딥러닝 세미나

Tobig's 5기 박이삭

Audio 분석 기초

Contents

Unit 01 | 음성인식?

- 카카오, 바이두, 구글, 네이버, SK, NC_soft 등

여러 회사들이 있지만… 주목해야 될 회사는 ?

바이두

독자 적인 모델을 만들어 배포

RNN → Bi-directional RNN → Baidu

中 바이두, 음성인식 기술 애플 시리·구글 나우 보다 높아...기술 정확도 96% 2016년

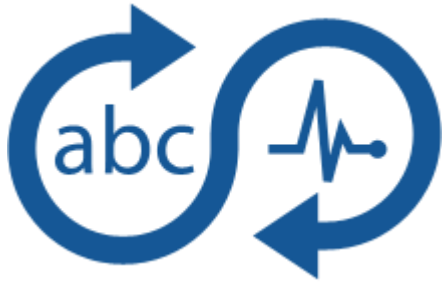
MK News - "바이두 음성인식 정확도 97%...車와 대화하는 시대 온다" 2017년

Kaggle 일반인 91% 2018년

바이두 : https://www.youtube.com/watch?v=g-sndkf7mCs&t=888s

Unit 01 | 음성인식?

- 음성인식 분야 – STT(speech to text) , TTS(text to speech)



활용 분야)

1. 자율주행 – (+ 강화학습)
2. Chat bot
3. AI 비서 (SK – Nugu, NAVER-clova(LG), Amazon – echo 등..

활용방안과 연구하는 기업이 많은 만큼!

이미지 인식(CNN, YOLO)보다 주목받는 사업이 되지 않을까 (개인적인 생각)

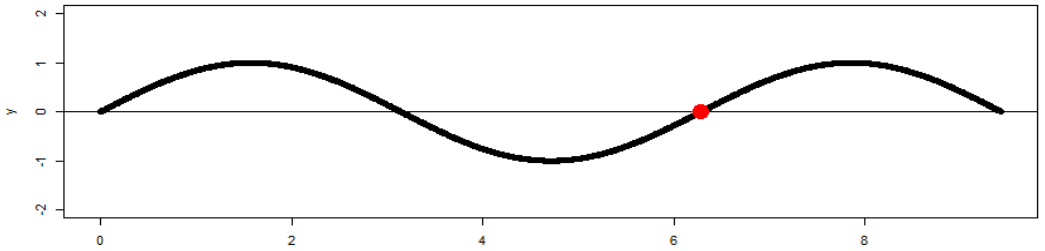
한국어 TTS: https://deview.kr/2017/schedule/182

삼각함수

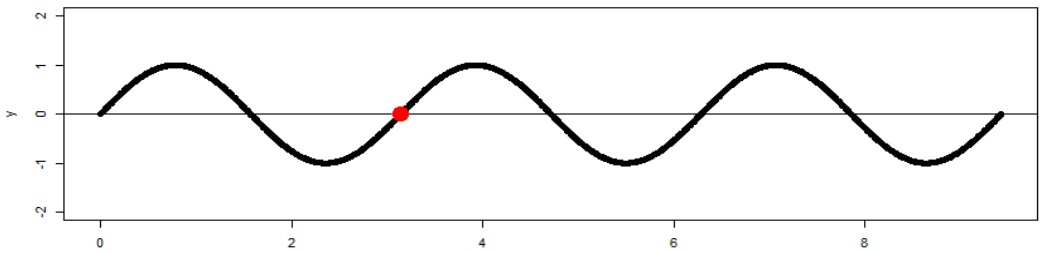
+내적, 허수, 오일러

Unit 02 | 삼각 함수

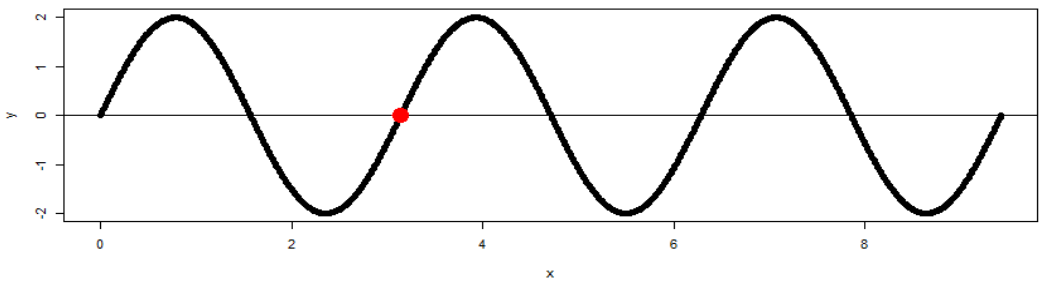
$\sin(x)$



$\sin(2x)$



$2\sin(2x)$



Sine 파는 (sinusoid, sine wave)

- 주기(frequency)
- 크기(amplitude)
- Cosine, sine 모두 사인파

$\sin(x)$ 의 주기는 보통 2π라고 한다.

