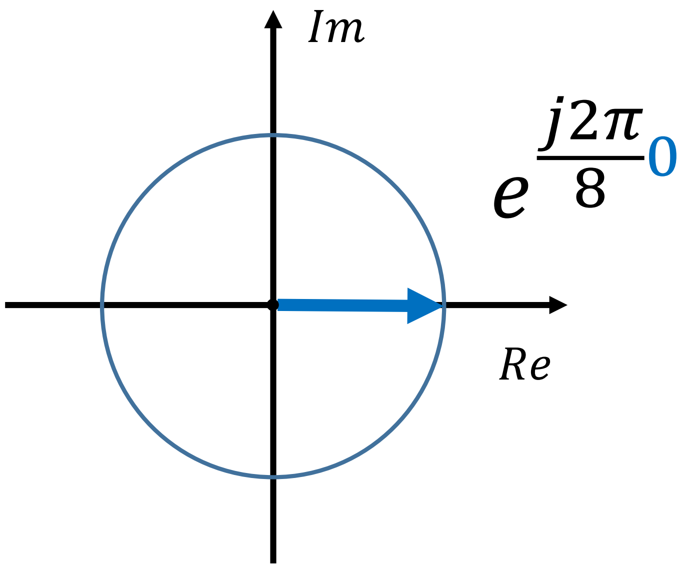
DFT Basis function

Example $N = 8$ Fs $= 16Hz$

$$e^{\frac{j2\pi}{8}0n}, n = 0,1,\cdots,7$$



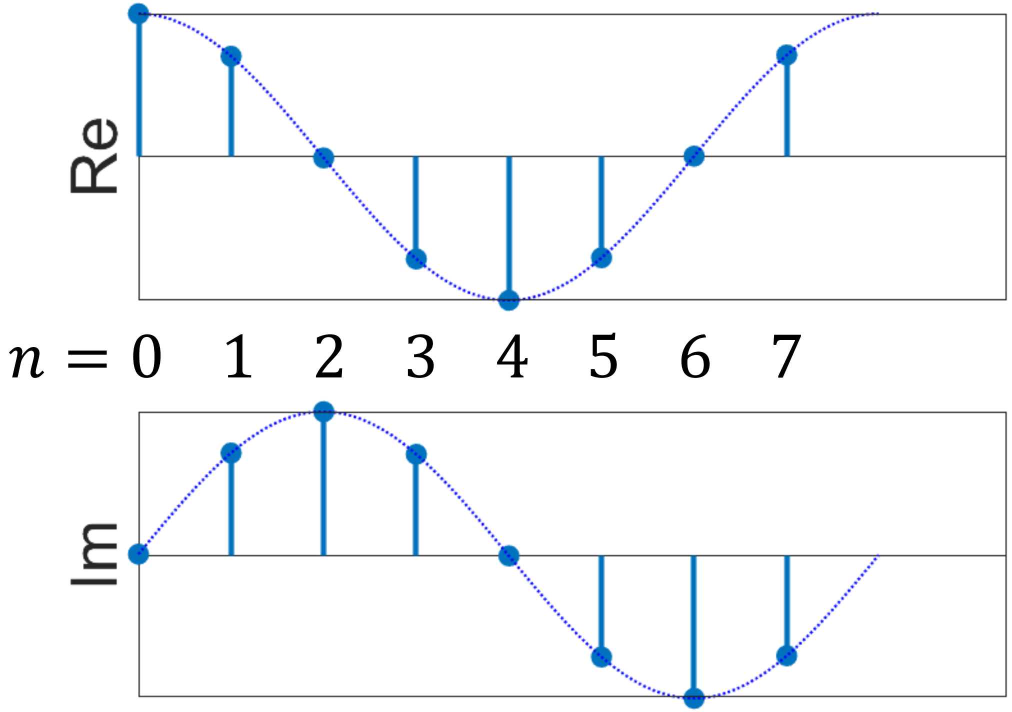
$$e^{\frac{j2\pi}{8}k}, k = 0$$



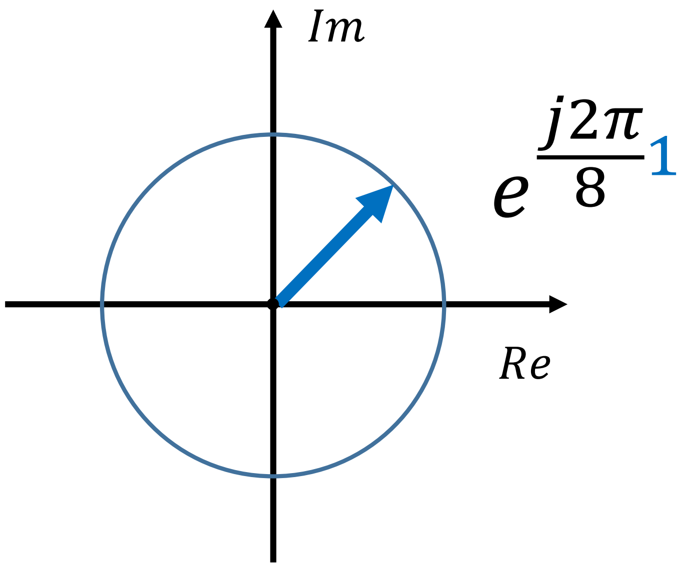
DFT Basis function

Example $N = 8$ | $\text{Fs} = 16Hz$

$$e^{\frac{j2\pi}{8}1n}, n = 0,1,\cdots,7$$



$$e^{\frac{j2\pi}{8}k}, k = 1$$

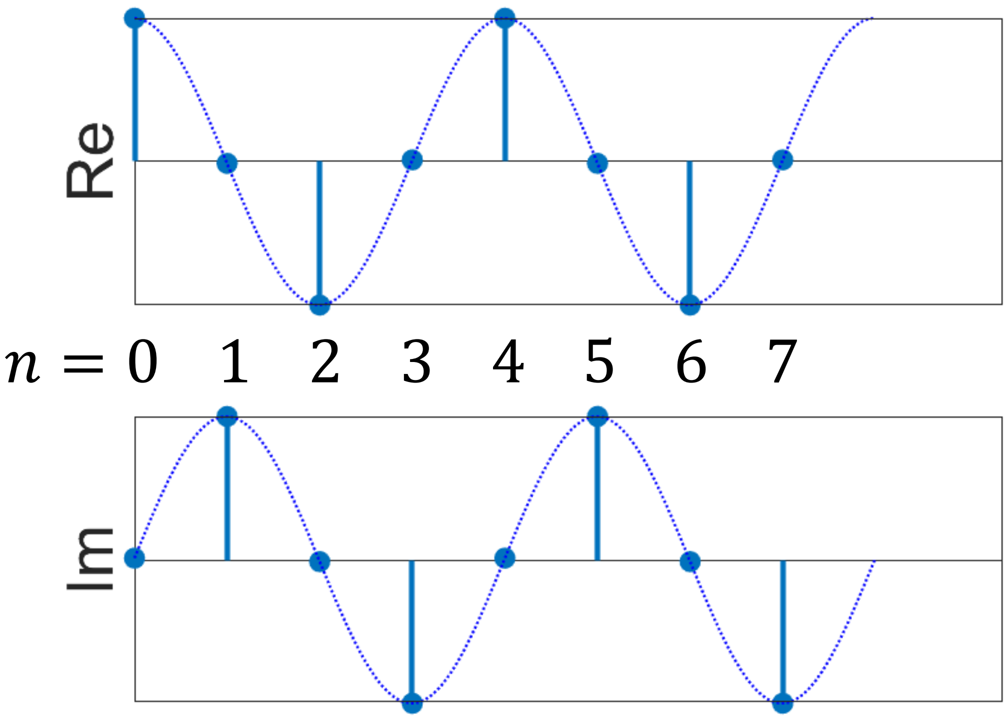


DFT Basis function

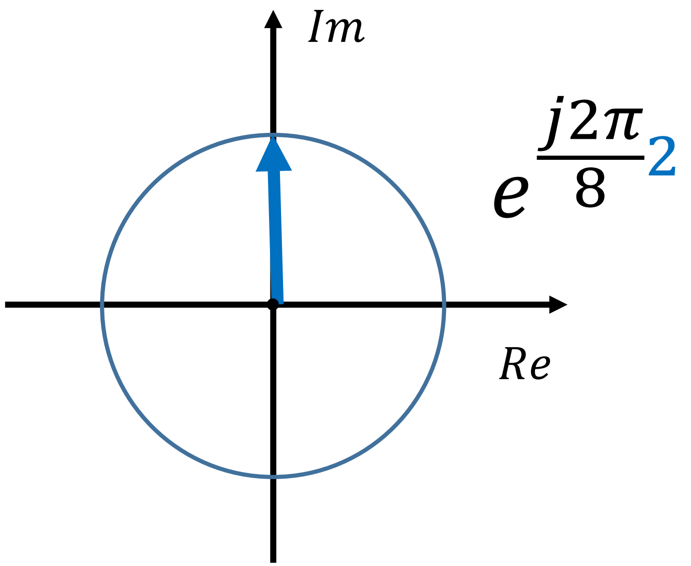
Example $N = 8$

Fs $= 16Hz$

$$e^{\frac{j2\pi}{8}2n}, n = 0,1,\cdots,7$$



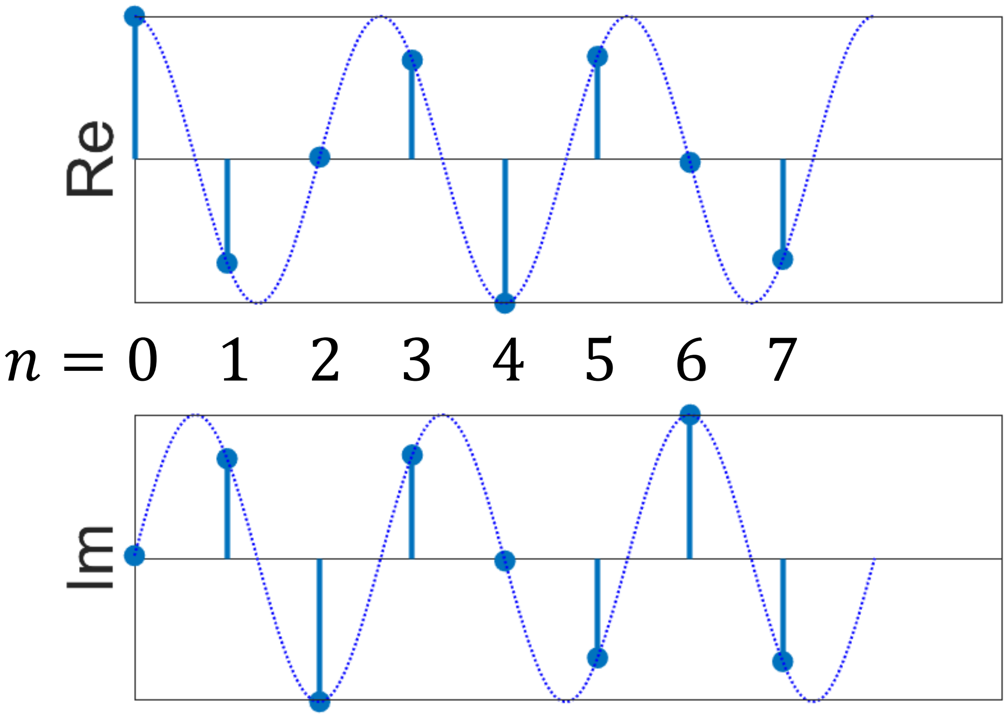
