디지털 음성 처리

강사 : 김정인

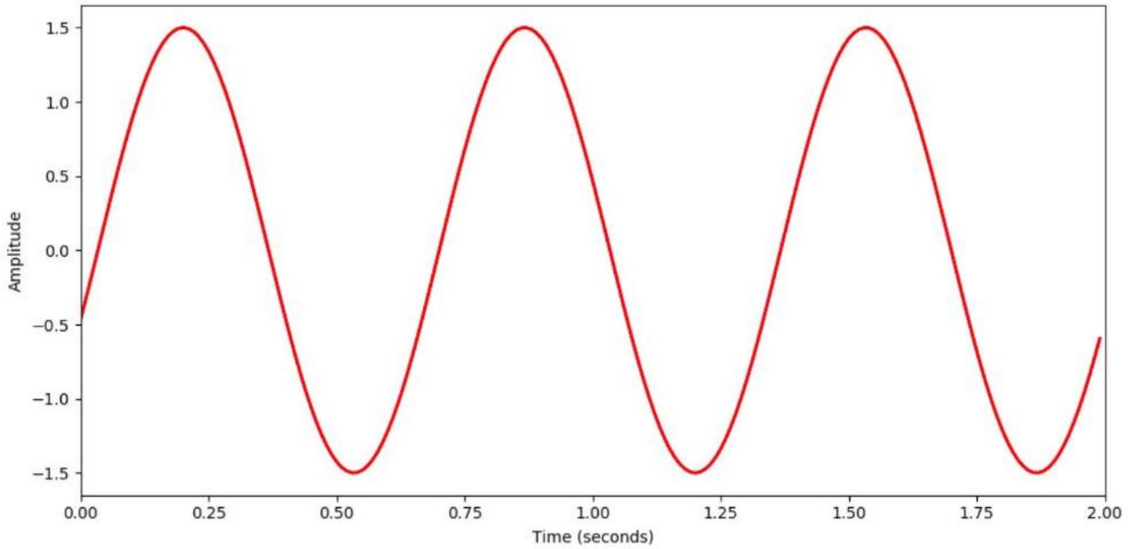
CONTENTS

01
기초 용어

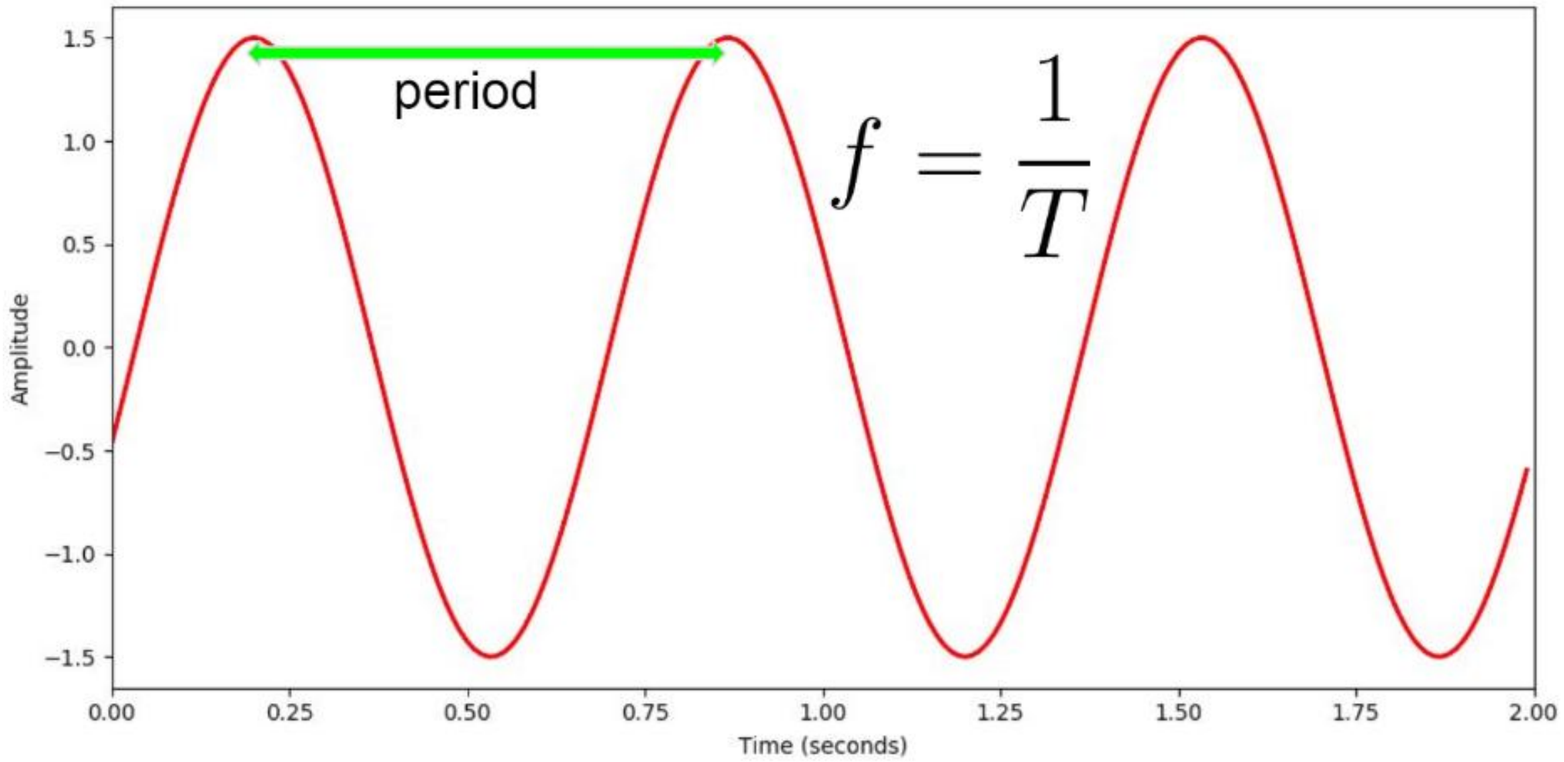
Frequency: 주파수 (높이)
Amplitude: 진폭 (세기)
Phase: 위상 (맵시)

Audio?