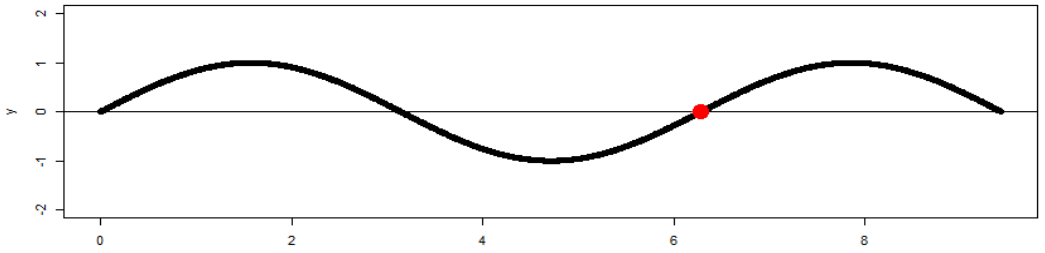
$\sin(2x)$의 경우 $\frac{2\pi}{2} = \pi$

$2\sin(2x)$의 amplitud는 $[-2,2]$

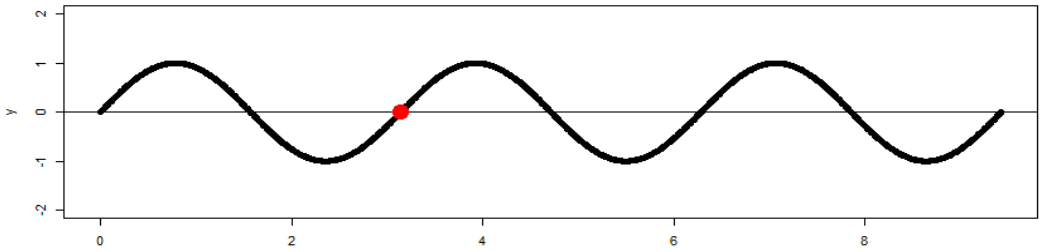
Analog 신호에서는 삼각함수가 통하지만…

Unit 02 | 삼각 함수

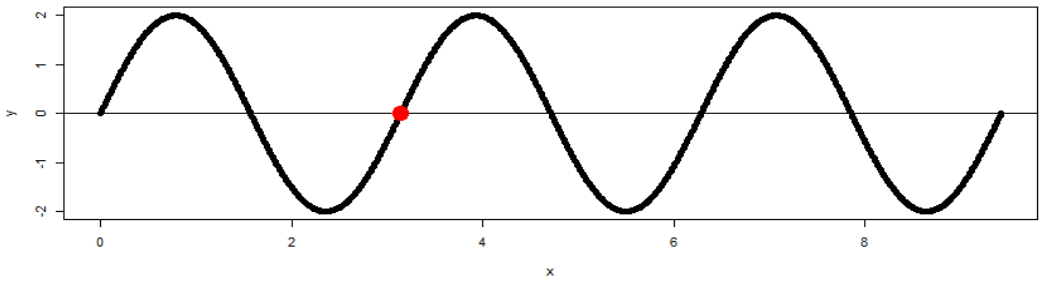
$\sin(x)$



$\sin(2x)$



$2\sin(2x)$

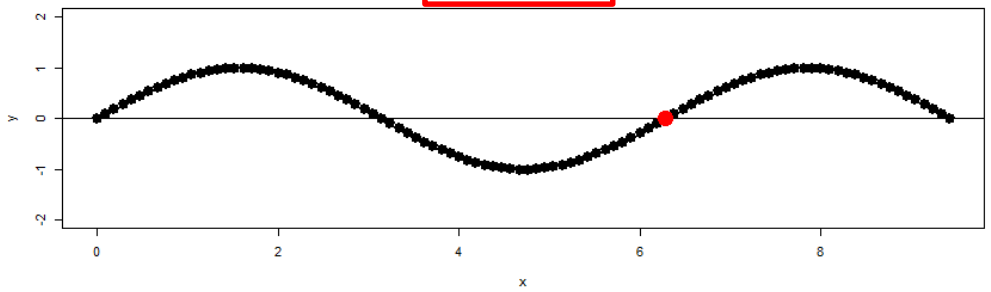


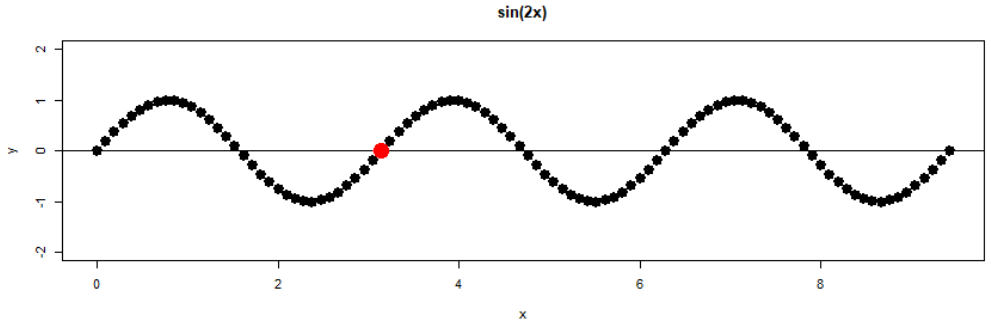
Analog

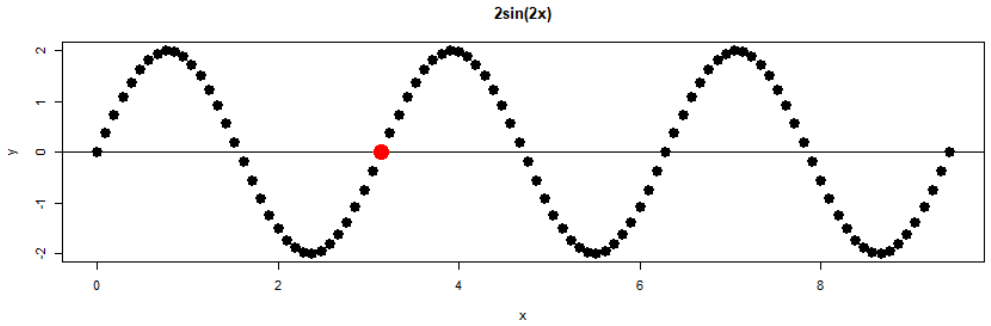
VS

Digital

sr = 100





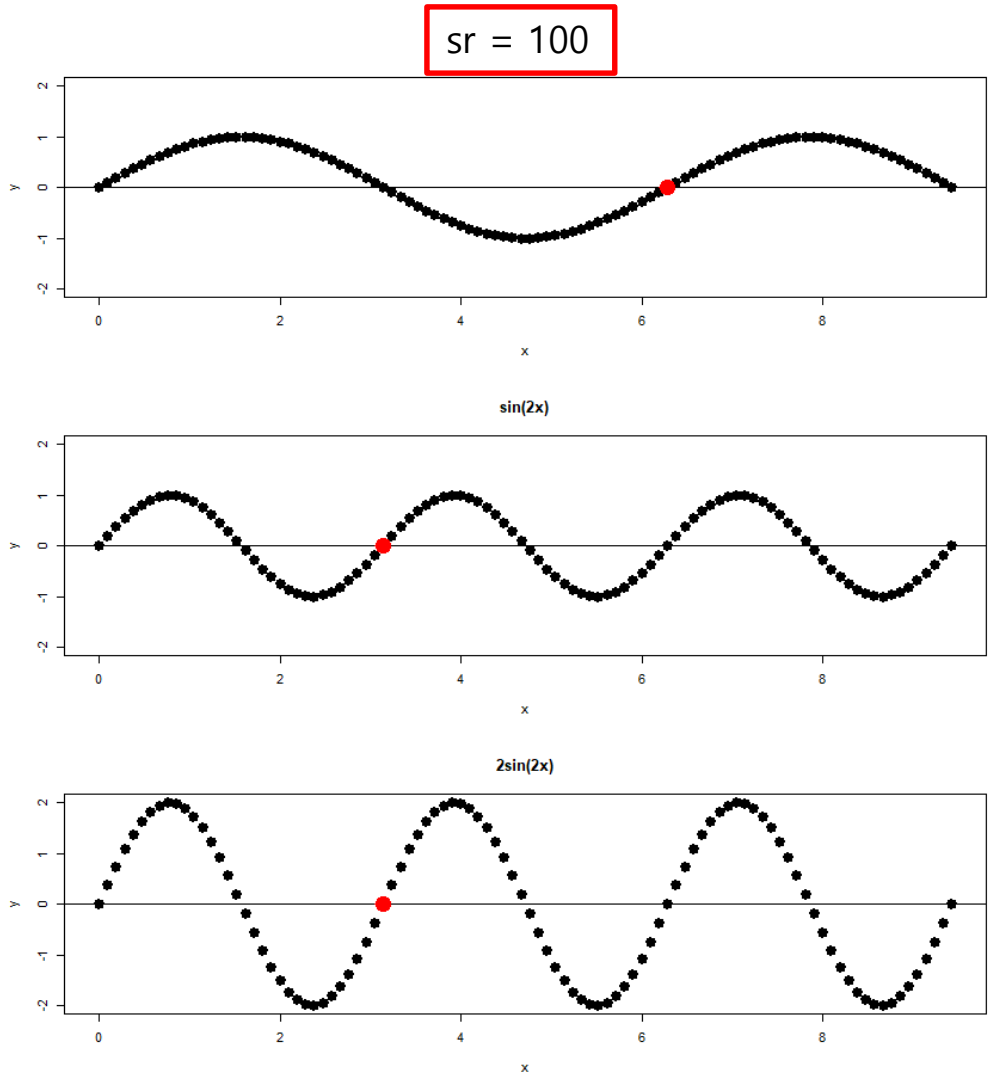


Unit 02 | 삼각 함수

컴퓨터상 연속을 다룰 수 없기에 적당한 sampling을 통해 신호를 이해하고 있다.

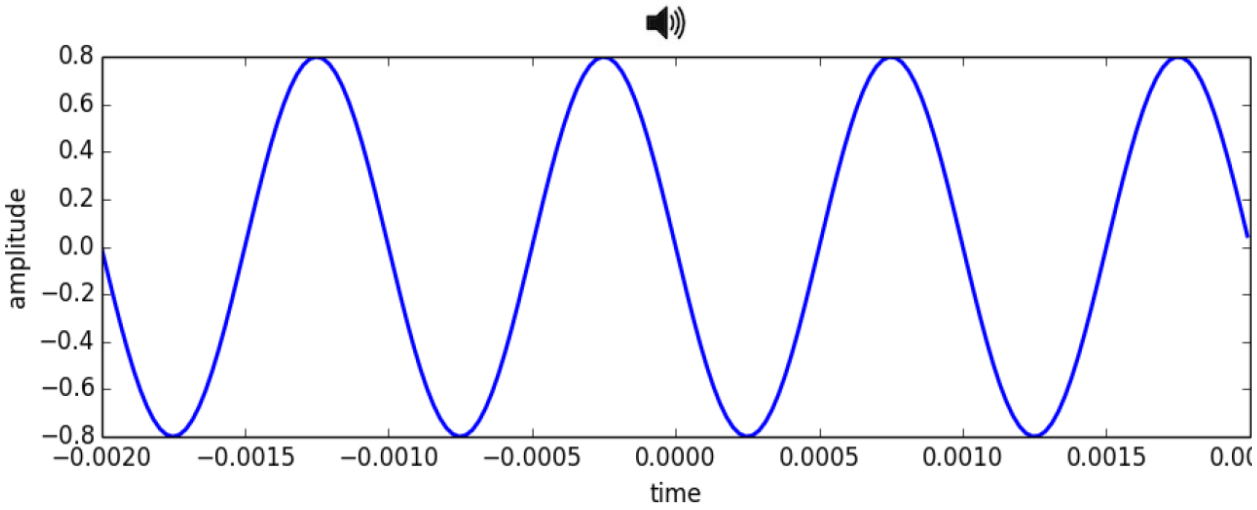
Sample추출 횟수를 Sampling rate(sr)라고 함.

오른쪽 그래프는
sr = 100
Amplitude = 1 & 2
Frequency = 2π & π



Unit 02 | 삼각 함수

Sinewave plot




```
A = .8
f0 = 1000
phi = np.pi/2
fs = 44100
t = np.arange(-.002, .002, 1.0/fs)
x = A * np.cos(2*np.pi*f0*t+phi)
```

1초 audio파일에 sr = 10 인 경우 T = 0.1, 0.2, ..

Sinusoidal functions (sinewaves)

$$x[n] = A\cos(\omega nT + \phi) = A\cos(2\pi f nT + \phi)$$

A : amplitude
ω : angular frequency in radians/seconds
$f = \omega / 2\pi$: frequency in Hertz (cycles/seconds)
ϕ : initial phase in radians
n : time index
$T = 1/f_s$: sampling period in seconds $(t = nT = n/f_s)$

출처 - Cosera: audio-signal-processing

Complex numbers

Q. Audio에 허수가?

->Sin 혹은 Cos을 테일러 급수로 근사 시킬수도 있지만,
오일러 공식을 이용한다면 간단하게 마무리 할 수 있기 때문.

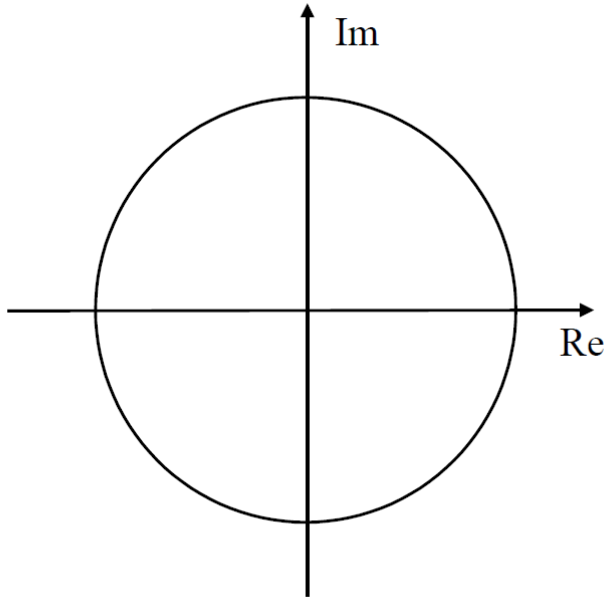
a, b: real numbers

$j = \sqrt{-1}$: imaginary unit

Complex plane:

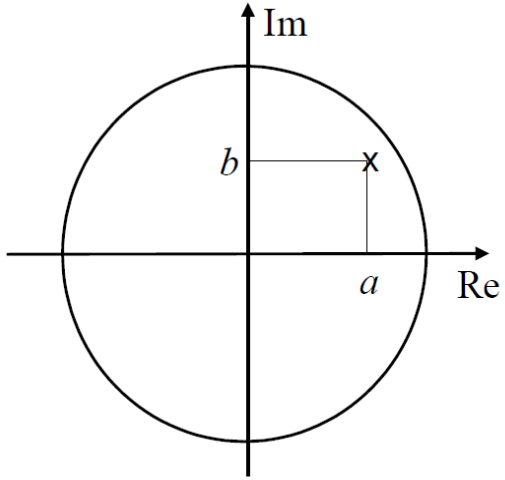
Re (real axis)

Im (imaginary axis)



Rectangular form:

$$(a + jb)$$



Polar form:

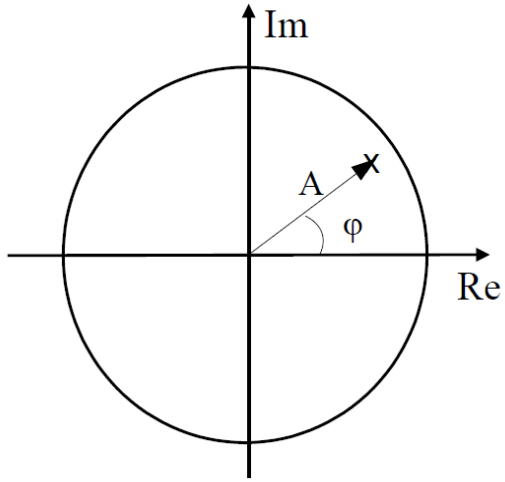
$$A = \sqrt{a^2 + b^2}$$

$$\phi = atan\,2\left(\frac{b}{a}\right)$$

where:

$$\text{if } (a>0) \quad atan\,2\left(\frac{b}{a}\right) = \tan^{-1}\left(\frac{b}{a}\right)$$

$$\text{else if } (a<0) \quad atan\,2\left(\frac{b}{a}\right) = \tan^{-1}\left(\frac{b}{a}\right) - pi$$