$$e^{\frac{j2\pi}{8}k}, k = 2$$



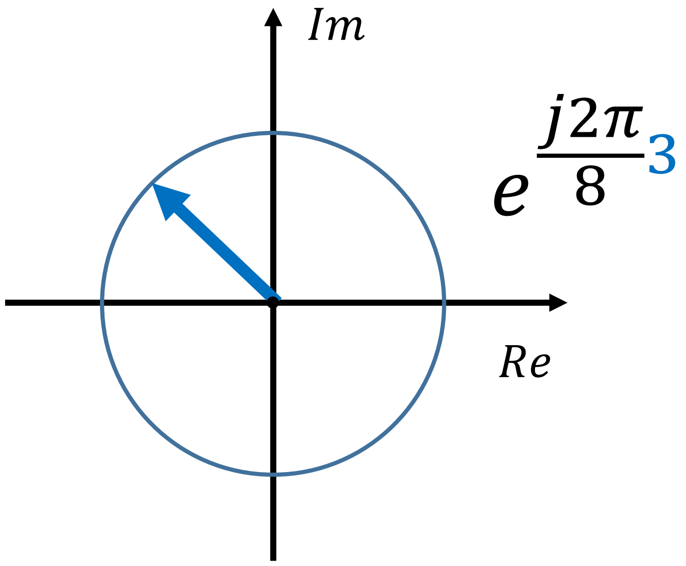
DFT Basis function

Example $N = 8$ | Fs $= 16Hz$

$$e^{\frac{j2\pi}{8}3n}, n = 0,1,\cdots,7$$



$$e^{\frac{j2\pi}{8}k}, k = 3$$

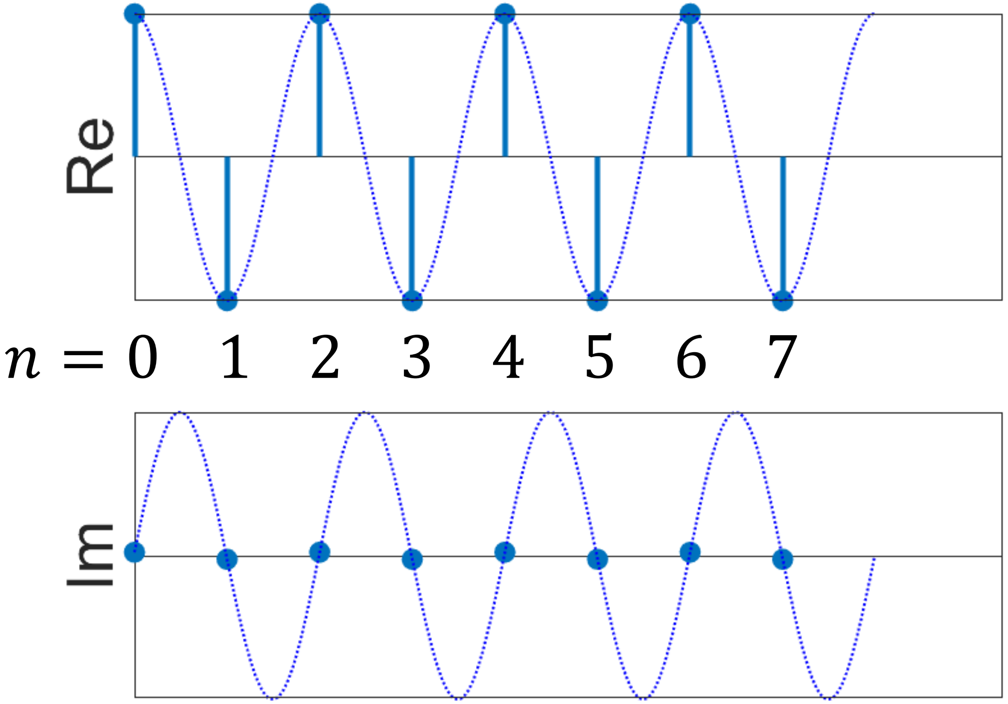


DFT Basis function

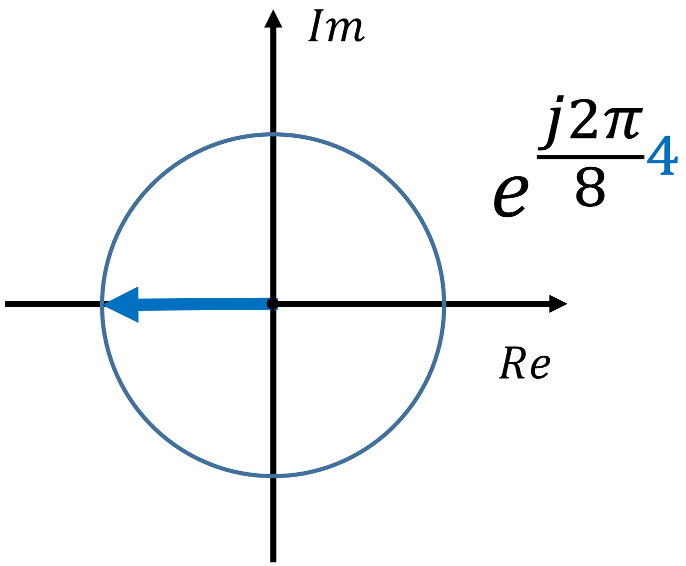
Example $N = 8$

Fs $= 16Hz$

$$e^{\frac{j2\pi}{8}4n}, n = 0,1,\cdots,7$$



$$e^{\frac{j2\pi}{8}k}, k = 4$$

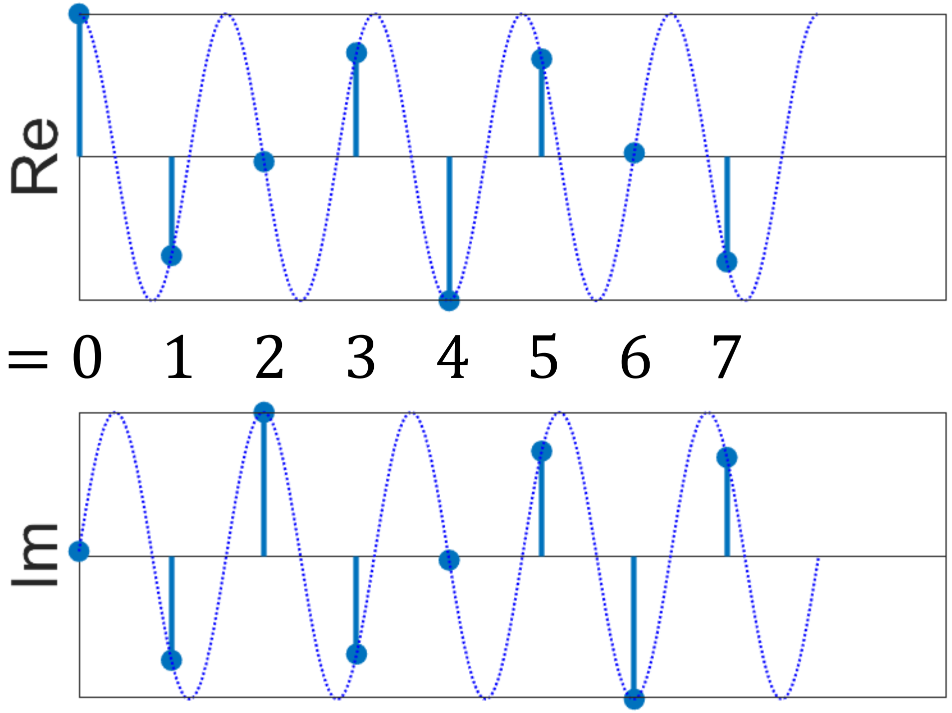


DFT Basis function

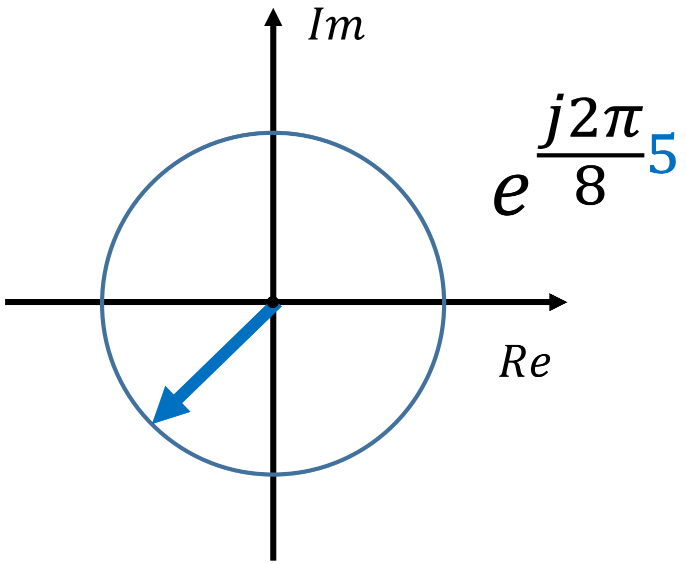
Example $N = 8$

$Fs = 16Hz$