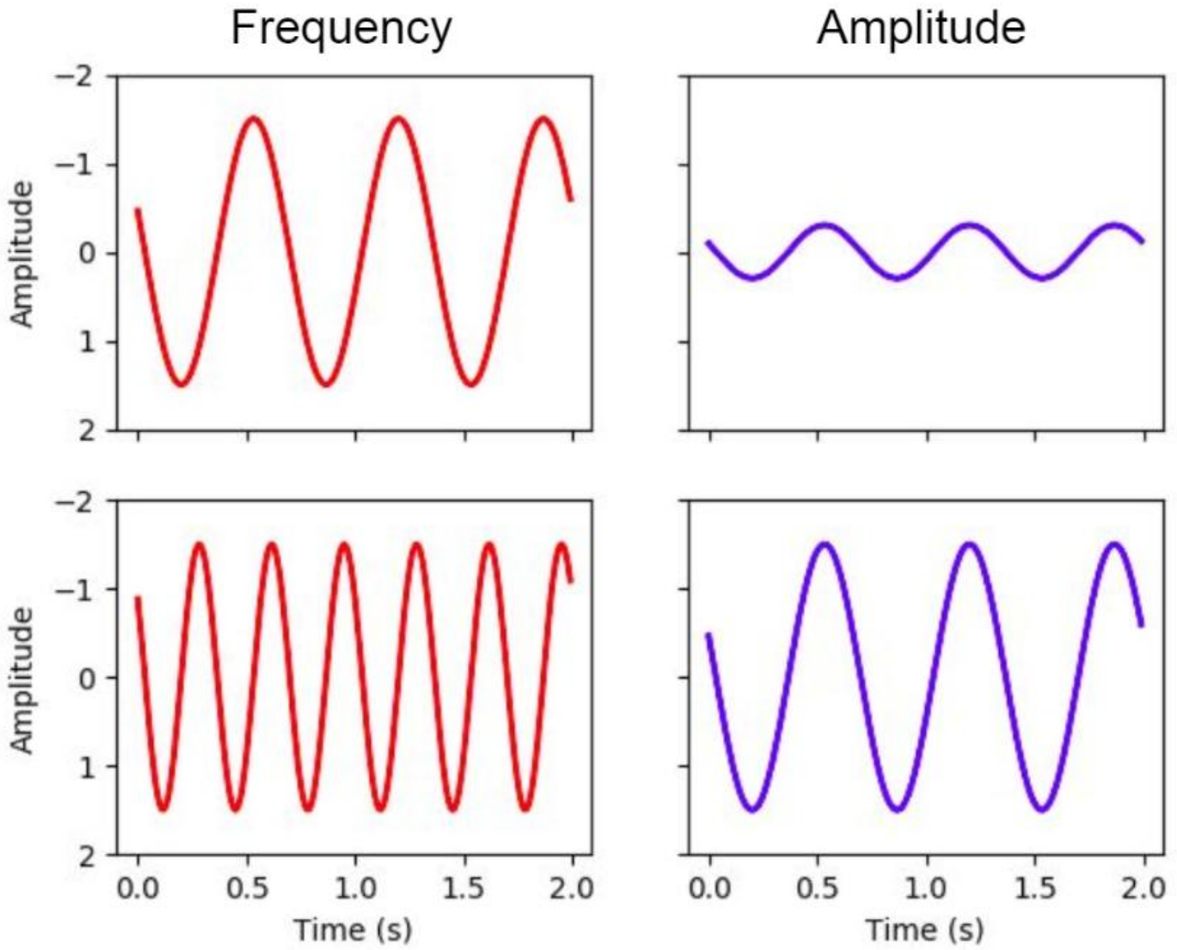
Audio 데이터를 다루기 위해선, audio가 무엇인지 어떻게 표현되는지 알아야한다.
기본적으로, audio는 어떤 물체가 진동하면서 발생한다. 예를 들어 목소리의 경우 공기 분자가 진동을 하면서 발생한다. 즉 매질인 공기 분자가 얼마나 크게 흔들렸는지에 따라 형성되는 이러한 공기압의 진폭이, waveform 형태를 띄게 되어 우리가 흔히 보는 그래프가 그려진다.



Y축은 Amplitude(진폭), X축은 Time(sec) 이다.



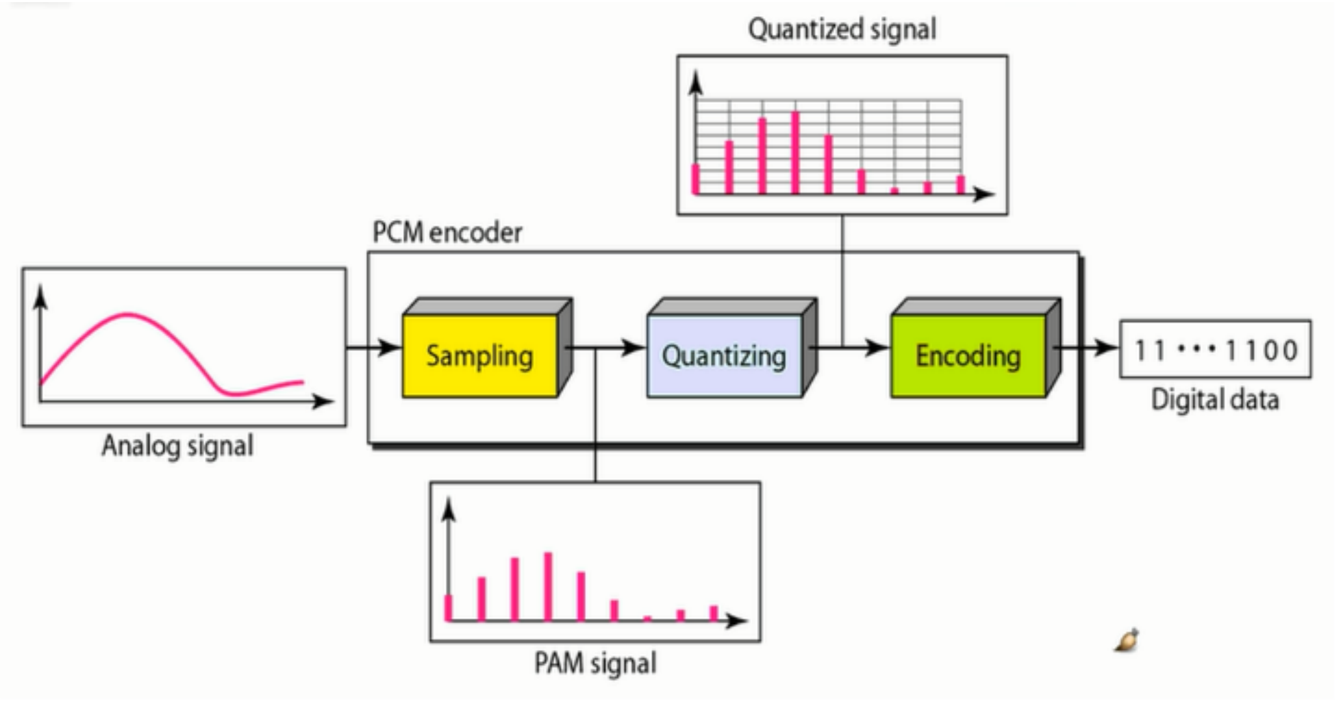
특정 지점에서 다음 등장하는 그 값까지를 period(주기)라고 정의한다. 이는 frequency(주파수) 개념으로 확장된다. Frequency는 Hz 단위를 사용하며, 1초에 100번 period(주기)가 발생하는, 즉 100번 진동하는 소리를 100Hz로 정의한다. 그래서 f = 1/T 라는 수식이 성립한다. 사람의 가청 Frequency는 약 20Hz ~ 20KHz.