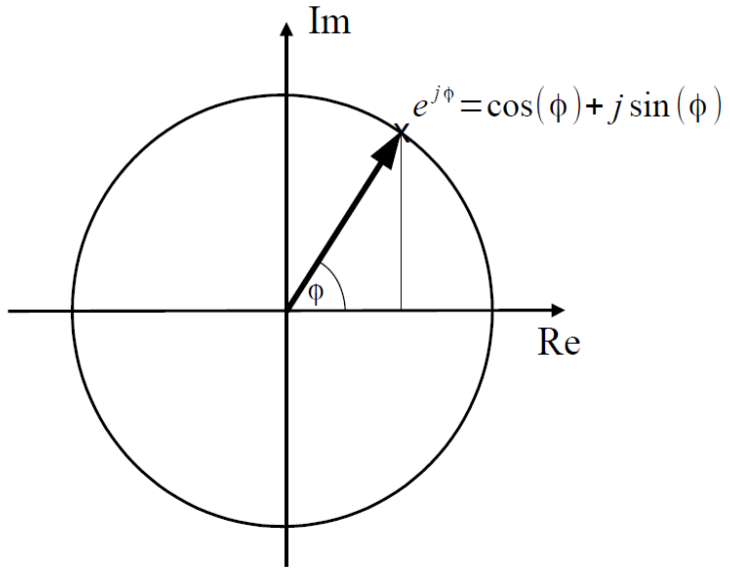


Euler's formula

$$e^{j\phi} = \cos\phi + j\sin\phi$$

$$\cos\phi = \frac{e^{j\phi} + e^{-j\phi}}{2}$$

$$\sin\phi = \frac{e^{j\phi} - e^{-j\phi}}{2j}$$



Complex sinewave

$$\bar{x}[n] = A e^{j(\omega nT+\phi)} = A e^{j\phi} e^{(j\omega nT)} = X e^{j(\omega nT)}$$
$$= A\cos(\omega nT+\phi) + j\, A\sin(\omega nT+\phi)$$

Real sinewave:

$$x[n] = A\cos(\omega nT+\phi) = A\left(\frac{e^{j(\omega nT+\phi)} + e^{-j(\omega nT+\phi)}}{2}\right)$$
$$= \frac{1}{2} X e^{j(\omega nT)} + \frac{1}{2} X^* e^{-j(\omega nT)} = \frac{1}{2}\bar{x}[n] + \frac{1}{2}\bar{x}^*[n]$$
$$= \Re\{\bar{x}[n]\}$$

Unit 04 | DFT

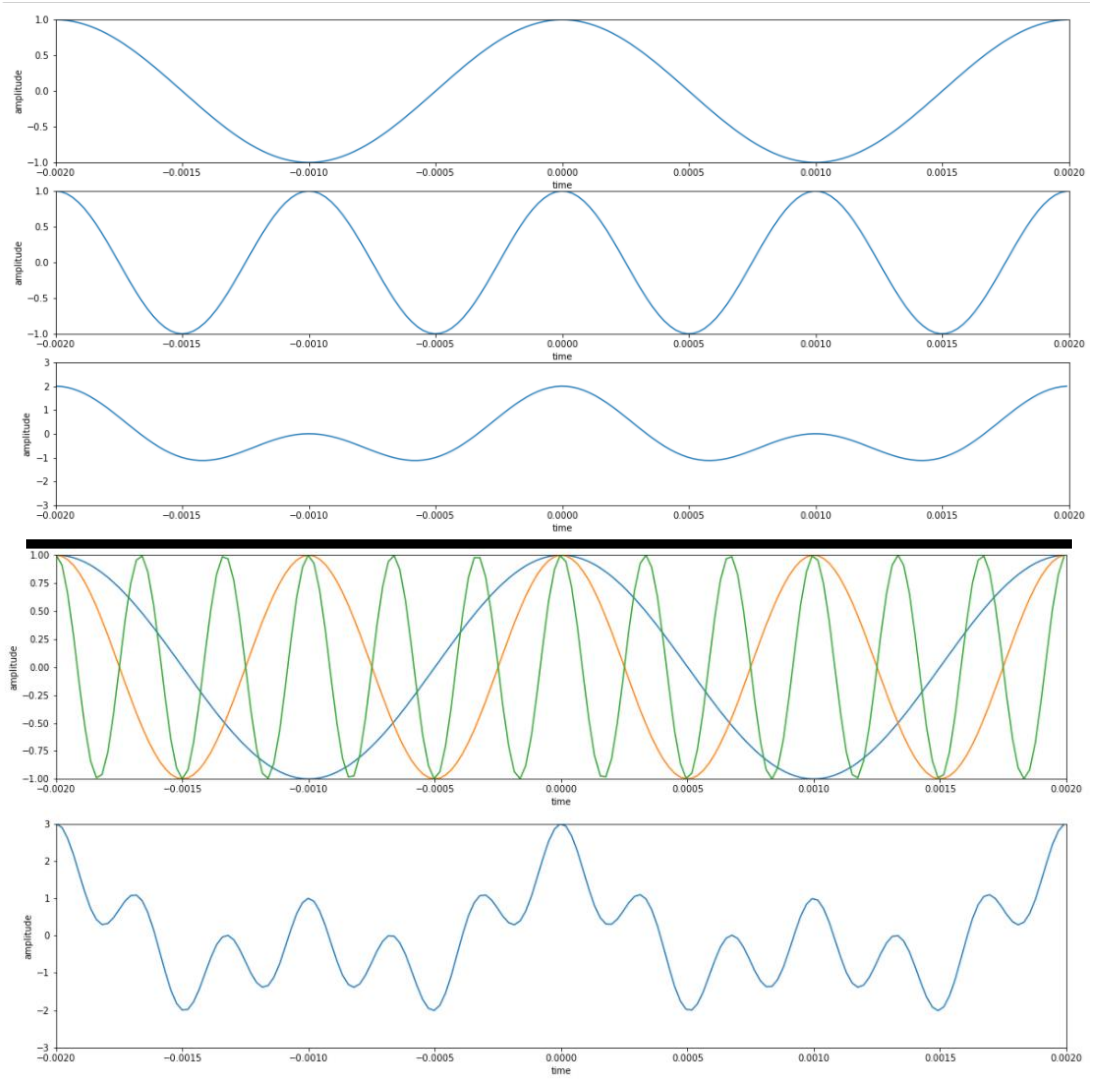
문제는... 실제 음성은 sin wave 처럼 간단하지 않다.

- Sin파 처럼 매끄럽지 않고 (울퉁불퉁)
- 여러 주파수가 합쳐져 있고 (위의 원인)
- 합쳐진 주파수가 몇 개나 될지 모른다.

→ 합성파를 분리할 수 없을까?

DFT(**D**iscrete **F**ourier **T**ransform)

- 주파수(Frequency)를 찾는 역할



Unit 04 | DFT

문제는... 실제 음성은 sin wave 처럼 간단하지 않다.

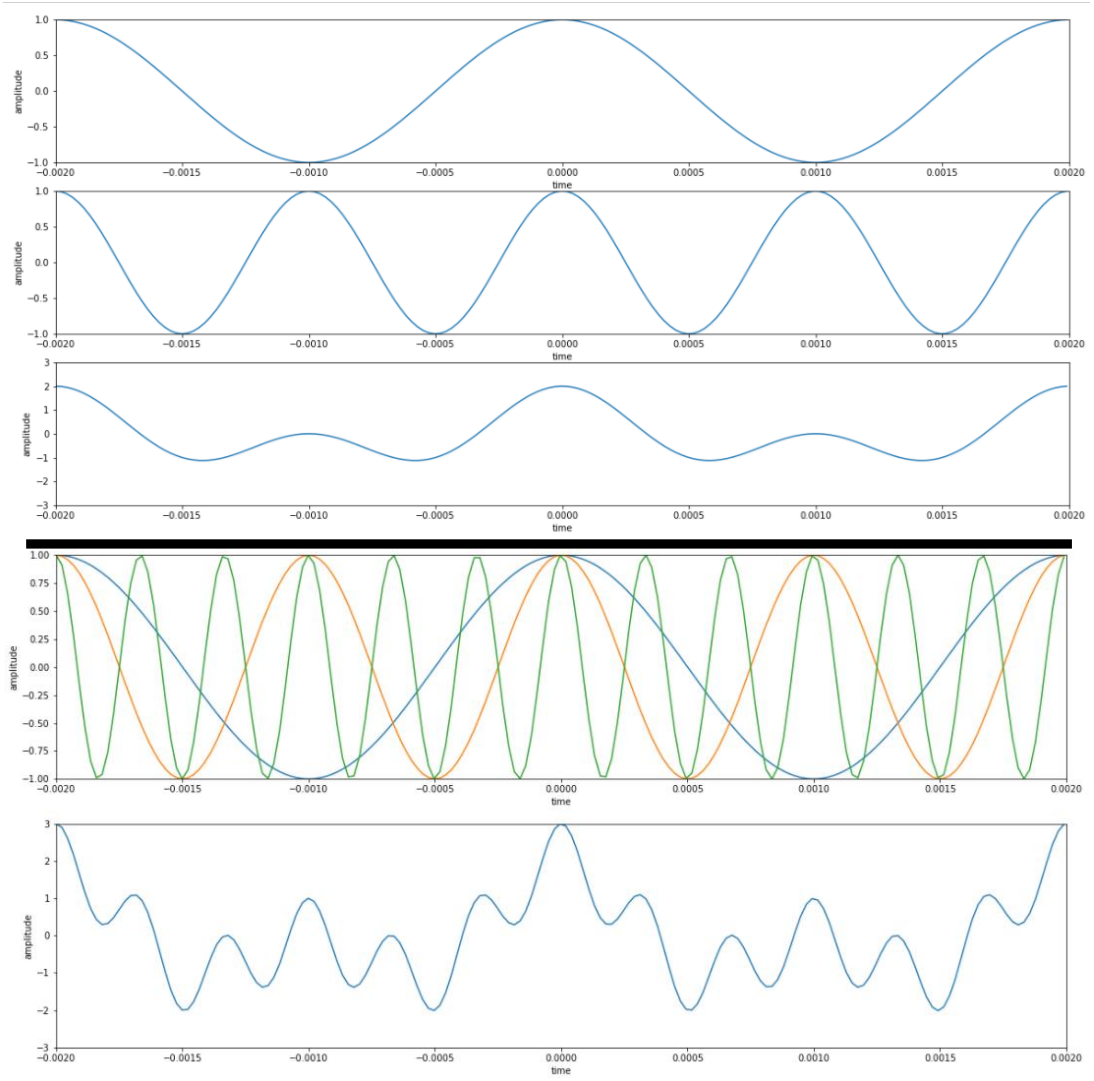
- Sin파 처럼 매끄럽지 않고 (울퉁불퉁)
- 여러 주파수가 합쳐져 있고 (위의 원인)
- 합쳐진 주파수가 몇 개나 될지 모른다.