$$e^{\frac{j2\pi}{8}5n}, n = 0,1,\cdots,7$$



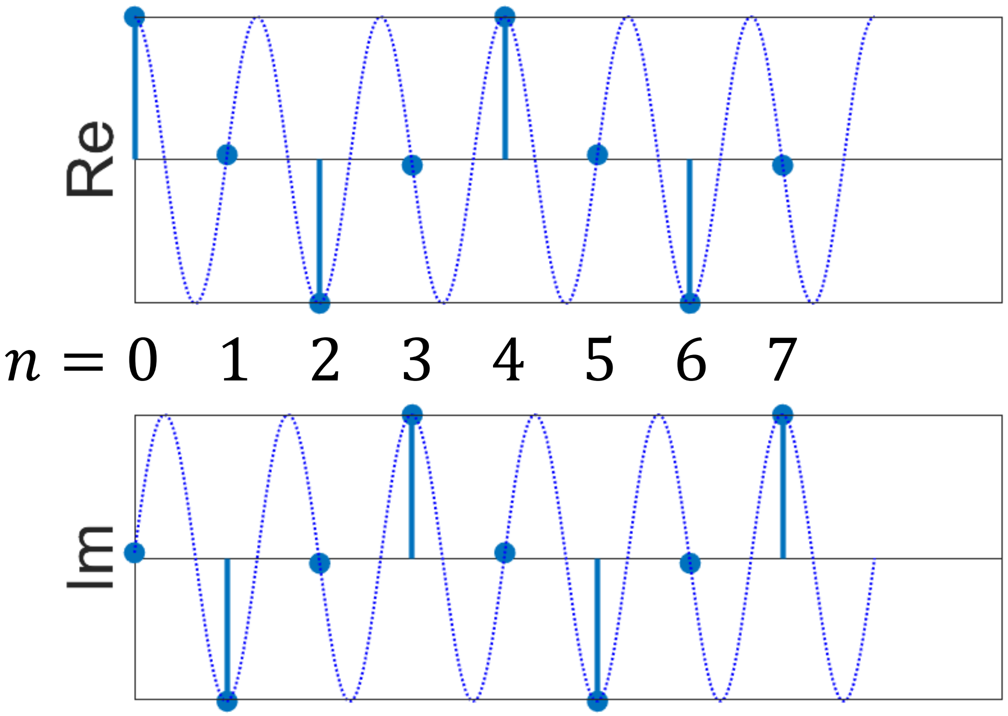
$$e^{\frac{j2\pi}{8}k}, k = 5$$



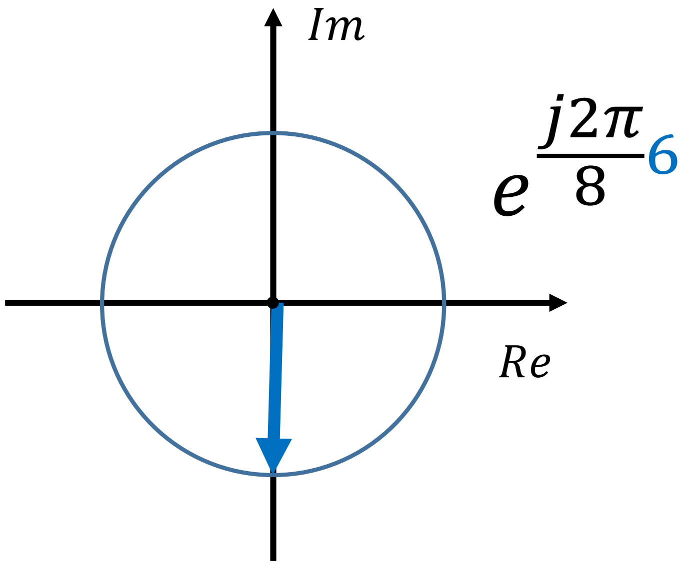
DFT Basis function

Example $N = 8$ | Fs $= 16Hz$

$$e^{\frac{j2\pi}{8}6n}, n = 0,1,\cdots,7$$



$$e^{\frac{j2\pi}{8}k}, k = 6$$

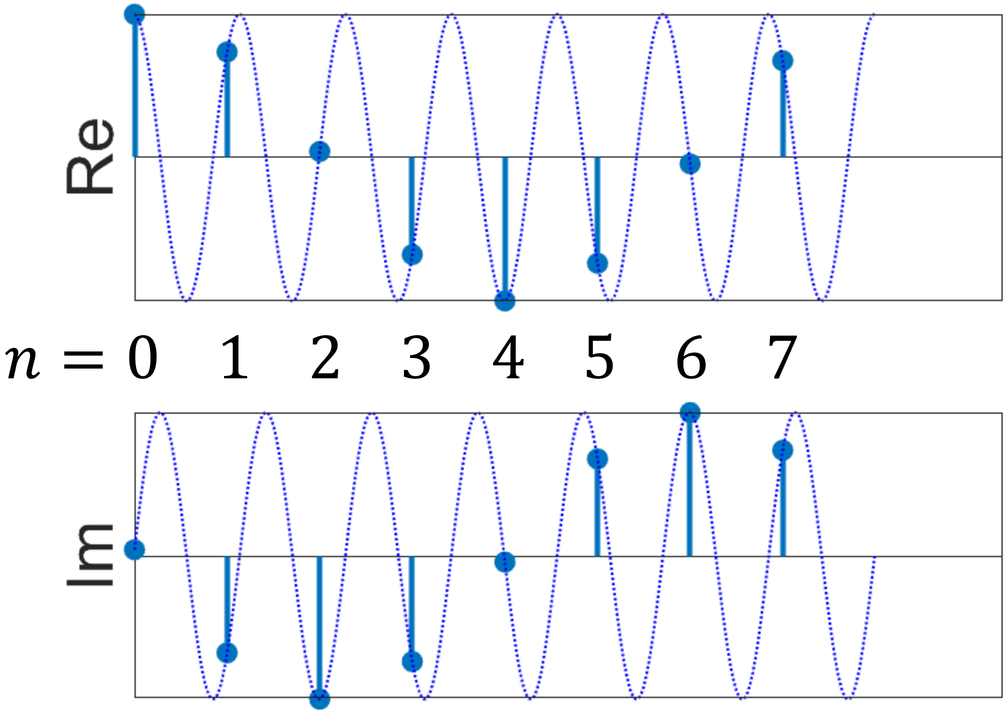


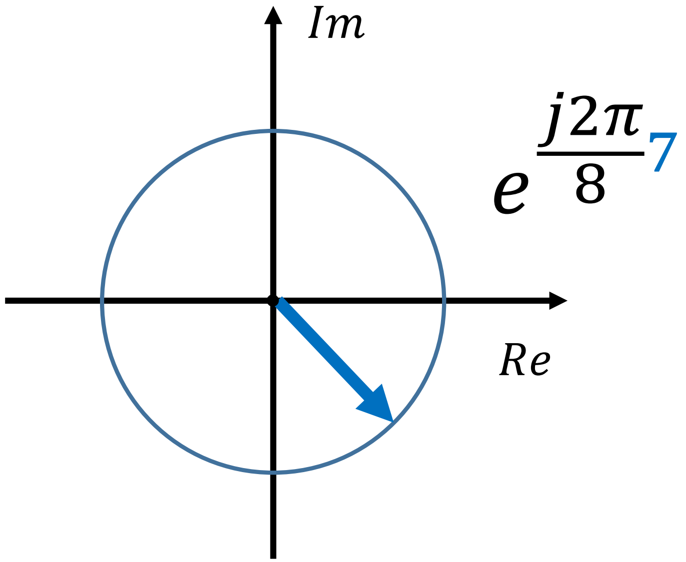
DFT Basis function

Example $N = 8$ Fs $= 16Hz$

$$e^{\frac{j2\pi}{8}7n}, n = 0,1,\cdots,7$$

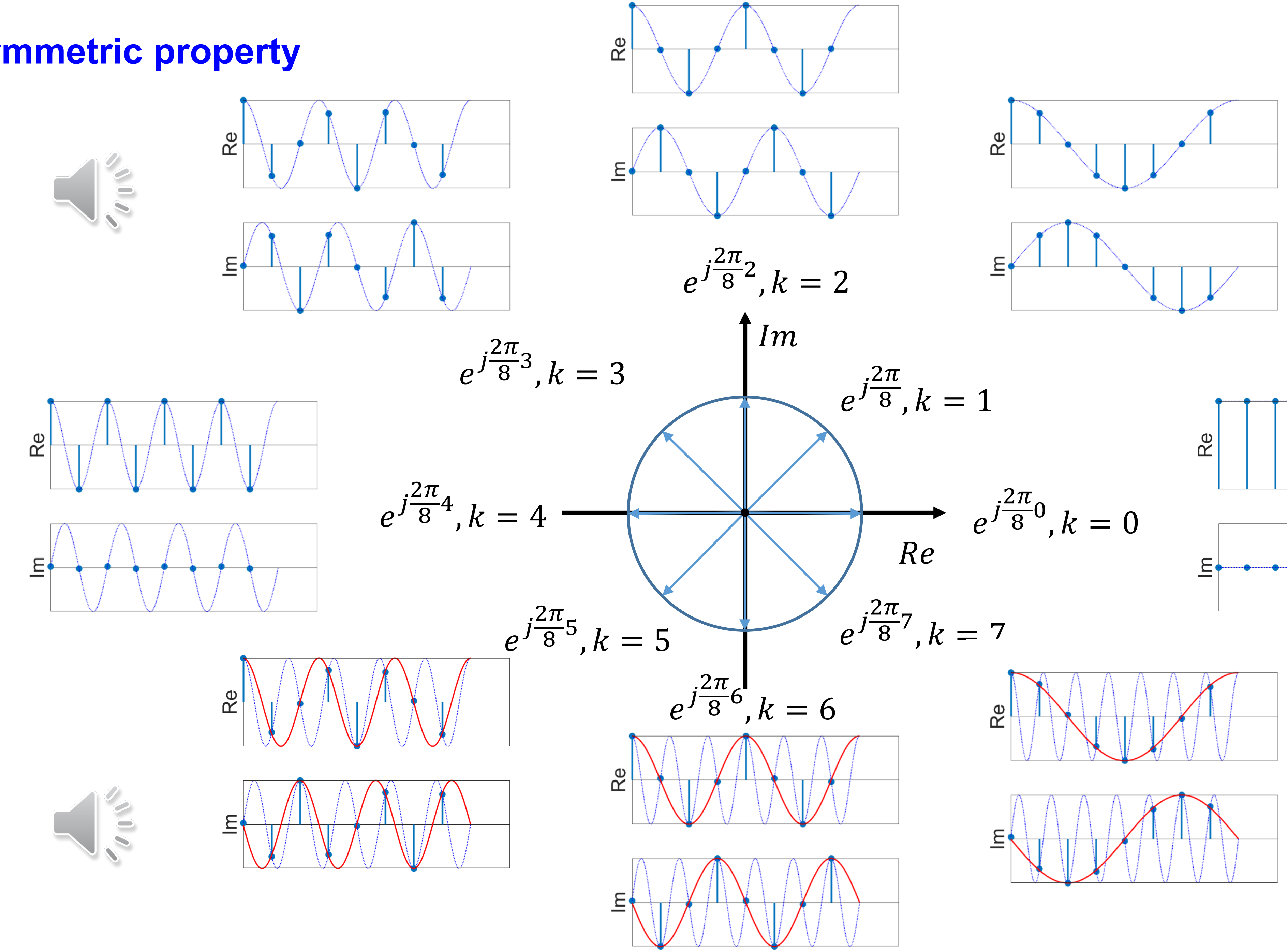
$$e^{\frac{j2\pi}{8}k}, k = 7$$

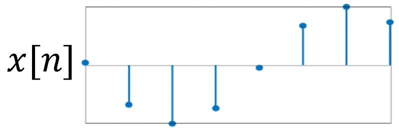