High frequency, 즉 초당 진동 수(주기)가 많은 소리의 경우 High pitch(높은 소리)를 가진다. Amplitude는 소리의 크기(Loudness)와 관련되어 있다. 선형적이진 않지만, 대체로 Amplitude가 크면 소리가 크다.

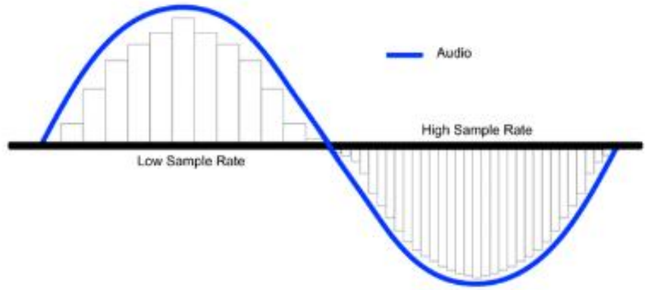
02
컴퓨터가 소리를 인식하는 방식

연속적인 아날로그 신호를 표본화(Sampling), 양자화(Quantizing), 부호화(Encoding)을 거쳐 이진 디지털 신호(Binary Digital Signal)로 변화시켜 인식하게 됨



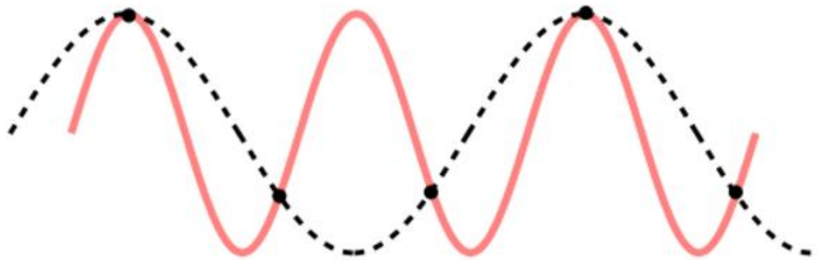
표본화(Sampling)

샘플링 단계에서 초당 샘플링 횟수를 정하는데, 이를 Sampling rate라고 함
“1초에 연속적인 시그널을 몇 개의 숫자로 표한할 것인가?”



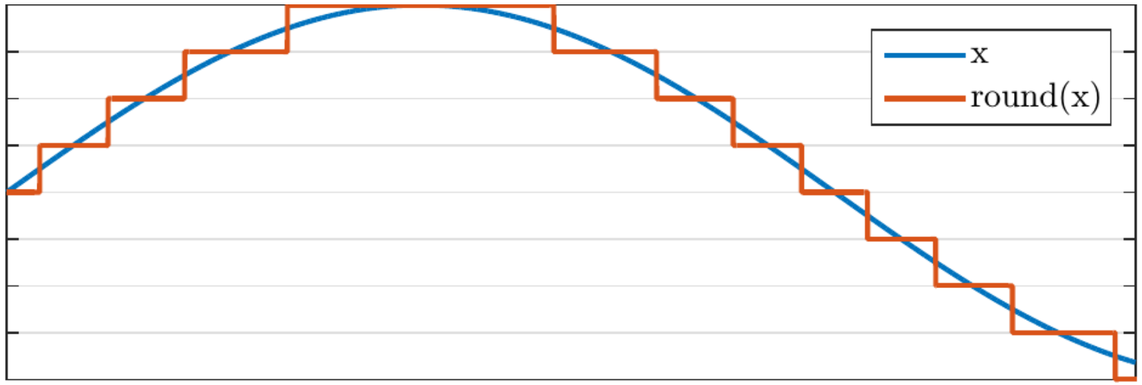
sampling rate가 클수록 즉, 자주 sampling할 수록 원본 데이터와 비슷할 것
→ 그러나 그만큼 저장해야 하는 데이터의 양이 늘어나게 됨

sampling rate가 작게 되면 아래와 같이 aliasing이 일어나 원본 데이터로 복원하는 데 어려움을 겪음



양자화(Quantizing)

양자화 단계에서는 amplitude의 real value를 기준으로 시그널의 값을 조절함
Amplitude를 이산적인 구간으로 나누고, signal 데이터의 Amplitude를 반올림하여 저장
(보통 bit로 나타냄)



- B bit의 Quantization : $-2^{B-1} \sim 2^{B-1} - 1$
- Audio CD의 Quantization (16 bits) : $-2^{15} \sim 2^{15} - 1$

위 값들은 보통 -1.0 ~ 1.0 영역으로 scaling 하기도 함

03
인간이 소리를 인지하는 방식

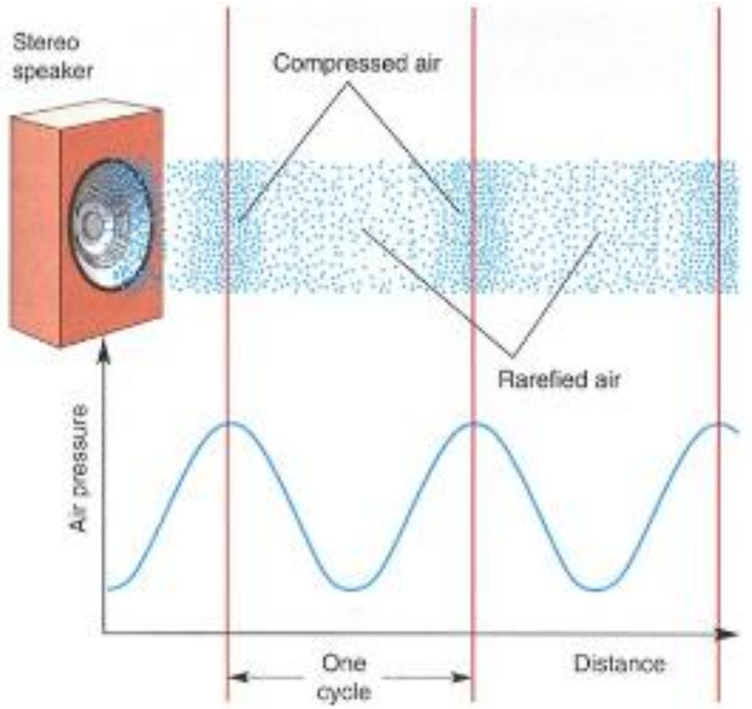
소리의 물리량

소리는 일반적으로 진동으로 인한 공기의 압축으로 생성됨.