→ 합성파를 분리할 수 없을까?

DFT(**D**iscrete **F**ourier **T**ransform)

- 주파수(Frequency)를 찾는 역할

How??



Discrete Fourier Transform

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

n : discrete time index (normalized time, $T=1$)
k : discrete frequency index
$\omega_k = 2\pi k/N$: frequency in radians
$f_k = f_s \frac{k}{N}$: frequency in Hz (f_s : sampling rate)

$\boldsymbol{x[n]}$: n번째 sample audio

$\boldsymbol{s_k}$: 오일러 공식 속 k번째 주파수(frequency)

N : 주파수(frequency) 탐색 범위

ex) sr = 100 , N = 4 일 때, <0, 25, 50, 75> 만 탐색
ex) sr = 100 , N = 100 일 때, 전부 탐색

DFT: complex exponentials

$$s_k^* = e^{-j2\pi kn/N} = \cos(2\pi kn/N) - j\sin(2\pi kn/N)$$

for $N=4$, thus for $n=0,1,2,3$; $k=0,1,2,3$

$$s_0^* = \cos(2\pi\times 0\times n/4) - j\sin(2\pi\times 0\times n/4) = [1,1,1,1]$$
$$s_1^* = \cos(2\pi\times 1\times n/4) - j\sin(2\pi\times 1\times n/4) = [1,-j,-1,j]$$
$$s_2^* = \cos(2\pi\times 2\times n/4) - j\sin(2\pi\times 2\times n/4) = [1,-1,1,-1]$$
$$s_3^* = \cos(2\pi\times 3\times n/4) - j\sin(2\pi\times 3\times n/4) = [1,j,-1,-j]$$

DFT: scalar product

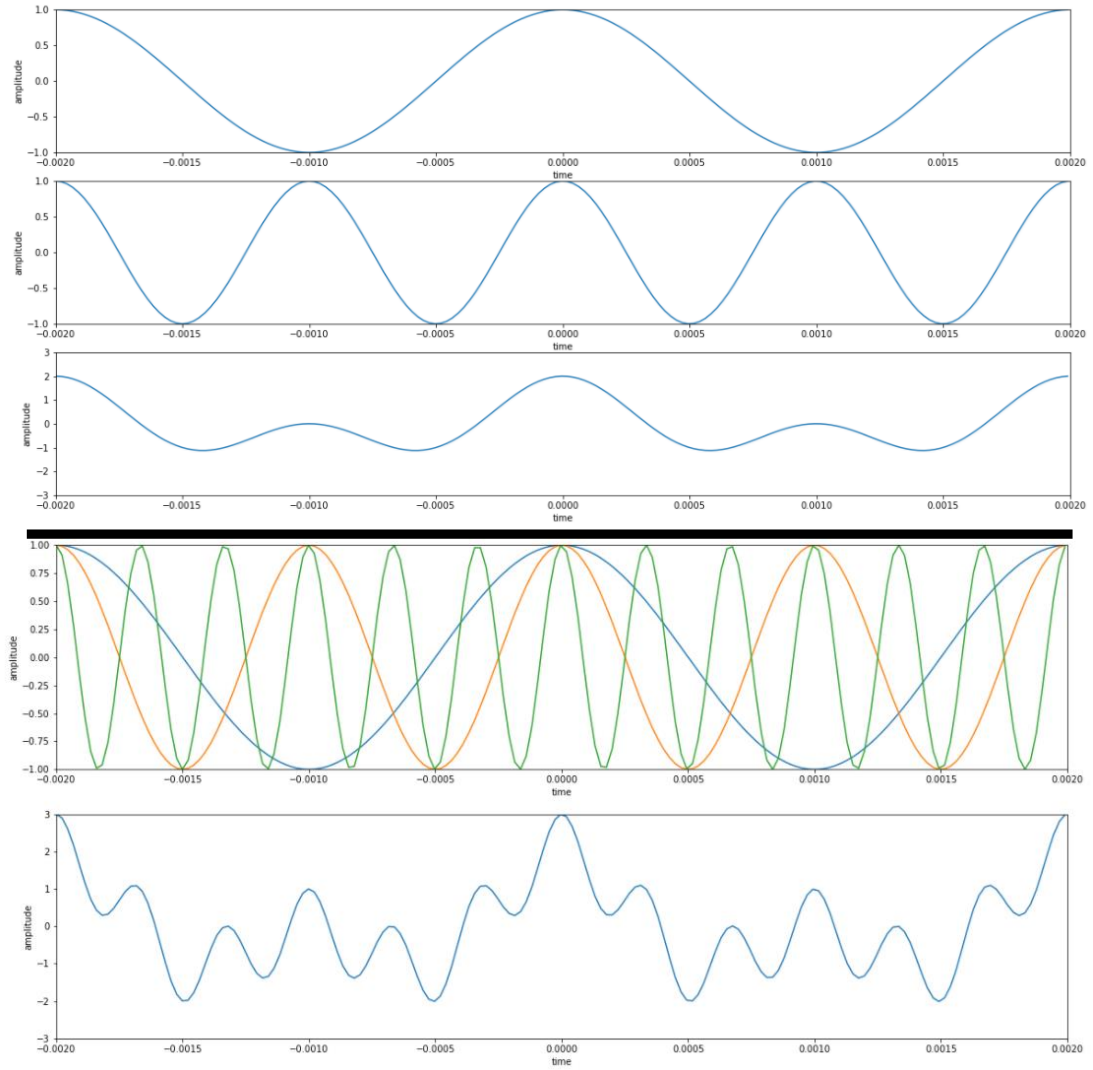
$$\langle x, s_k \rangle = \sum_{n=0}^{N-1} x[n] s_k^*[n] = \sum_{n=0}^{N-1} x[n] e^{-j 2\pi kn/N}$$

Example:

$$x[n] = [1, -1, 1, -1]; N = 4$$

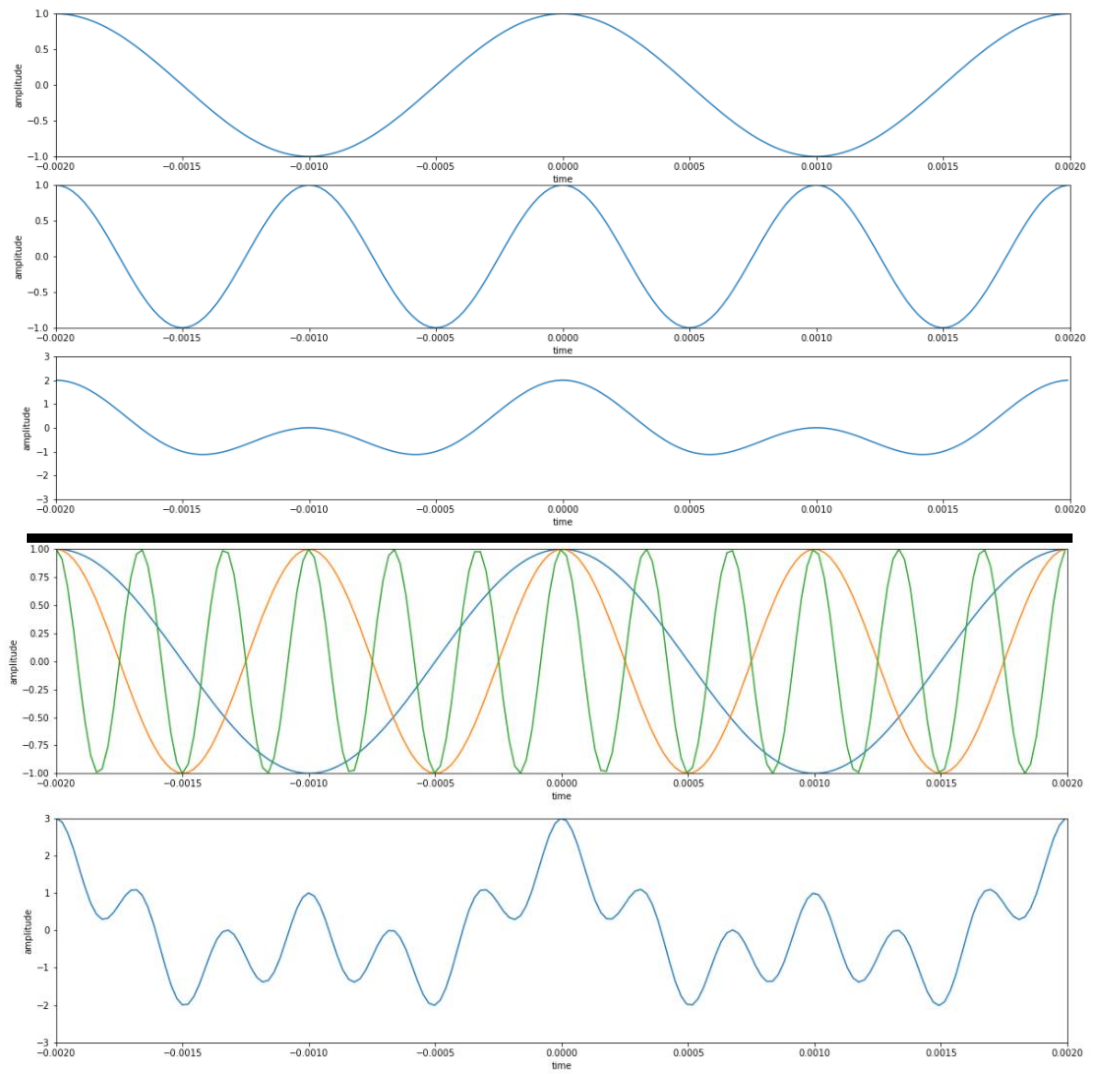
$$\langle x, s_0 \rangle = 1 \times 1 + (-1) \times 1 + 1 \times 1 + (-1) \times 1 = 0$$
$$\langle x, s_1 \rangle = 1 \times 1 + (-1) \times (-j) + 1 \times (-1) + (-1) \times j = 0$$
$$\langle x, s_2 \rangle = 1 \times 1 + (-1) \times (-1) + 1 \times 1 + (-1) \times (-1) = 4$$
$$\langle x, s_3 \rangle = 1 \times 1 + (-1) \times j + 1 \times (-1) + (-1) \times (-j) = 0$$

Unit 04 | DFT

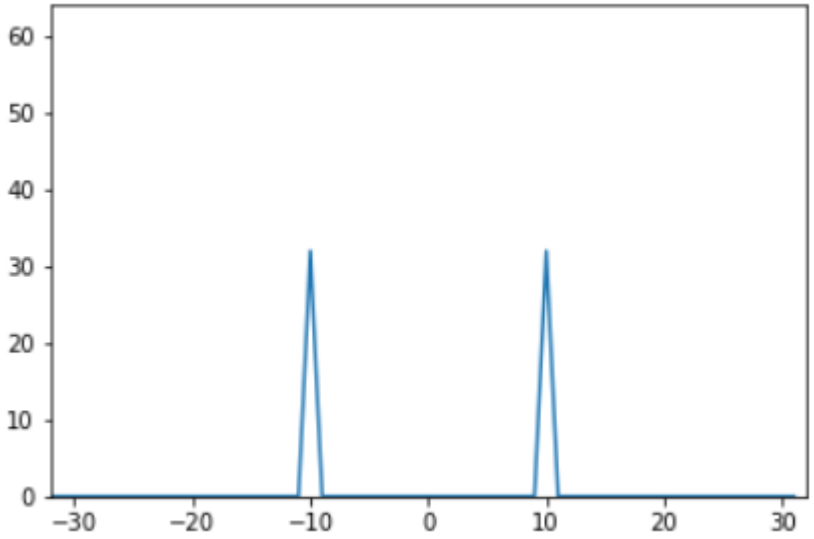


그럼, 이제 합성파들을 분리 해 볼까요?