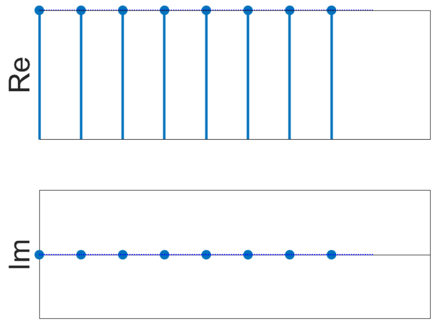
Symmetric property

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}$$





$e^{j\frac{2\pi}{8}0}, k = 0$



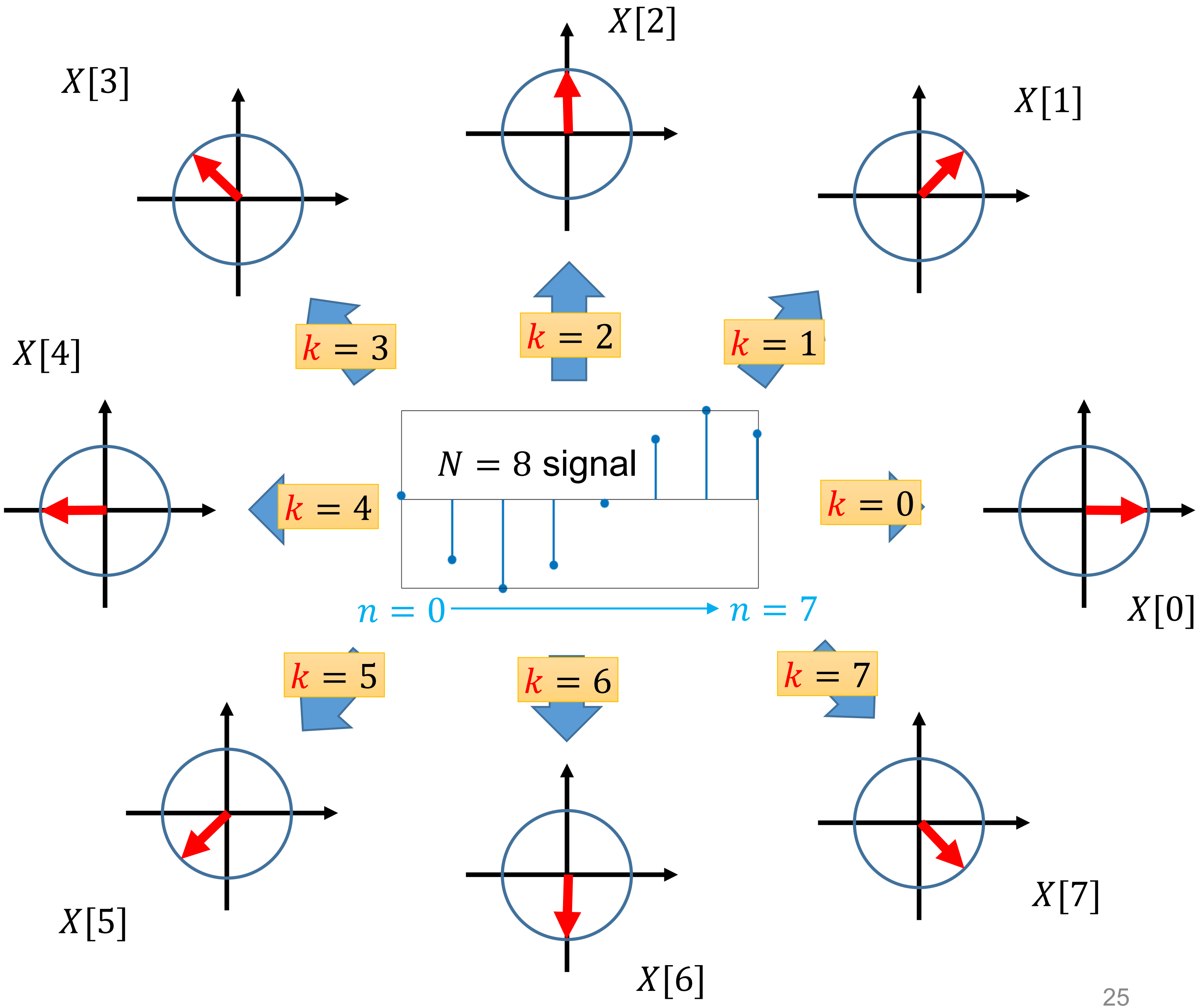
This interpolation Explains why aliasing happens!

The key insight of DFT is a form of pattern matching between $x[n]$ and each basis function!

DFT Formula:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}$$

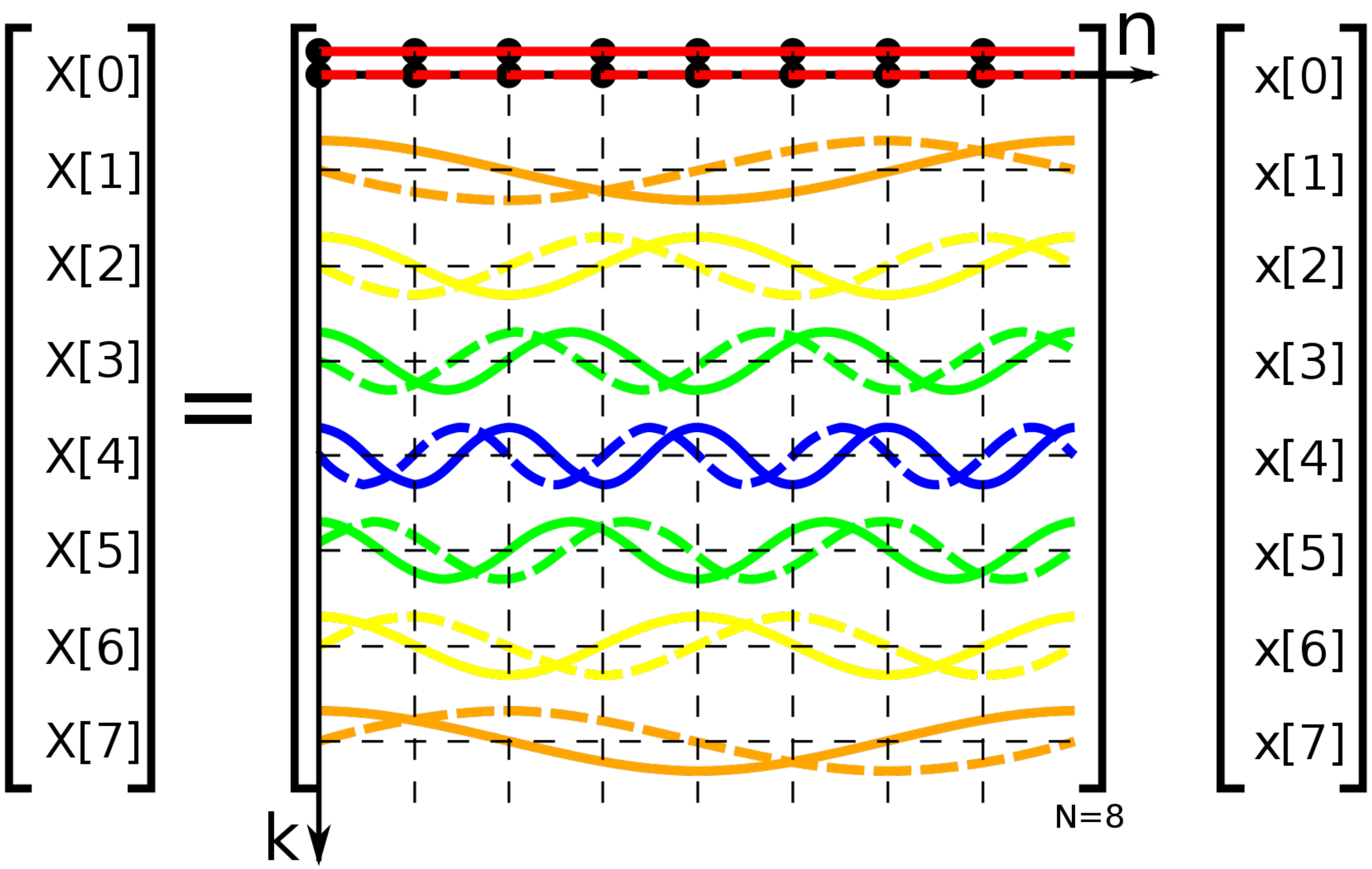
Compute the ***similarity*** between $x[n]$ and each basis function



The purpose of DFT is to determine the frequency components of an unknown signal $x[n]$!