그 압축이 얼마나 되었느냐에 따라서 표현되는 것이 바로 Wave(파동)

파동은 진동하며 공간/매질을 전파해 나가는 현상으로,
질량의 이동은 없지만 에너지/운동량의 운반은 존재.

파동에서 얻을 수 있는 물리량은 크게 아래 세 가지가 있음



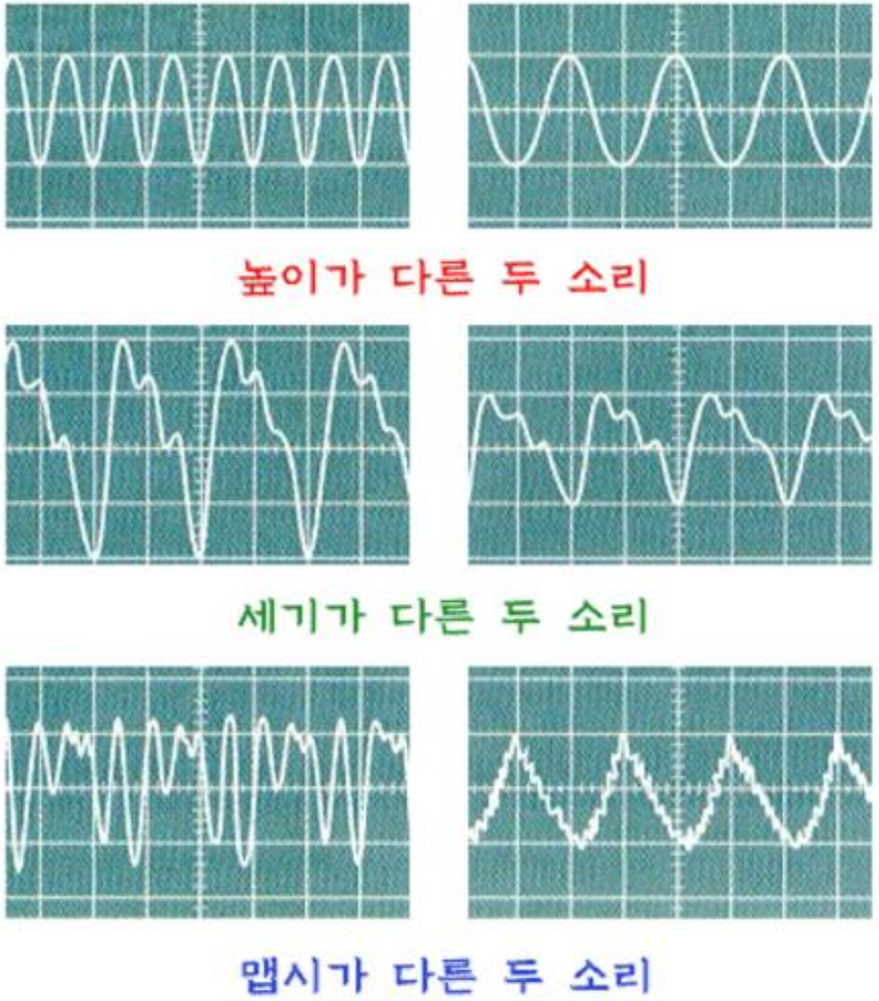
Frequency: 주파수 (높이), Amplitude: 진폭 (세기), Phase: 위상 (맵시)

물리 음향

- Intensity : 소리 진폭의 세기
- Frequency : 소리 떨림의 빠르기
- Tone-Color : 소리 파동의 모양

심리 음향

- Loudness : 소리 크기
- Pitch : 음정, 소리의 높낮이 / 진동수
- Timbre : 음색, 소리 감각



Frequency

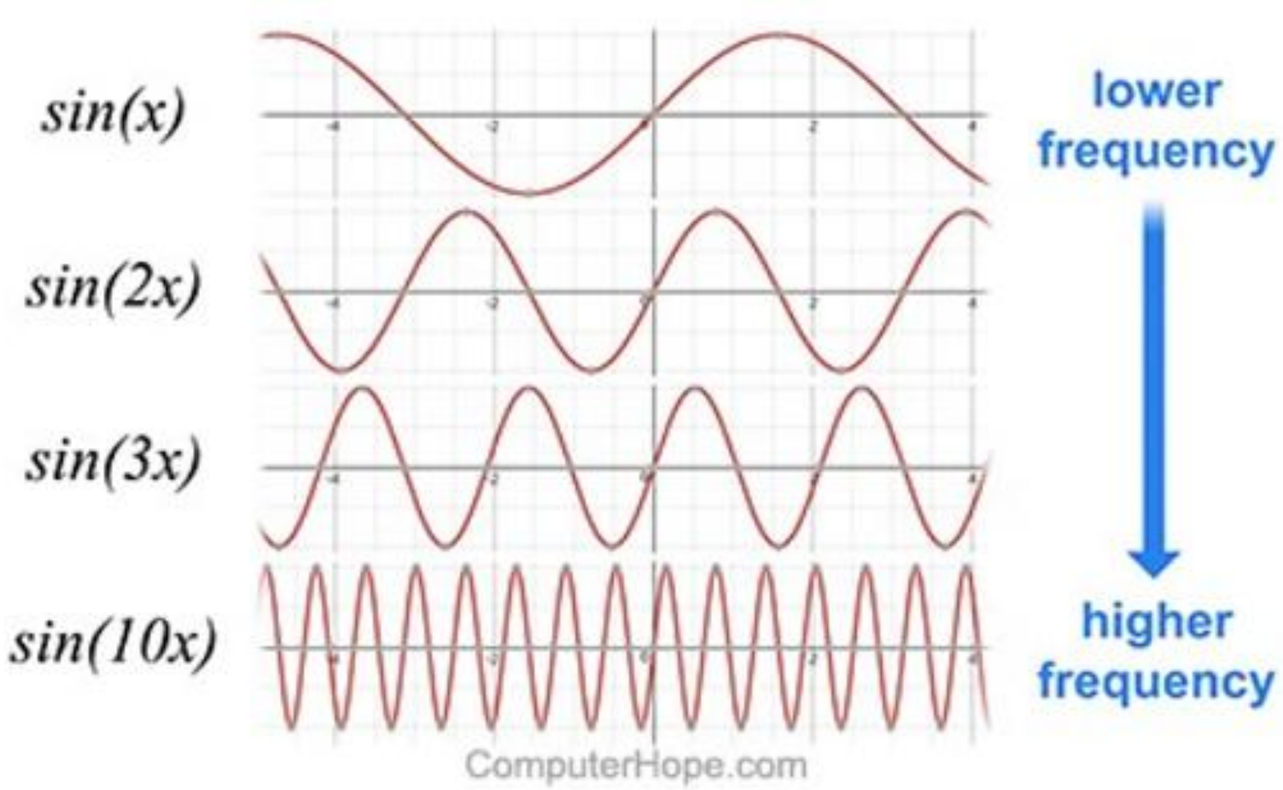
Frequency는 The number of compressed, 신호가 1초에 몇 번 진동했는지를 나타내는 수치.