Unit 04 | DFT



그럼, 이제 합성파들을 분리 해 볼까요?



X축이 시간 → 주파수(frequency)로 변했습니다.

이제부터 이런 그림을 spectrum 이라고 명명하겠습니다.

(Y축 : Amplitude)

문제는... 노래 한 곡 4분, 회의록 10분, 전화통화 2시간

꽤 긴 시간동안 이뤄진 녹음을 한꺼번에 표시한다??

Frequency가 Audio분석에 중요한 것임을 알았지만...

- 회의를 하다 노래를 할 수도 있고
- 전화 중 노래를 할 수도 있고

→ 아무리 긴 시간의 audio signal이 들어와도 제대로 분석 하기 위해선, 작은 시간동안 frequency의 변화를 살펴야 되지 않을까?

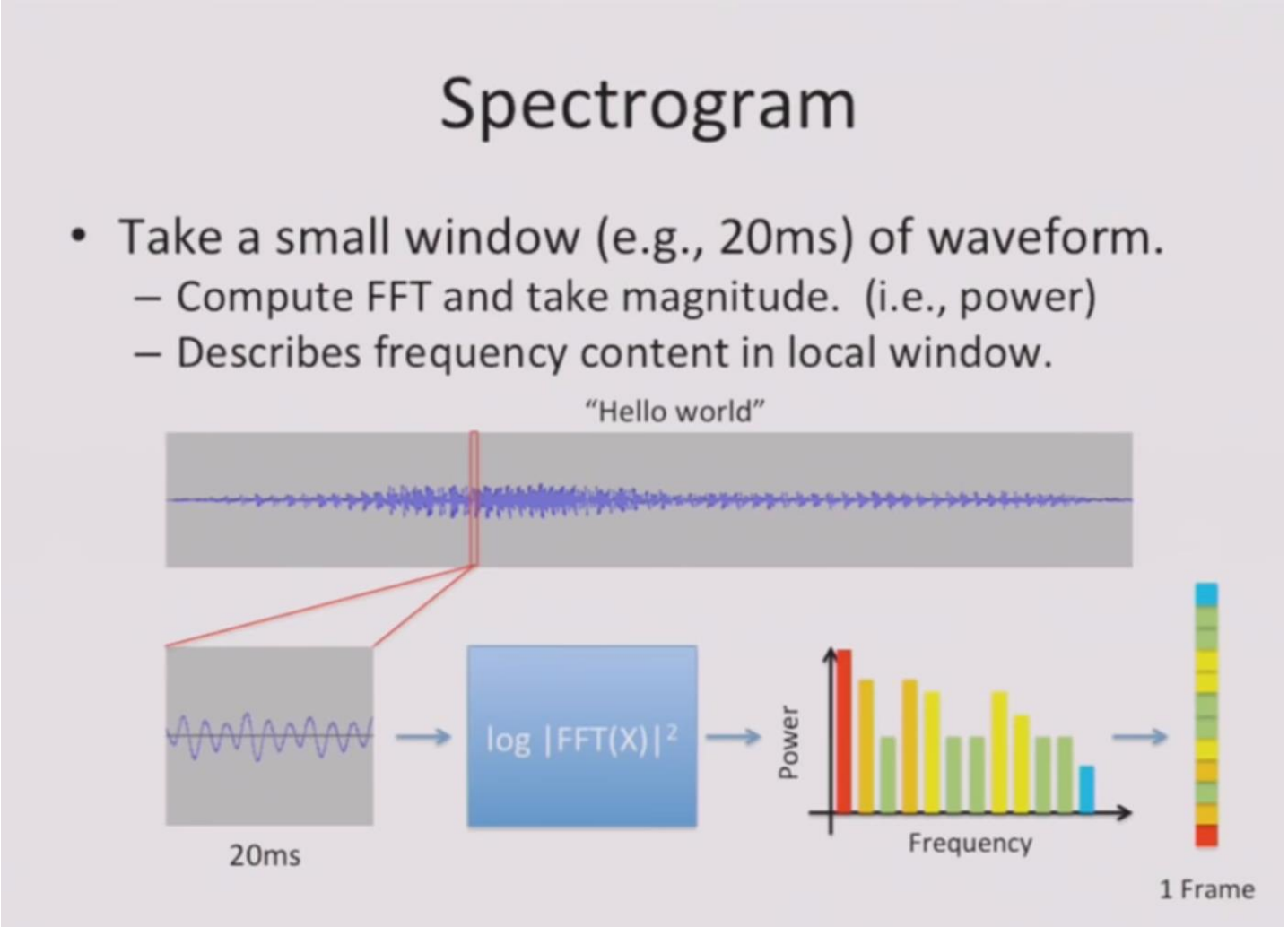
문제는... 노래 한 곡 4분, 회의록 10분, 전화통화 2시간 꽤 긴 시간동안 이뤄진 녹음을 한꺼번에 표시한다??

Frequency가 Audio분석에 중요한 것임을 알았지만...

- 회의를 하다 노래를 할 수도 있고
- 전화 중 노래를 할 수도 있고

→ 아무리 긴 시간의 audio signal이 들어와도 제대로 분석 하기 위해선, 작은 시간동안 frequency의 변화를 살펴야 되지 않을까?