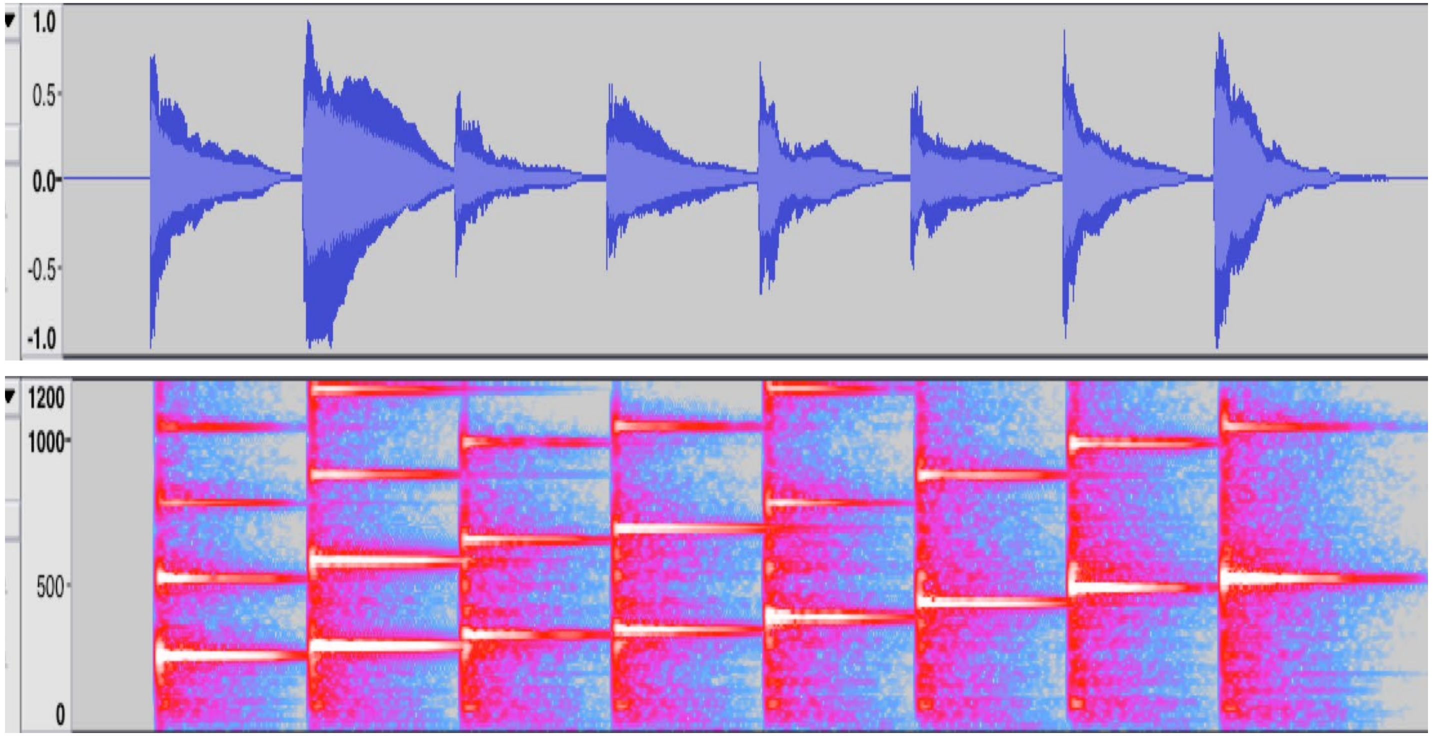
DFT Formula: $X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn}$

A Matrix representation



Audio representations

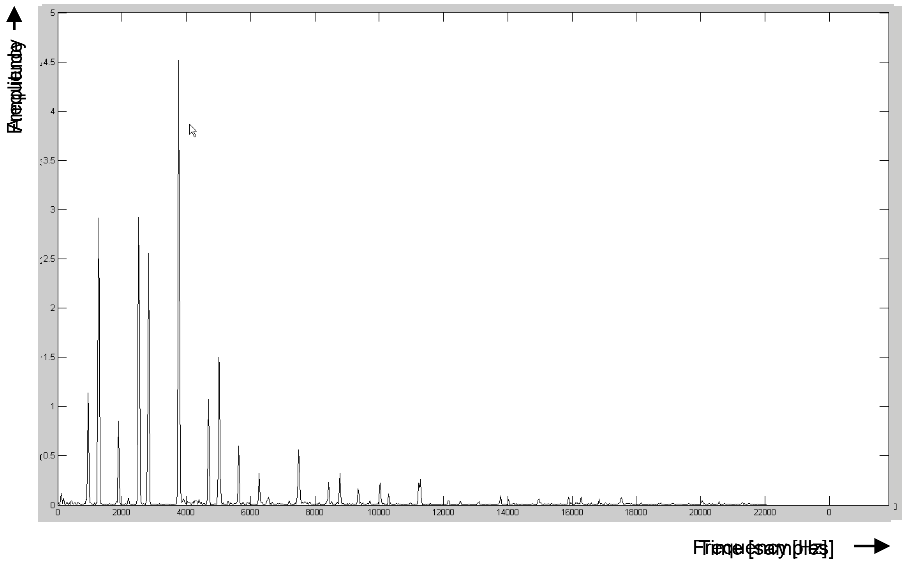
Audio representation in time and frequency domain



Waveform

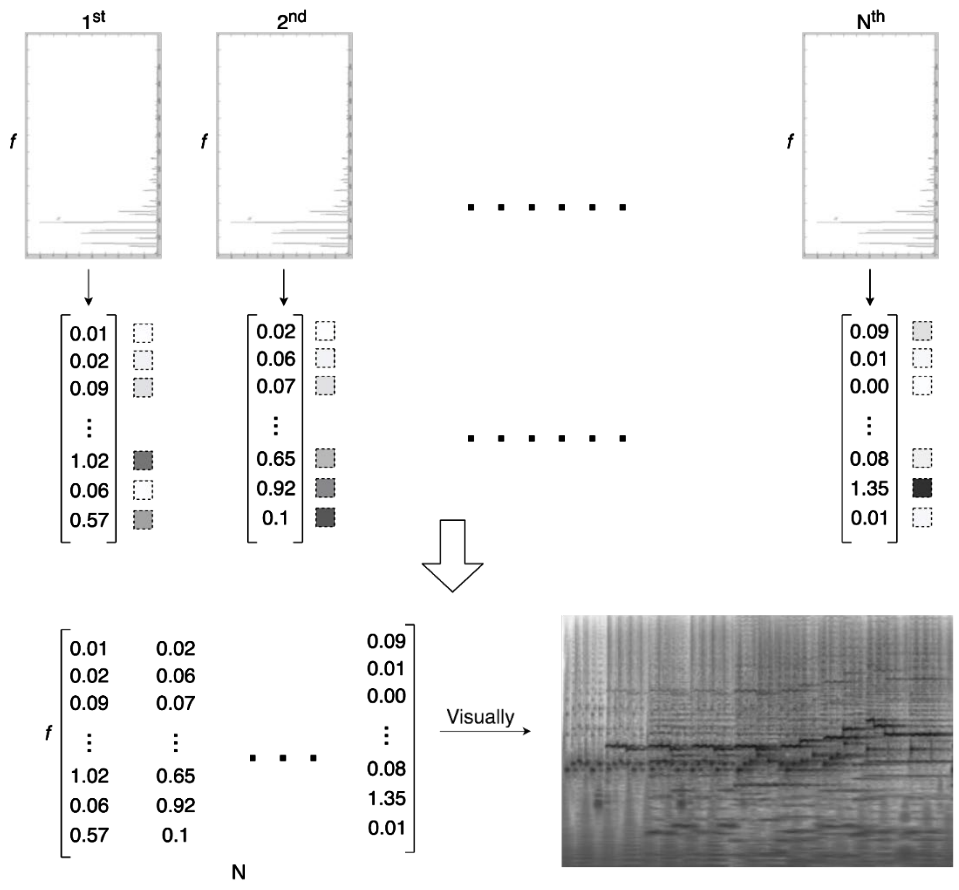
Spectrogram

How is a spectrogram constructed?



DFT

How is a spectrogram constructed?



DFT with audio

Recall that the DFT gives us a complex number for each frequency bin $X[k]$

For audio, we usually care about the log-magnitude spectrum:

$$|X[k]| = \sqrt{Re(X[k])^2 + Im(X[k])^2}$$

$$|X[k]|_{dB} = 20\log_{10}(|X[k]| + \epsilon)$$

Given

- *X[k]*: complex numbers corresponding to frequency bin k
- *Re*: real part
- *Im*: imaginary part
- *ε*: very small number; prevents log(0)

Phase φ[k] are much less commonly-used

$$\varphi[k] = -\tan^{-1}\left(\frac{Im(X[k])}{Re(X[k])}\right)$$

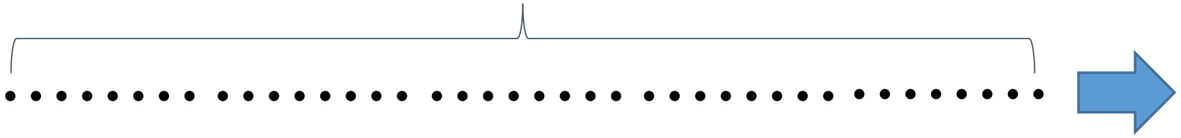
Trade-off of temporal & frequency resolution

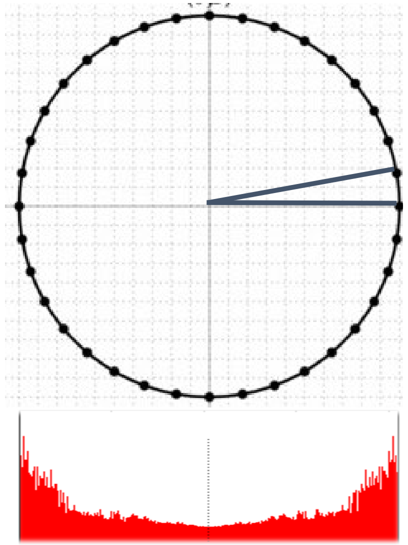
DFT gives us information about frequency bins.