소리는 빠르게 진동할수록, 즉 주파수가 높을수록 음이 높게 들림

단위는 Hertz를 사용하며, 1Hertz는 1초에 한번 Vibration을 의미
※ 정상인의 가청 범위) 20Hz - 20kHz (20,000Hz)
※ 가장 민감하게 느끼는 주파수) 1kHz(1,000Hz) - 4kHz(4,000Hz)

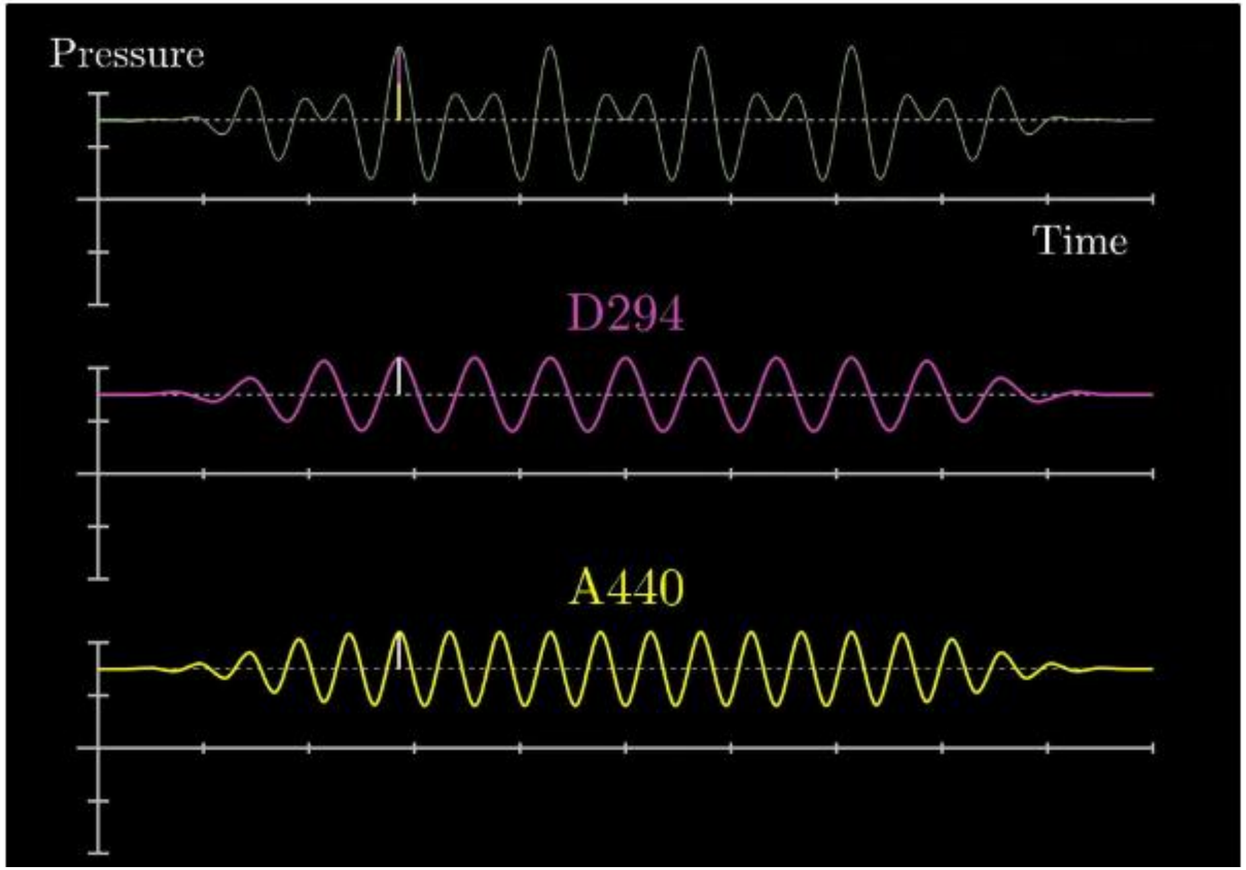
- 주기(period) : 파동이 한 번 진동하는데 걸리는 시간, 또는 그 길이. 일반적으로 sin함수의 주기는 $2pi/w$

- 주파수(frequency) : 1초 동안의 진동 횟수



Complex Wave(복합파)

우리가 사용하는 대부분의 소리들은 복합파
복합파는 복수의 서로 다른 정현파들의 합으로 이루어진 파형 (Wave)