FFT: Fast fourier transform



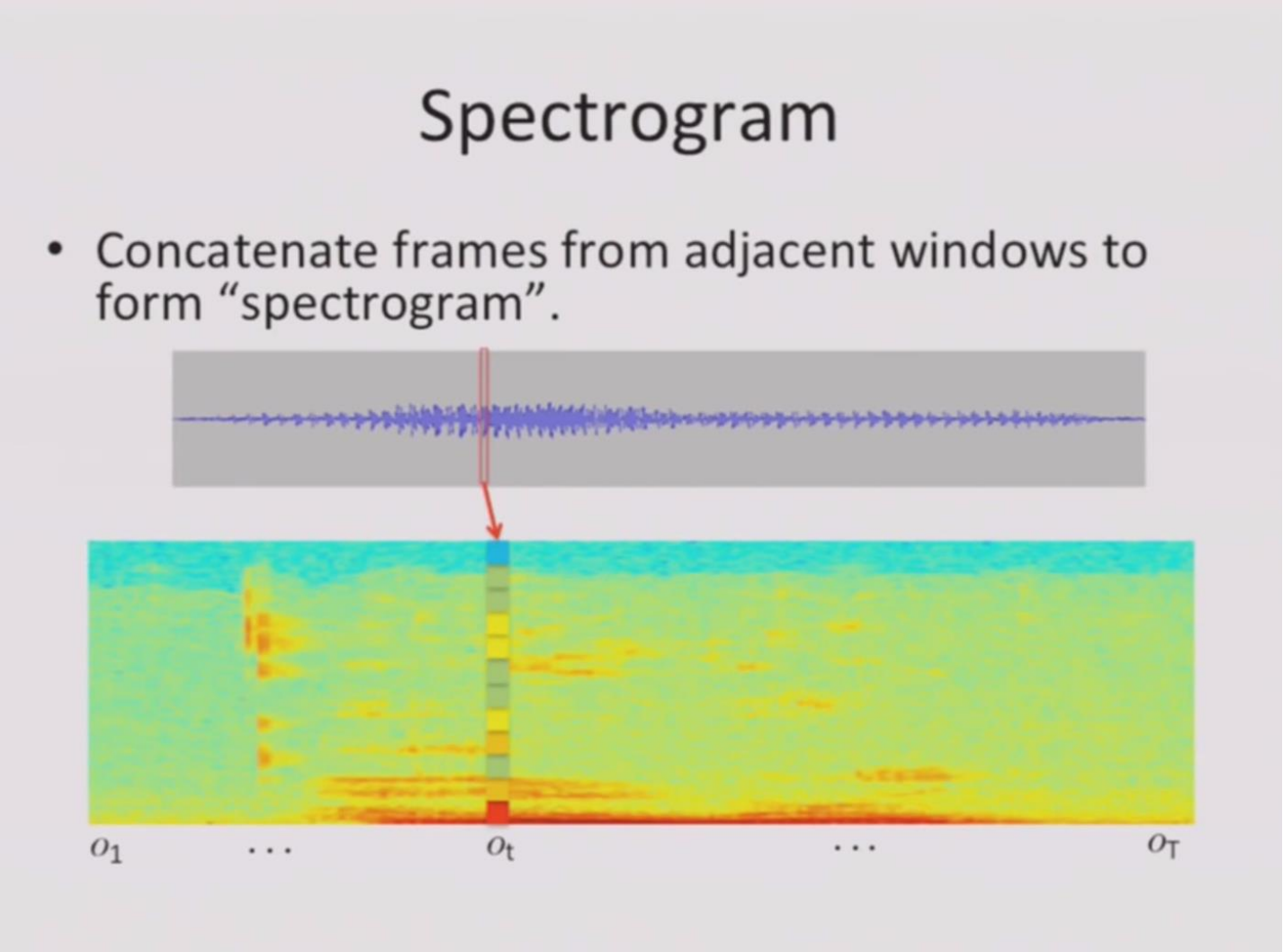
※짧은 구간안에서 소리는 변화하지 않는다고 가정

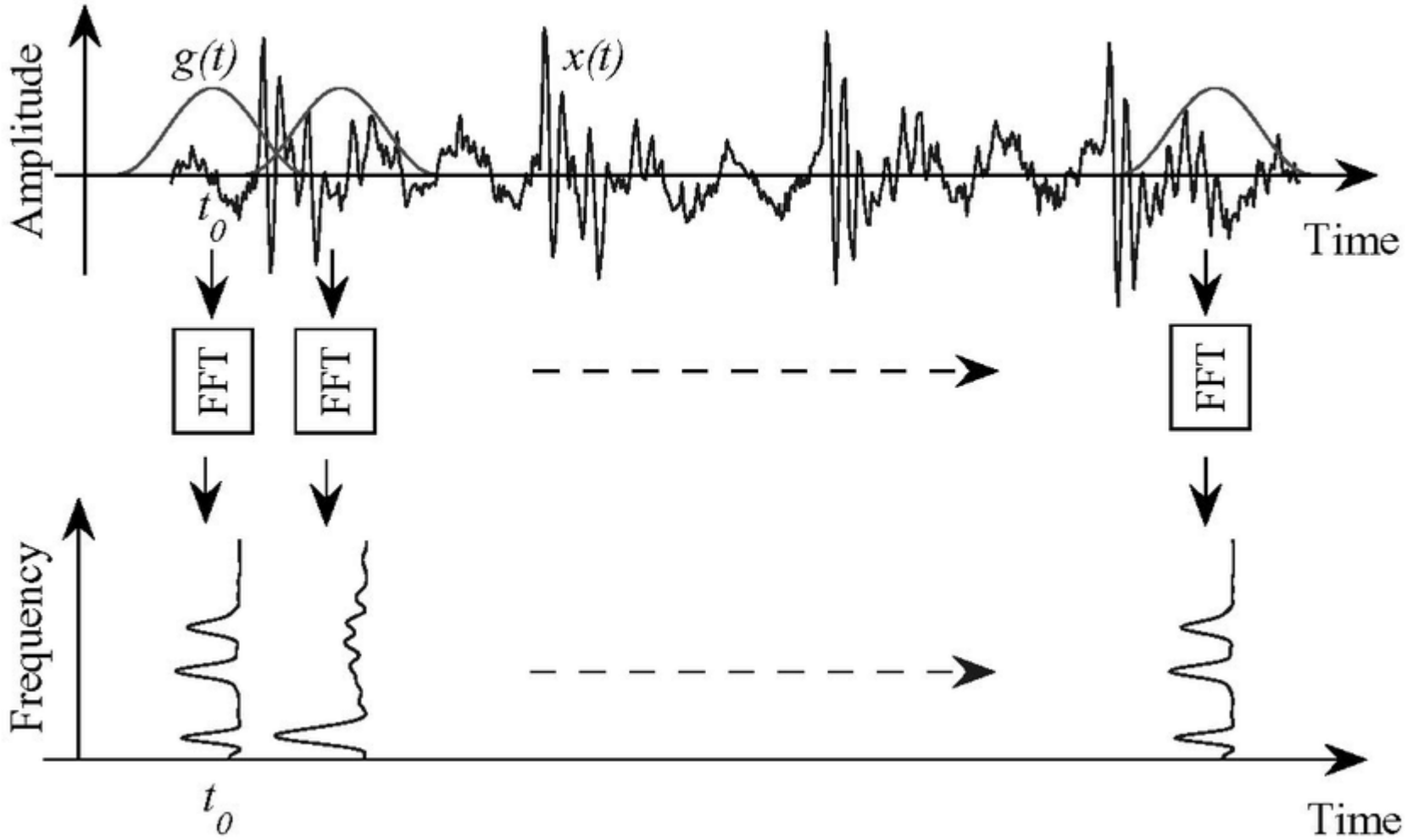
< spectrogram >

→ X: time, Y: frequency, Z: amplitude(진폭)