- N = 2048, 44100 Hz signal → each bin is 44100/2048 = 21 Hz wide; each buffer is 46.4 ms long

The larger the window size (temporal or time resolution), the smaller the width of frequency bins (frequency resolution)

$$X[k] = \sum_{n=0}^{N-1} x[n]\ e^{-j2\pi k\frac{n}{N}}$$





Mapping a time-domain input array to a complex frequency-domain representation

Audio analysis with DFT

Timbre: Time-Frequency domain representations

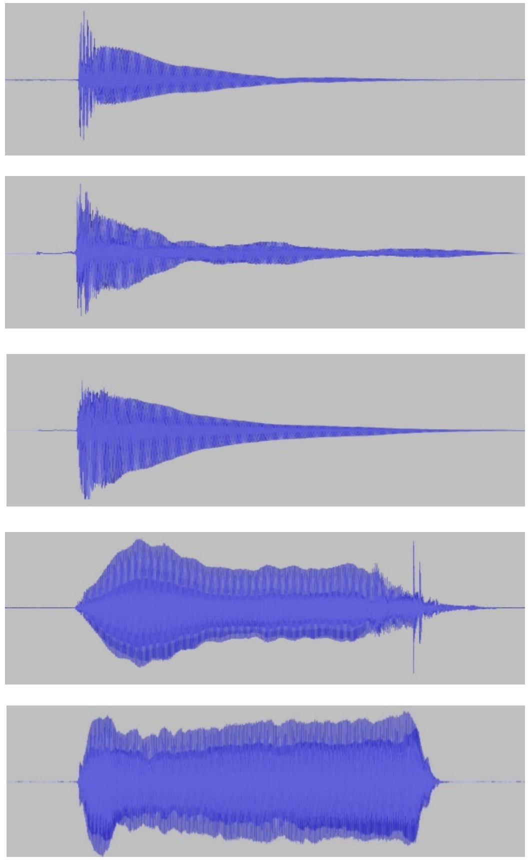
Classical Guitar

Acoustic Guitar

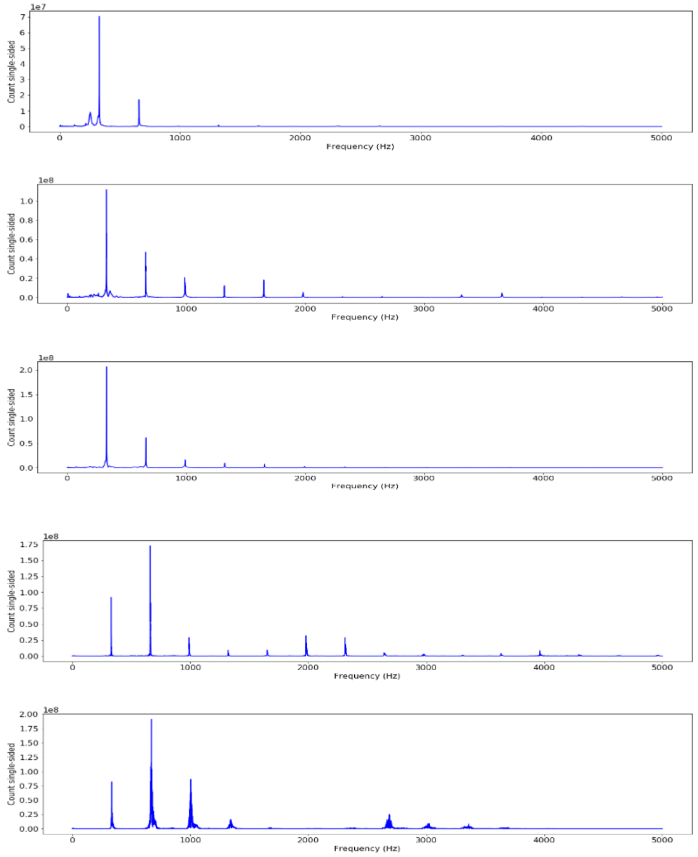
Piano

Violin

Human Voice



Raw Waveforms (time domain)



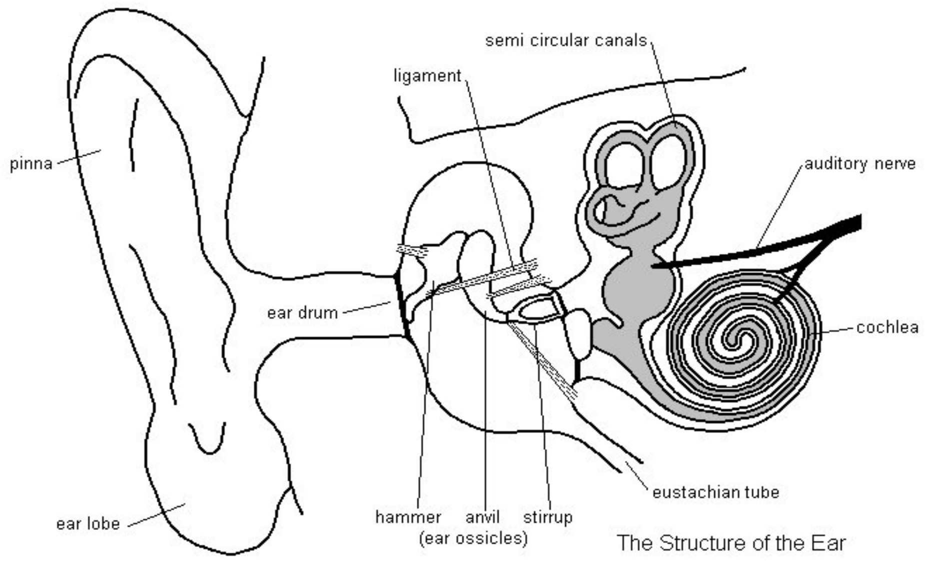
Spectrum (frequency domain)

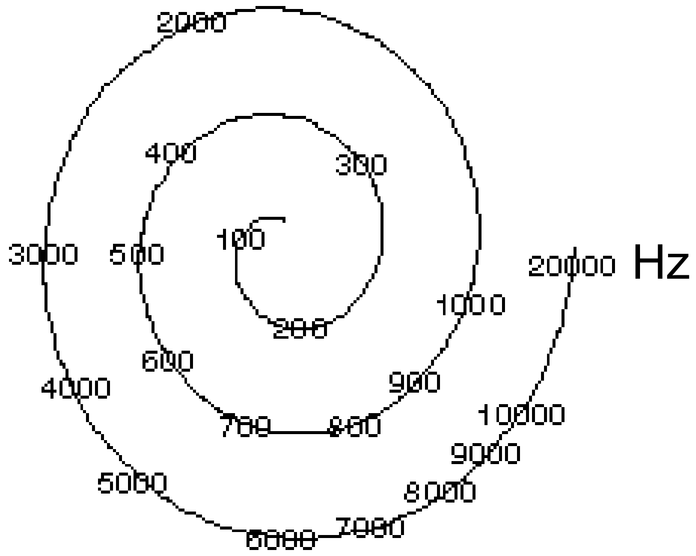
Time-Frequency Representations

Some cool Audio Analysis websites:

1. https://academo.org/demos/spectrum-analyzer/
2. https://musiclab.chromeexperiments.com/Spectrogram/
3. https://p5js.org/examples/sound-frequency-spectrum.html

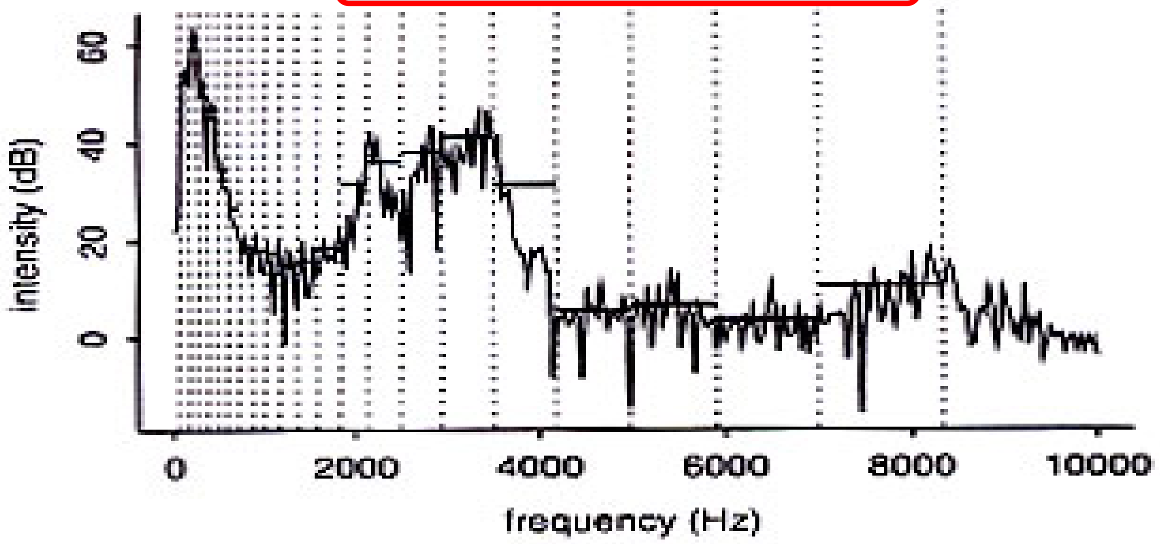
Fourier Transform in Our Ear – a First-order Approximation