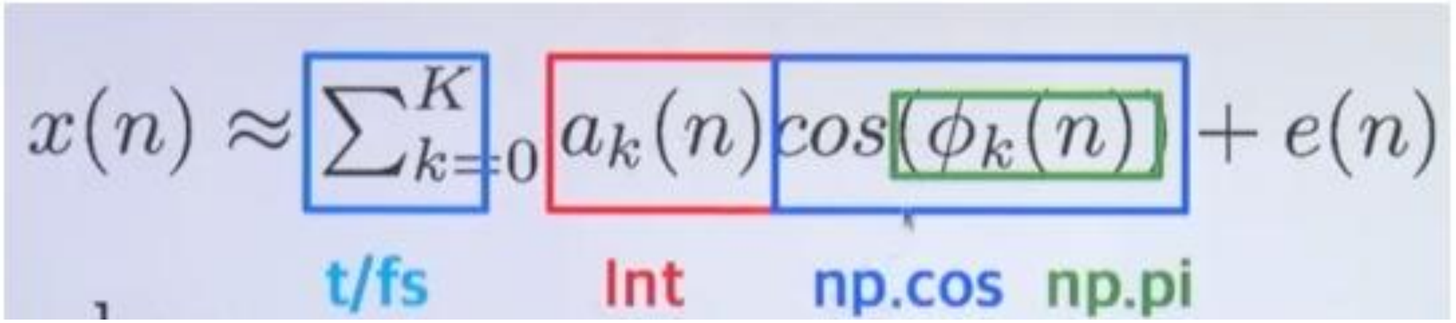
Sinusoid Wave (정현파)

정현파(Sinusoid)는 일종의 복소 주기함수 (복소수가 있는 주기 함수)

$$x(n) \approx \sum_{k=0}^{K} a_k(n) cos(\phi_k(n)) + e(n)$$

※ $a_k : amplitude, \phi_k : phase, e(n) : residual(noise)$

※ 복소수가 Phase 영역, 주기가 Frequency 영역이 됨, 크기가 Amplitude

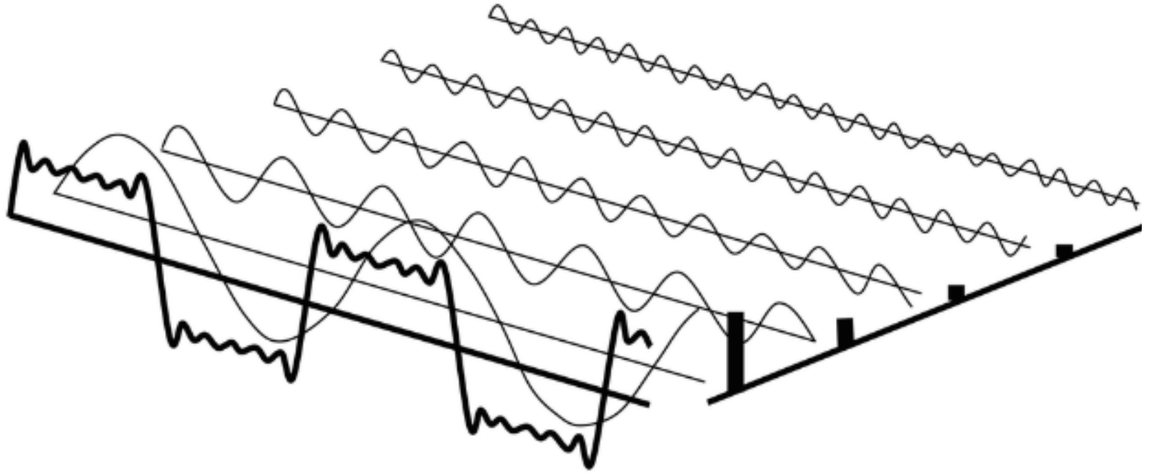


04
Fourier Transform

임의의 입력 신호를 다양한 주파수를 갖는 주기함수(복수 지수함수)의 합으로 분해하여 표현하는 것
그리고 각 주기함수들의 진폭을 구하는 과정을 퓨리에 변환이라고 함

Fourier Series

1807년 푸리에는 임의의 함수를 삼각함수의 (무한) 선형조합으로 표현할 수 있다고 주장



퓨리에 변환이 끝나면 실수부와 허수부로 리턴이 됩니다.
보통 실수부가 프리퀀시 영역대의 magnitude (주파수의 강도)를 뜻하고 , 허수부가 phase (주파수의 위상)를 뜻합니다.

이미지 상의 왼쪽에 위치한 신호는 입력신호(speech)로 보고, 이를 오른쪽과 같이 분해해서 보면 입력된 신호가 어떠한 frequency 영역대에서 활발하게 반응하는지 신호들의 세기를 알 수 있습니다.

즉, 푸리에 변환을 하면 신호에 어떤 주파수들이 포함되어 있는지와 그 강도를 나타내는 주파수 스펙트럼(Frequency Spectrum)을 얻을 수 있으며, 이는 단일 시점 또는 단일 시간 프레임에 대한 정보를 제공합니다.

그리고 이를 시간 축에 걸쳐 나열하여 시각화한 것이 Spectrogram이 보통 딥러닝에서 음성 신호를 변환하여 input으로 사용하는 데이터입니다.

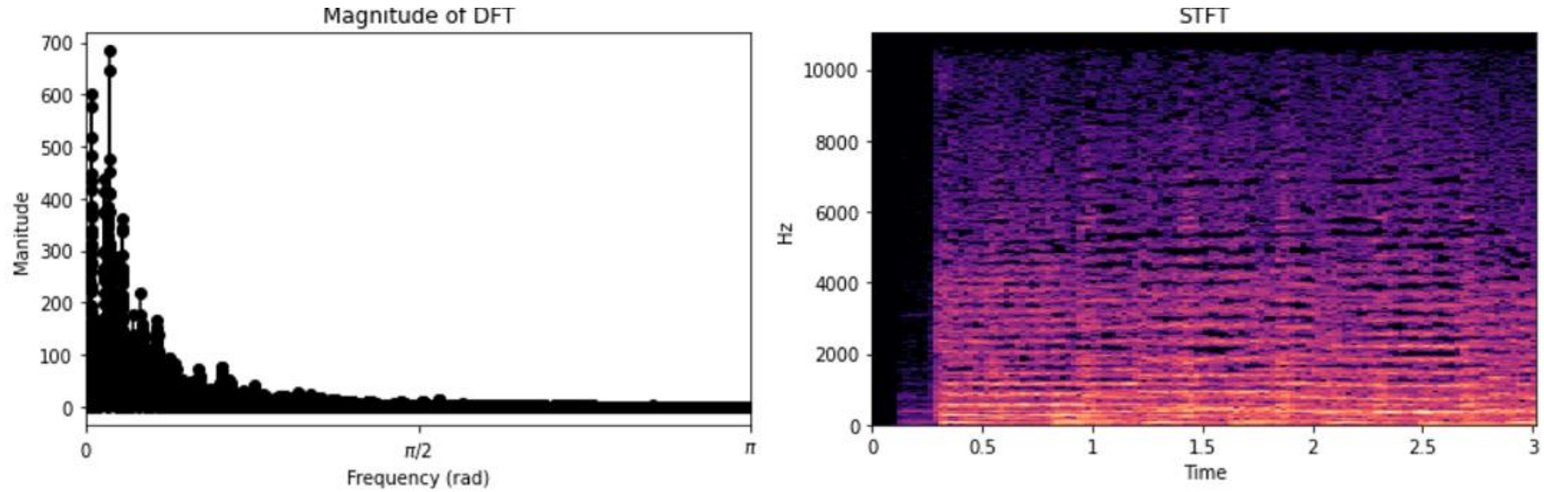
어떤 신호가 주기를 가지고 반복되고 있음
주기성에 대한 정보는 Real Axis(실수부)와 Imaginary Axis(허수부)로 나누어짐
푸리에 변환이 끝나면, 실수부와 허수부를 가지는 복소수가 얻어짐

$$z = a + bi$$

이러한 복소수의 절대값(실수부)은 spectrum magnitude (주파수의 강도)라고 부르며,
복소수가 가지는 phase를 phase spectrum (주파수의 위상)이라고 부름