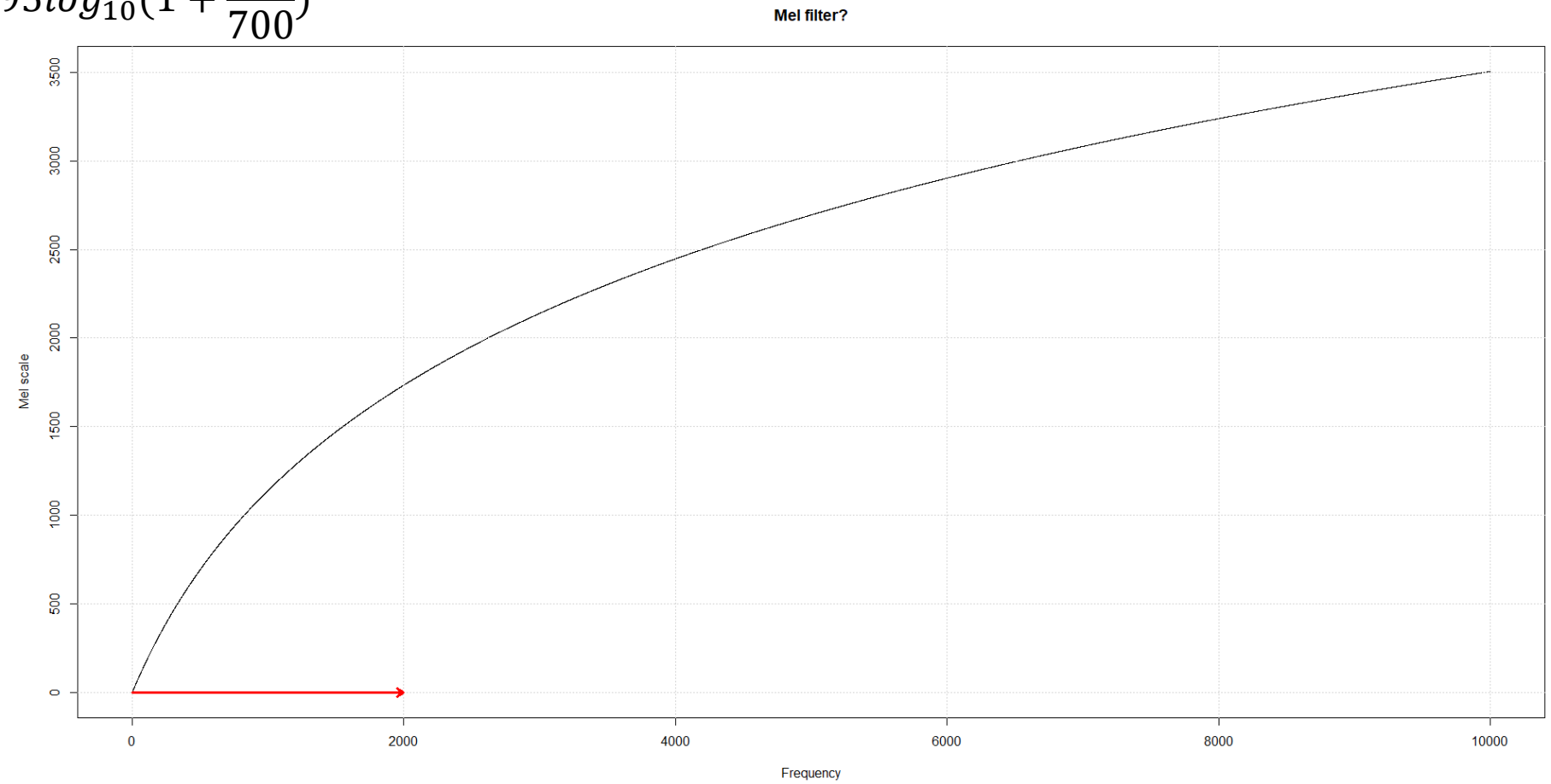
이제,

- 시간의 변화에 민감해 졌다.
- 각 시간별 주파수를 탐지 가능
- window size = 20ms 기본

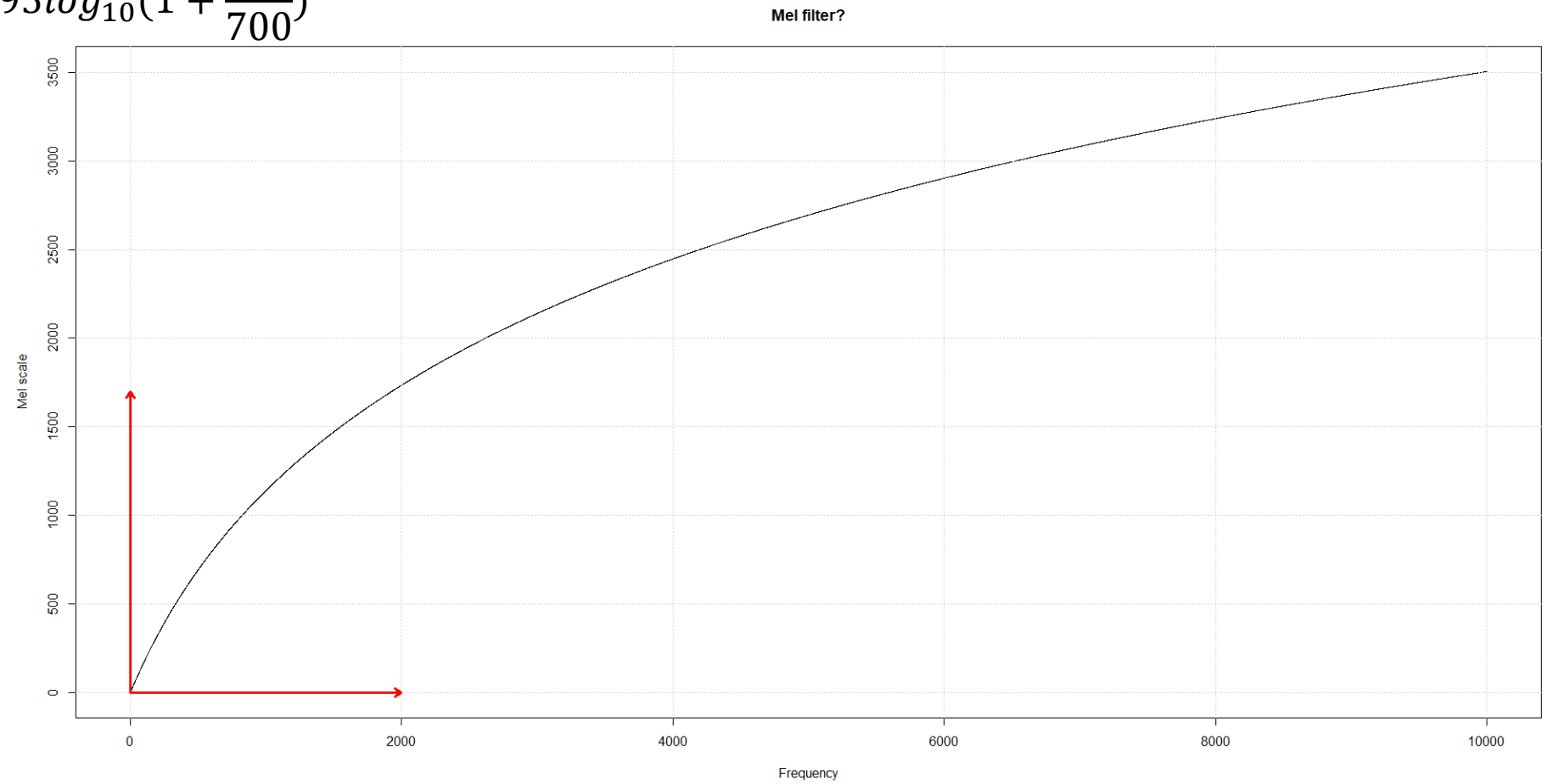



Amplitude
g(t)
x(t)
t_0
Time
FFT
FFT
FFT
Frequency
t_0
Time

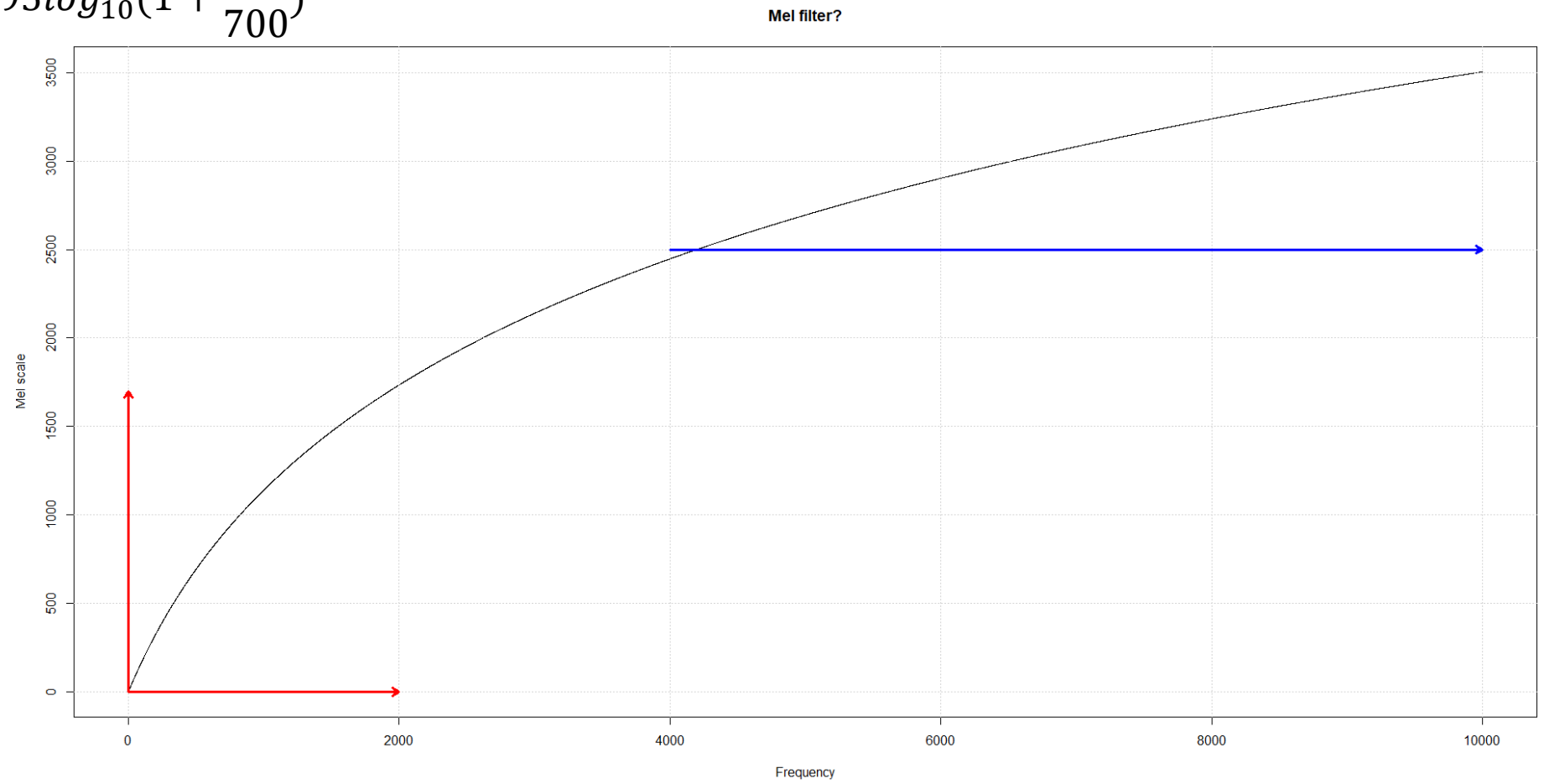
$$mel = 2595log_{10}(1 + \frac{f}{700})$$



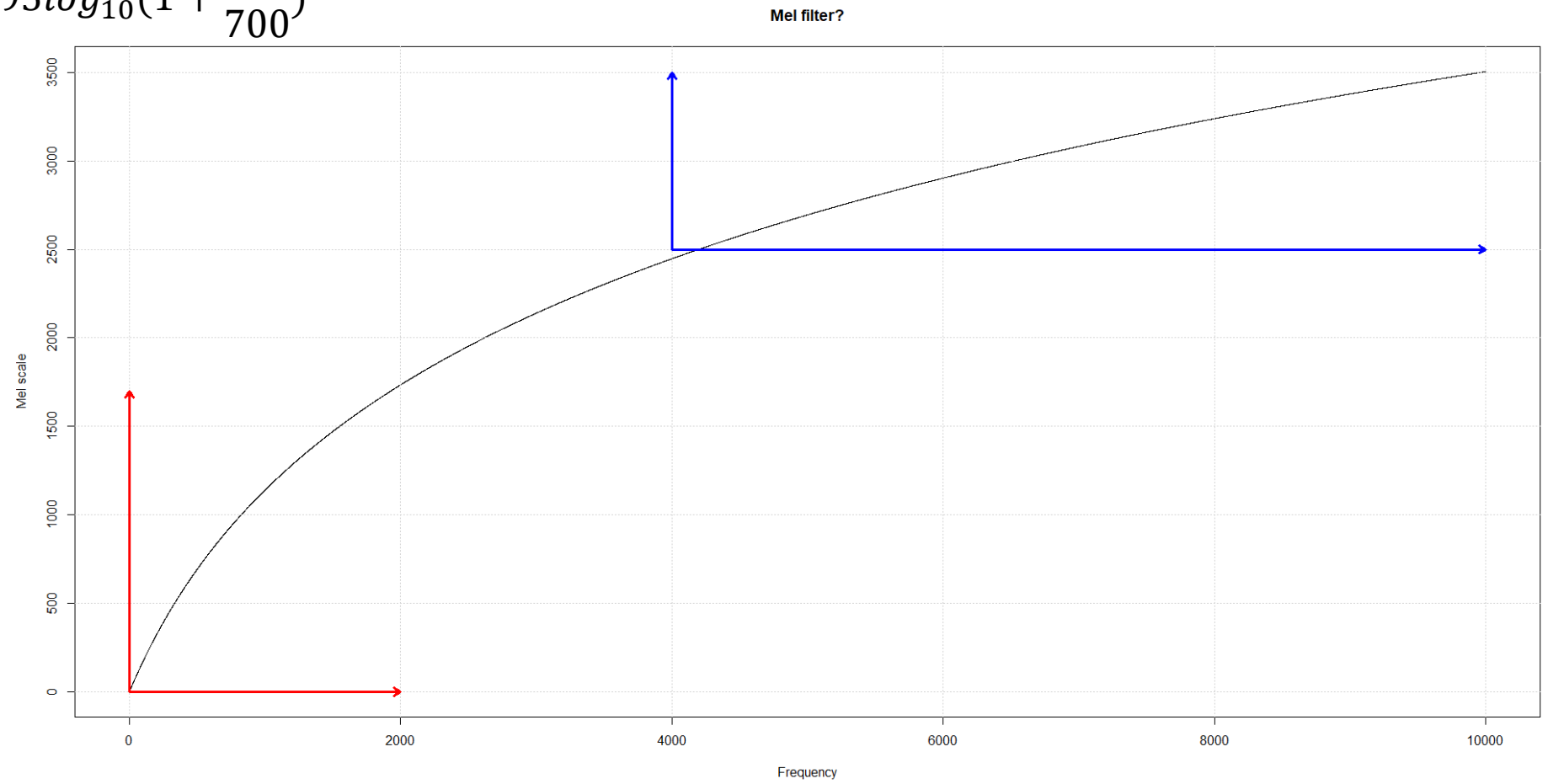
$$mel = 2595log_{10}(1 + \frac{f}{700})$$



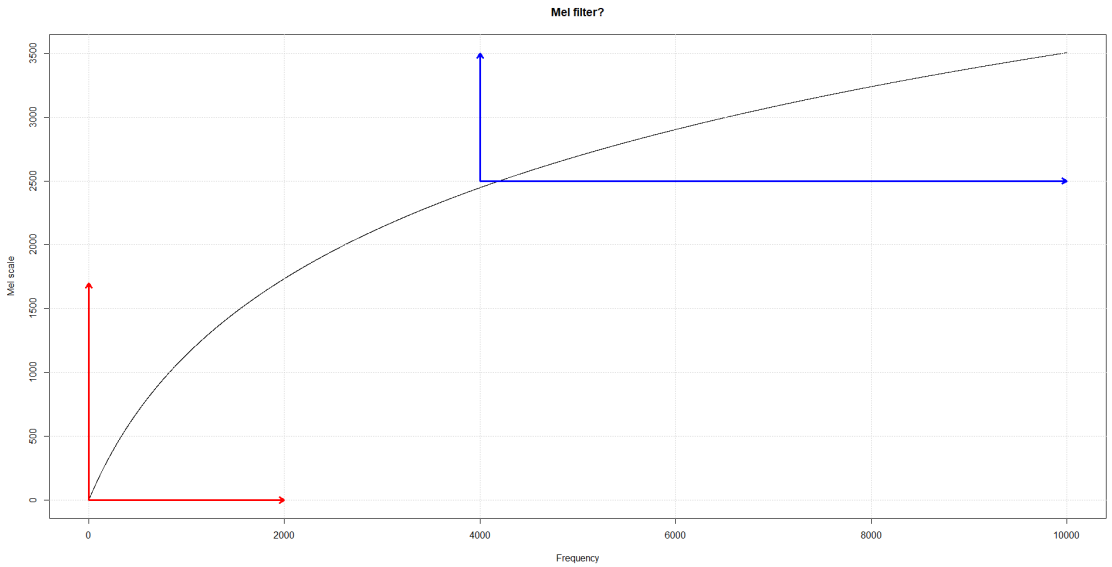
$$mel = 2595log_{10}(1 + \frac{f}{700})$$



$$mel = 2595log_{10}(1 + \frac{f}{700})$$



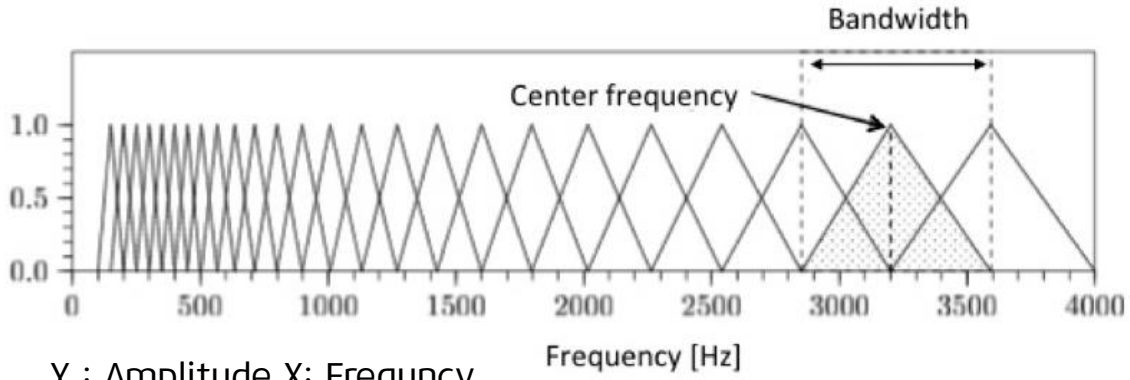
$$mel = 2595log_{10}(1 + \frac{f}{700})$$



- Mel scale를 통해 frequency들을 인간이 듣는 형태로 변형을 함.