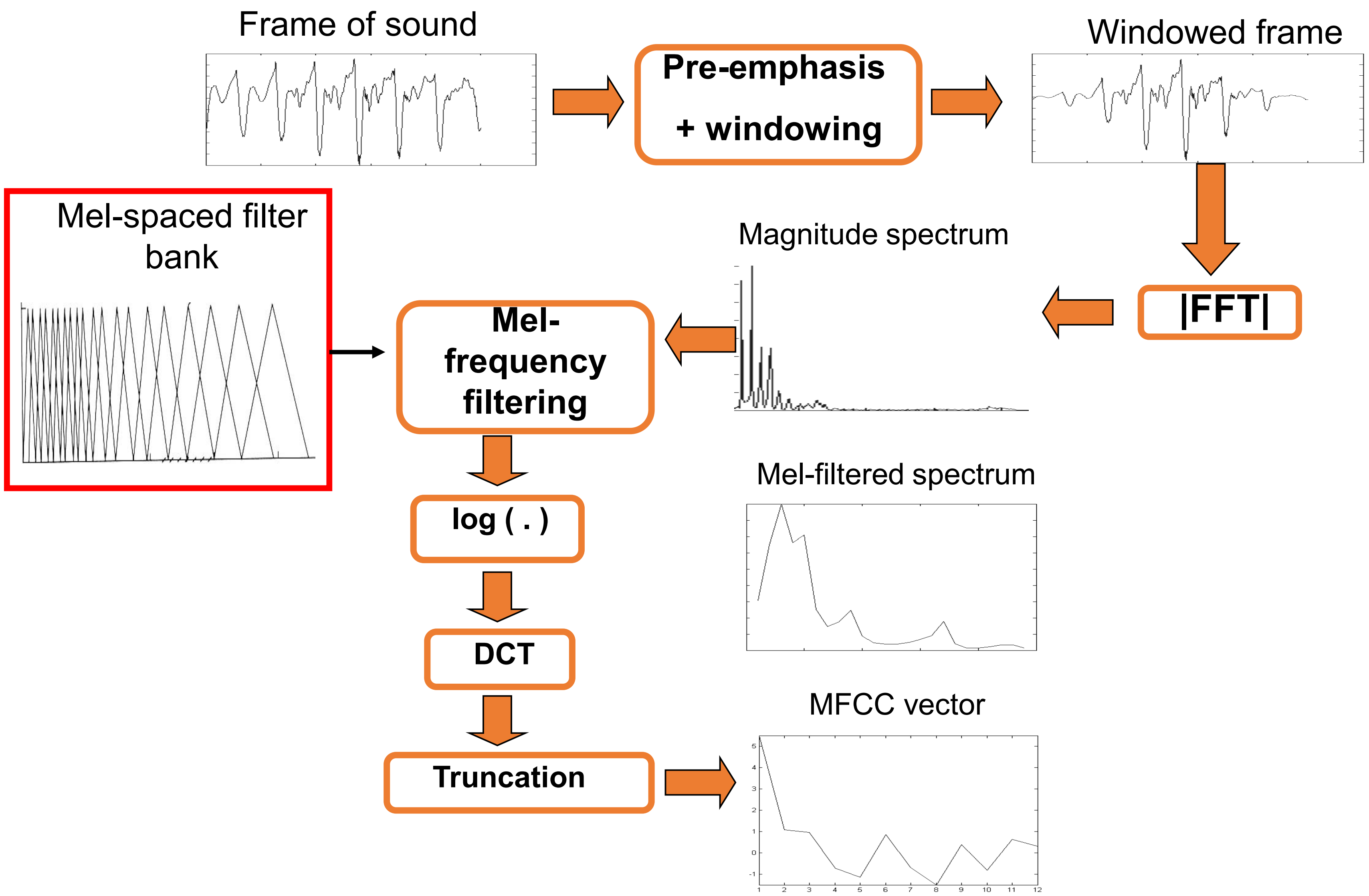
Critical band filtering

DFT: linear transform

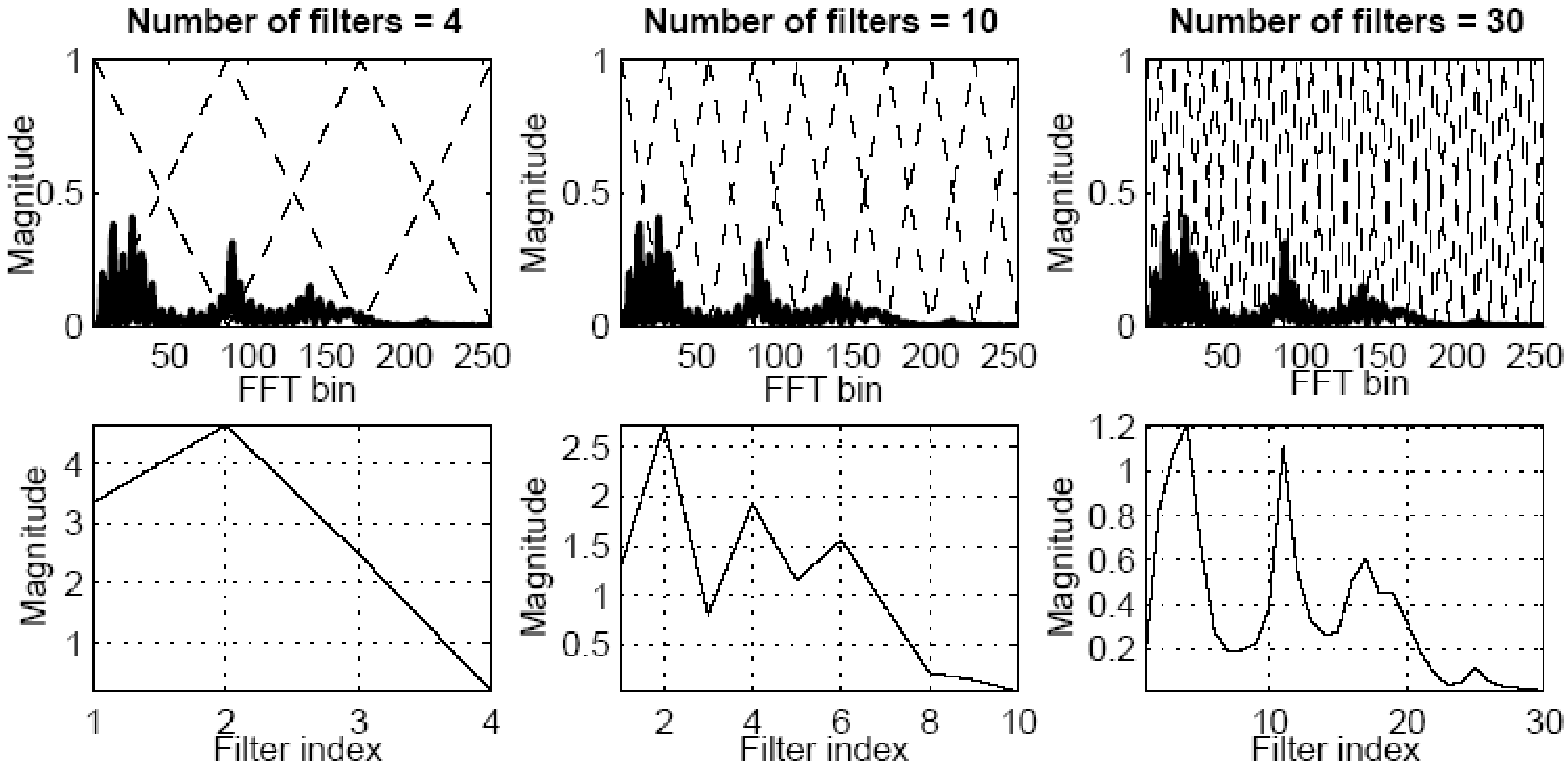


Ear: nonlinear transform

Mel-Frequency Cepstral Coefficients (MFCC) – a Perceptual Domain Feature



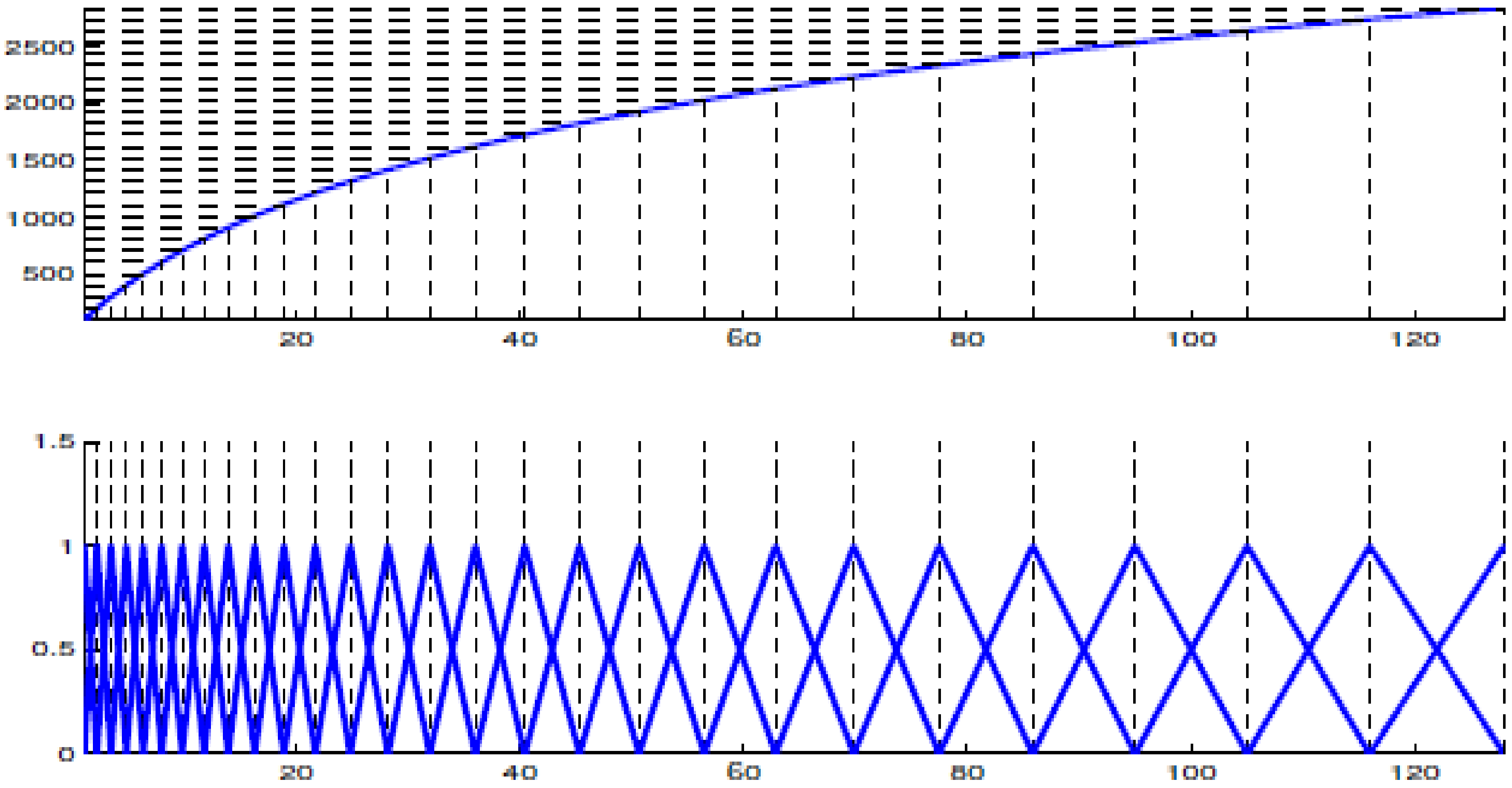
Spectral Envelope Estimation with a Filterbank