과거 waveform 같은 경우는 amplitude 밖에 볼 수 없었는데,
푸리에 변환을 통해 어떤 frequency 영역대가 강한지 시각화시켜주는 Spectrogram을 얻을 수 있게 됐습니다.

05
Spectrum vs. Spectogram



왼쪽 그림이 spectrum이고, 오른쪽 그림이 spectrogram, spectrum과 spectrogram의 가장 큰 차이는 시간 축의 유무

spectrum은 시간 축이 없기 때문에 특정 시간대의 snapshot을 찍어 소리의 에너지를 분석한 것입니다.

반면에 spectrogram은 시간에 따른 소리의 변화를 시각화한 것입니다.

스펙트럼(Spectrum) : [특정 시간 길이]의 음성 조각(프레임)이 각각의 주파수 성분들을 얼마만큼 갖고 있는지를 의미 (푸리에 변환을 사용하면 스펙트럼을 얻음)

스펙트로그램(Spectrogram) : 시간 변화에 따른 스펙트럼의 변화.
여러 개의 스펙트럼을 시간 축에 나열하면 시간 변화에 따른 스펙트로그램을 얻음

대부분의 신호는 시간에 따라 주파수가 변하게 되므로 유용한 방법입니다. 그러나, FFT(Fast Fourier Transform)는 신호의 전체 시간 범위에 대해 한 번에 수행되며, 시간적 변화에 따른 주파수의 변화를 포착하지 못합니다. 즉, FFT를 하면 Time domain에 대한 정보가 사라진다는 것입니다.

이러한 한계를 극복하기 위해 시간을 프레임별로 나눠서 FFT를 수행하는 STFT(Short-Time Fourier Transform)가 등장합니다.

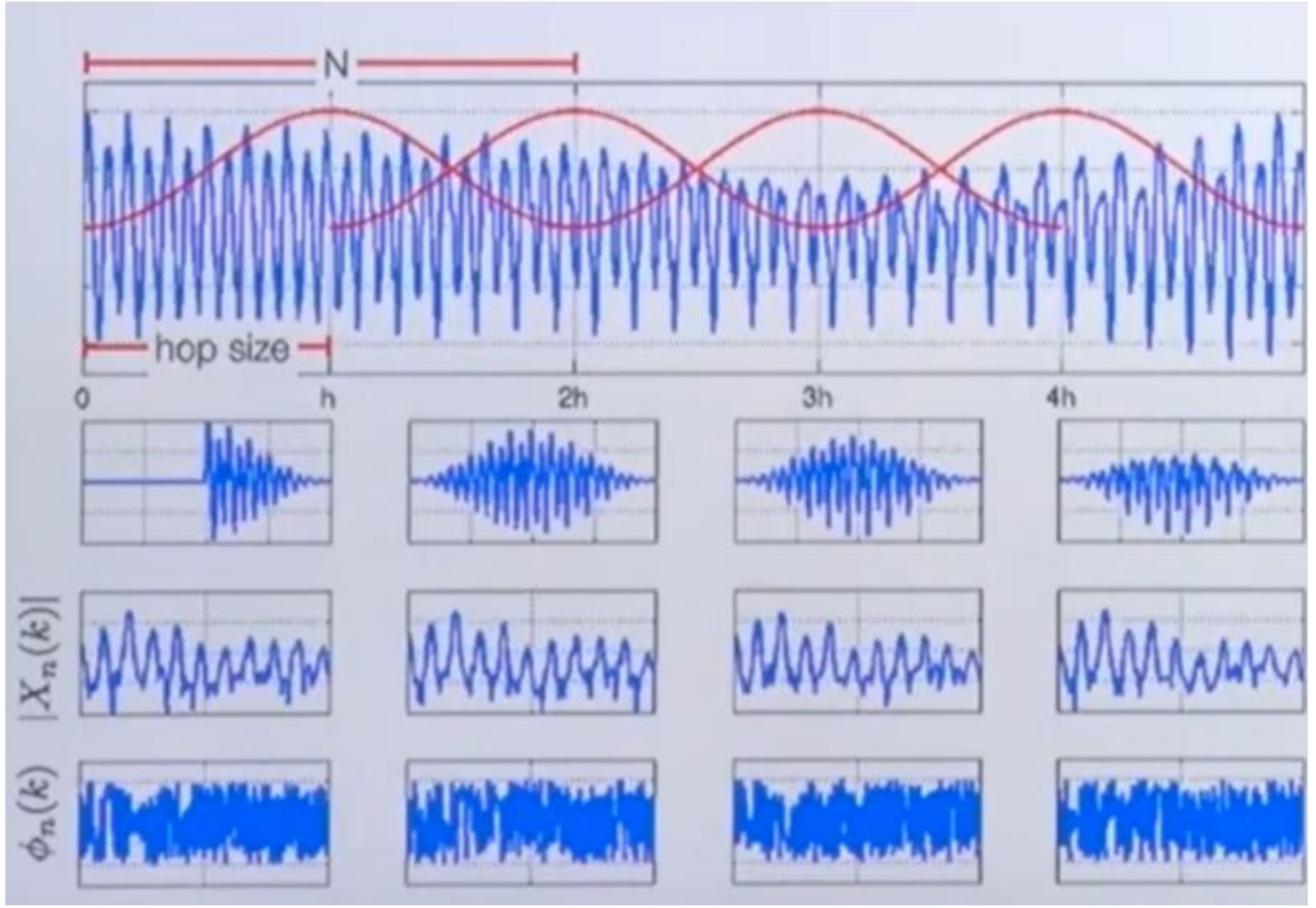
06

STFT (Short-Time Fourier Transform)

STFT는 신호를 시간의 짧은 구간으로 나누어 각 구간에 대해 푸리에 변환을 수행하는 기법입니다.