→ **문제는...** 사람은 100f 와 110f 를 구분하지 못한다.

$$mel = 2595log_{10}(1 + \frac{f}{700})$$



Y : Amplitude X: Frequncy

- Mel scale를 통해 frequency들을 인간이 듣는 형태로 변형을 함.

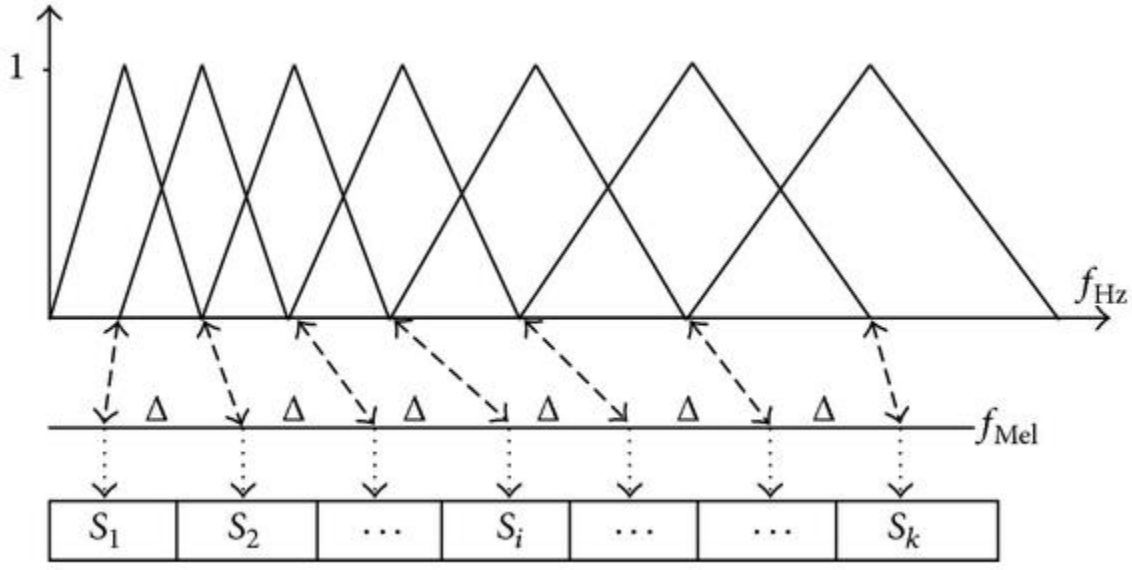
→ **문제는...** 사람은 100f 와 110f 를 구분하지 못한다.

인접 주파수들을 구분하지 못하기 때문에 음성인식에 적용하기 힘들다.

따라서 **구간** 을 나누어서 분석을 한다. Mel Filter Bank

- Filter별로 에너지($\sum_{k=0}^{N-1} |X_l[k]|^2$)를 측정

$$mel = 2595log_{10}(1 + \frac{f}{700})$$



- Mel scale를 통해 frequency들을 인간이 듣는 형태로 변형을 함.

→ **문제는…** 사람은 100f 와 110f 를 구분하지 못한다.

인접 주파수들을 구분하지 못하기 때문에 음성인식에 적용하기 힘들다.

따라서 **구간** 을 나누어서 분석을 한다. Mel Filter Bank

- Filter별로 에너지 $\sum_{k=0}^{N-1} |X_l[k]|^2$ 를 측정 $(l\ is\ mel\ filter)$
- 각 필터별, DCT(**D**iscrete **C**osine **T**ransform)를 취한다.

Mel frequency cepstral coefficients

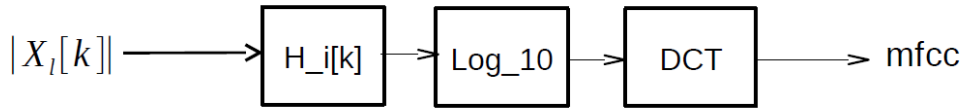
$$mfcc_l = DCT\left(\log_{10}\left(\sum_{k=0}^{N/2} |X_l[k]| H_i[k]\right)\right)$$

where

$|X[k]|$ is the positive magnitude spectrum

$H_i[k]$ is the mel scale filter bank for each filter i

$$DCT[m](\text{Discrete Cosine Transform}) = \sum_{n=0}^{N-1} f[n]\cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)m\right)$$



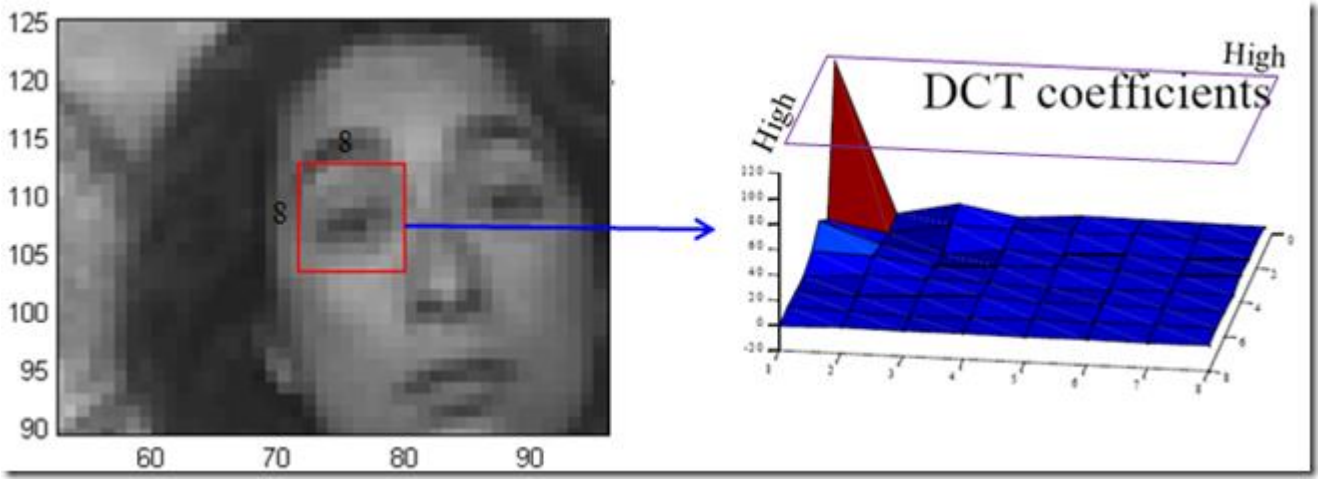
$\sum_{k=0}^{N-1} |X_l[k]|^2$ 의 꼴이 아닌 이유?

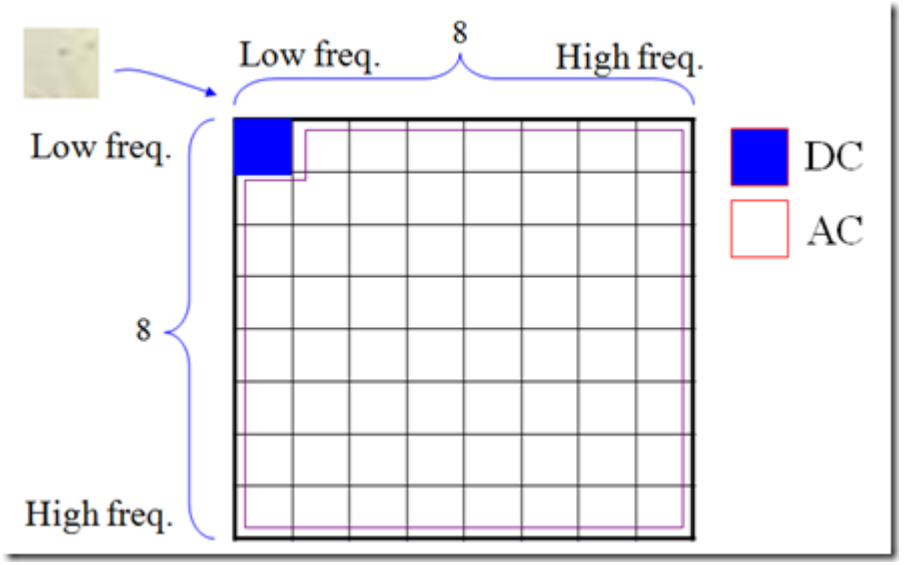
필터 별 에너지를 계산하기 위해.

DCT?

필터간 겹쳐진 영역이 존재하기 때문에 이를 분리하는 역할
(cos similarity와 반대 개념이라고 생각)

DCT?





DCT를 하는 이유 :

이미지에서 인접한 pixel은 비슷한 색상을 가진다.

64개의 pixel중 같은 색상이 낮은 주파수로 몰리게 됩니다.

색상의 변화가 있는 경우 높은 주파수로 위치하게 됩니다.

사람의 눈(귀)는 낮은 주파수(DC)에 민감하게 반응하지만 높은 주파수에선 아니기 때문에 높은 주파수 영역을 생략한다 해도 화질의 차이를 못 느낌.

→ 동영상, 사진 압축의 원리.

http://crynut84.tistory.com/61

그 밖에...

- Energy, RMS, Loudness
- Spectral centroid
- Pitch
- Chroma
- Onset

등등...feature들은 많이 있음!!!

https://www.coursera.org/learn/audio-signal-processing/lecture/ZRurD/audio-features

- Librosa 패키지 설명서: https://librosa.github.io/librosa/

들어주셔서 감사합니다.