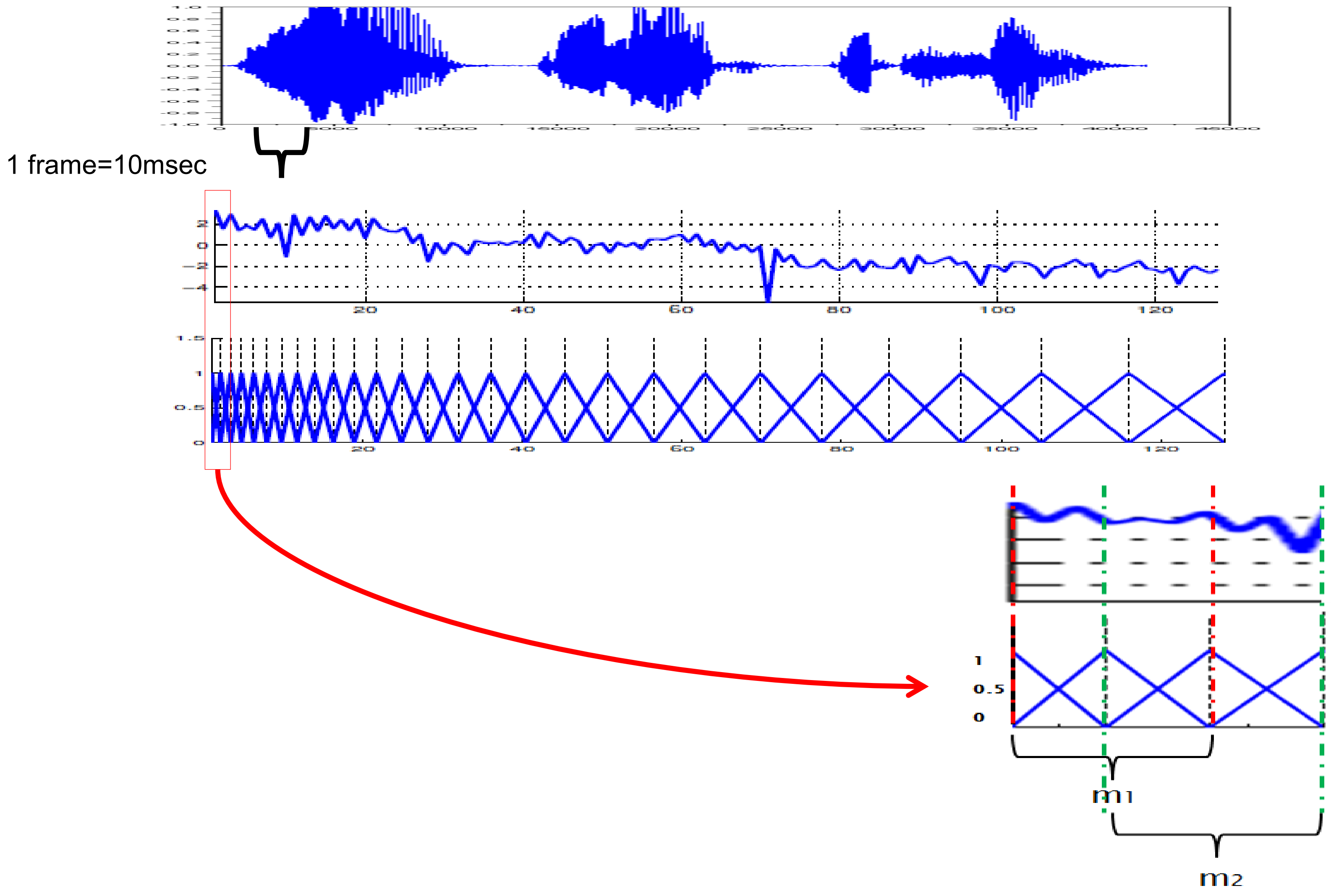
Each filter computes a weighted sum of the
FFT magnitude within that frequency band

Mel-Frequency Filterbank:

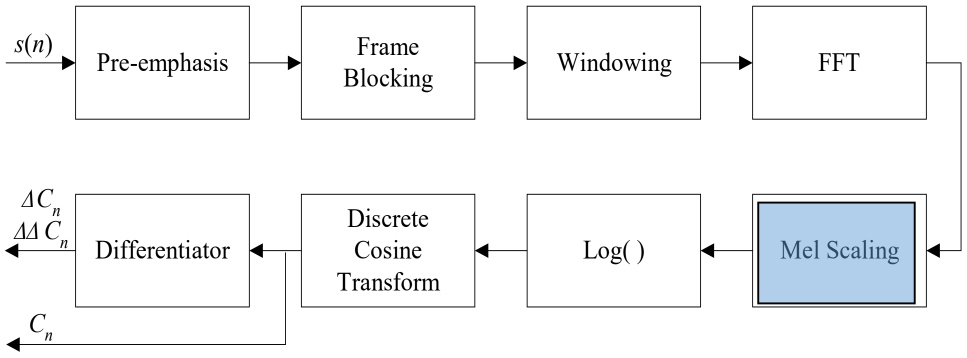
Mel Scale: $\quad \text{Mel}(f) = 1127 \log(1+f/700)$

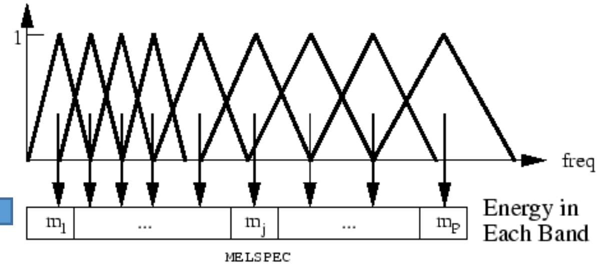


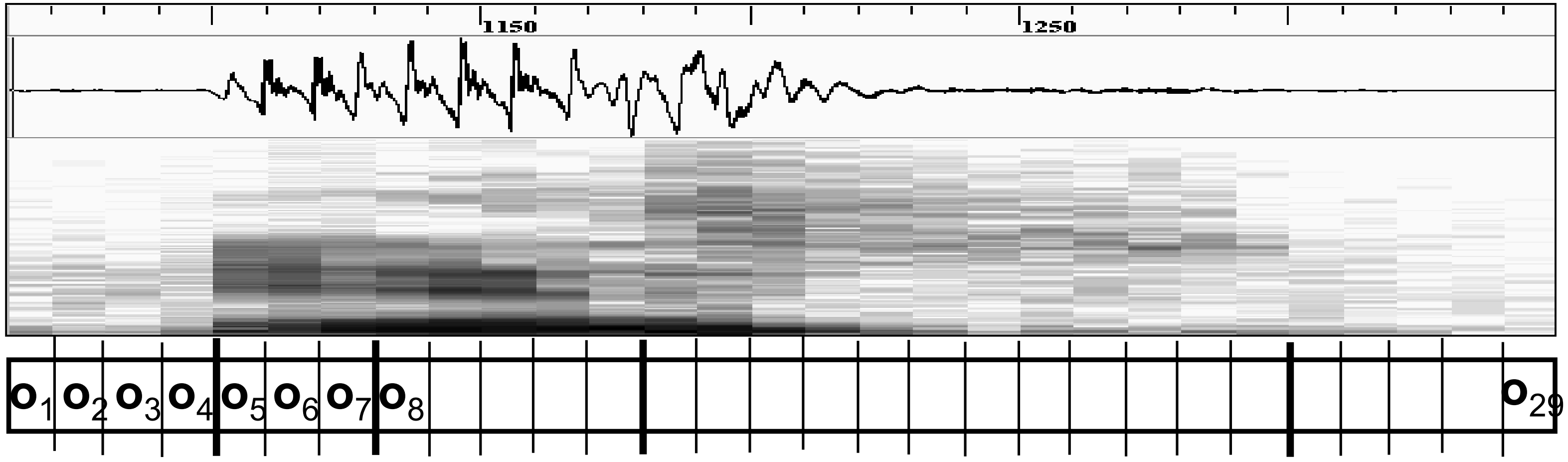
Mel-Frequency Filterbank:



MFCC Feature Vector (invented for ASR initially)







ML

Recap of DFT and Audio Representations

1) Recap of the key points of the first lecture

2) Recap of some insights of Discrete Fourier Transform (DFT)

3) Audio representation in time, frequency and perceptual domains