STFT는 각 시간 구간에 대해 푸리에 변환을 적용함으로써, 그 구간 내에서의 주파수 스펙트럼을 계산합니다.

STFT의 결과는 시간에 따른 주파수의 변화를 보여주는 스펙토그램을 생성하는 데 사용됩니다. 이를 통해 각각의 시간 구간에 대해 주파수 분석을 수행하기 때문에, 시간에 따라 어떻게 주파수 성분이 변하는지를 시각적으로 파악할 수 있게 해줍니다.



즉, STFT를 수행하면, (time, frequncy, magnitude) 꼴로 spectrogram이 나오게 됩니다.

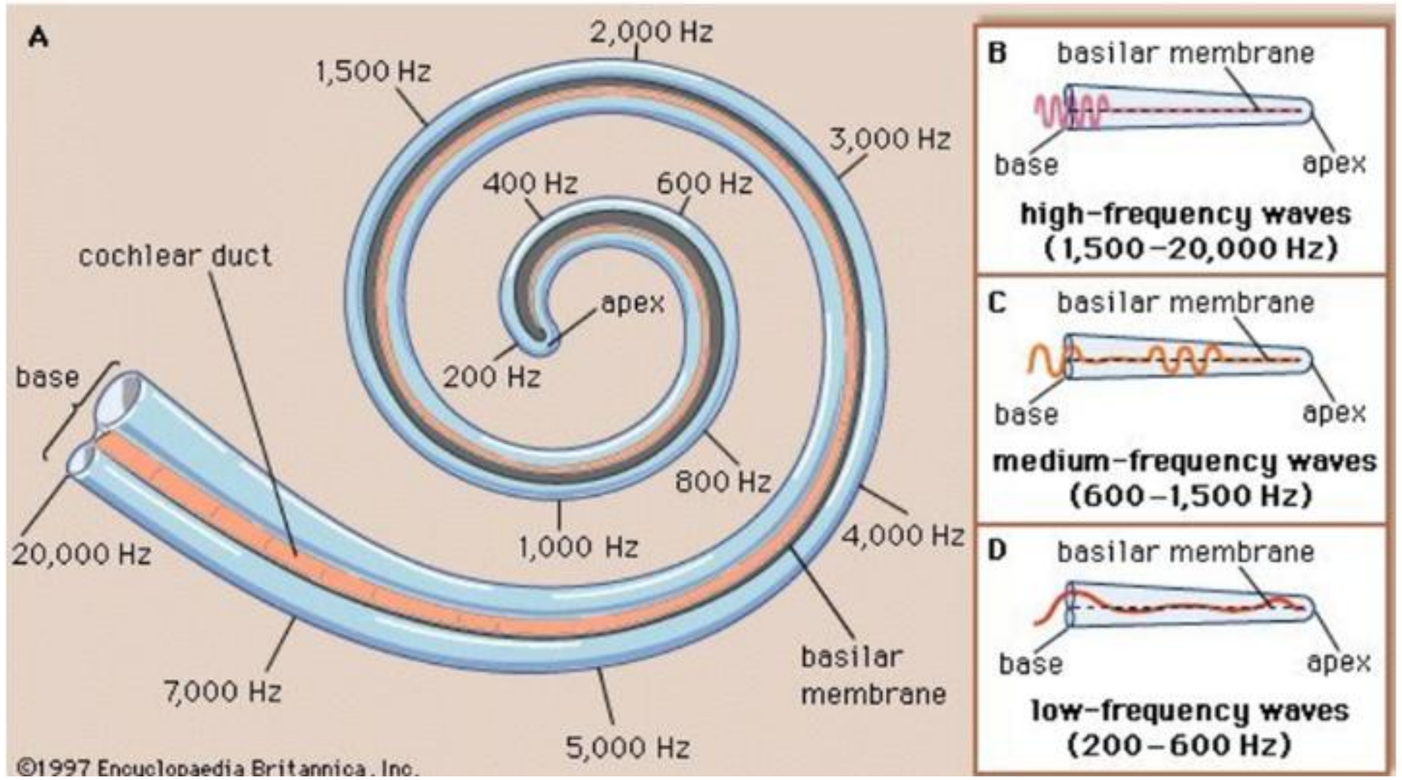
※ 시각화할 때는 일반적으로 time을 x축으로, frequency를 y축으로, magnitude를 색깔로 표현

07
Spectrogram vs. Mel Spectrogram

스펙토그램과 멜 스펙토그램은 모두 음성 신호의 시간에 따른 주파수 분포를 시각화한 것이지만, 두 가지의 주요 차이점은 주파수 스케일링 방식에 있습니다.
멜 스펙토그램의 핵심 아이디어는 인간의 청각 특성을 반영한 것인데, 이는 인간이 낮은 주파수에서는 더 민감하고, 주파수가 높아질수록 그 민감도가 감소하기 때문입니다.

Spectrogram

- 선형 주파수 스케일: 스펙토그램은 신호의 단시간 푸리에 변환(Short-Time Fourier Transform, STFT) 결과를 사용하여 생성됩니다. 여기서 각각의 시간 프레임에 대해 계산된 주파수 스펙트럼은 선형적인 주파수 스케일을 가집니다. 즉, 스펙토그램에서 주파수 축은 모든 주파수 구간을 동일한 간격으로 나타냅니다.



- 멜 스케일 : 멜 스케일은 인간의 청각 특성을 수학적으로 모델링한 것으로, 낮은 주파수에서는 세밀한 주파수 변화를, 높은 주파수에서는 상대적으로 넓은 주파수 변화를 포착할 수 있도록 설계되었습니다. 이를 사용하여 주파수 축을 변환한 것이 멜 스펙토그램입니다. (이는 달팽이관의 가장 안쪽 청각 세포는 저주파 대역을 인지하며, 바깥쪽 청각 세포는 고주파 대역을 인지한다는 점을 모델링한 것입니다.)

결론적으로, 멜 스펙토그램은 선형 주파수 스케일의 스펙토그램에 비해 인간의 청각 특성을 더 잘 반영한 주파수 스케일을 사용합니다. 이러한 특성 때문에 멜 스펙토그램은 음성 신호를 처리하는 딥러닝 모델의 입력으로 널리 사용되며, 특히 인간의 언어와 음악에 대한 태스크에서 더 성능이 좋은 결과를 낼 수 있습니다.

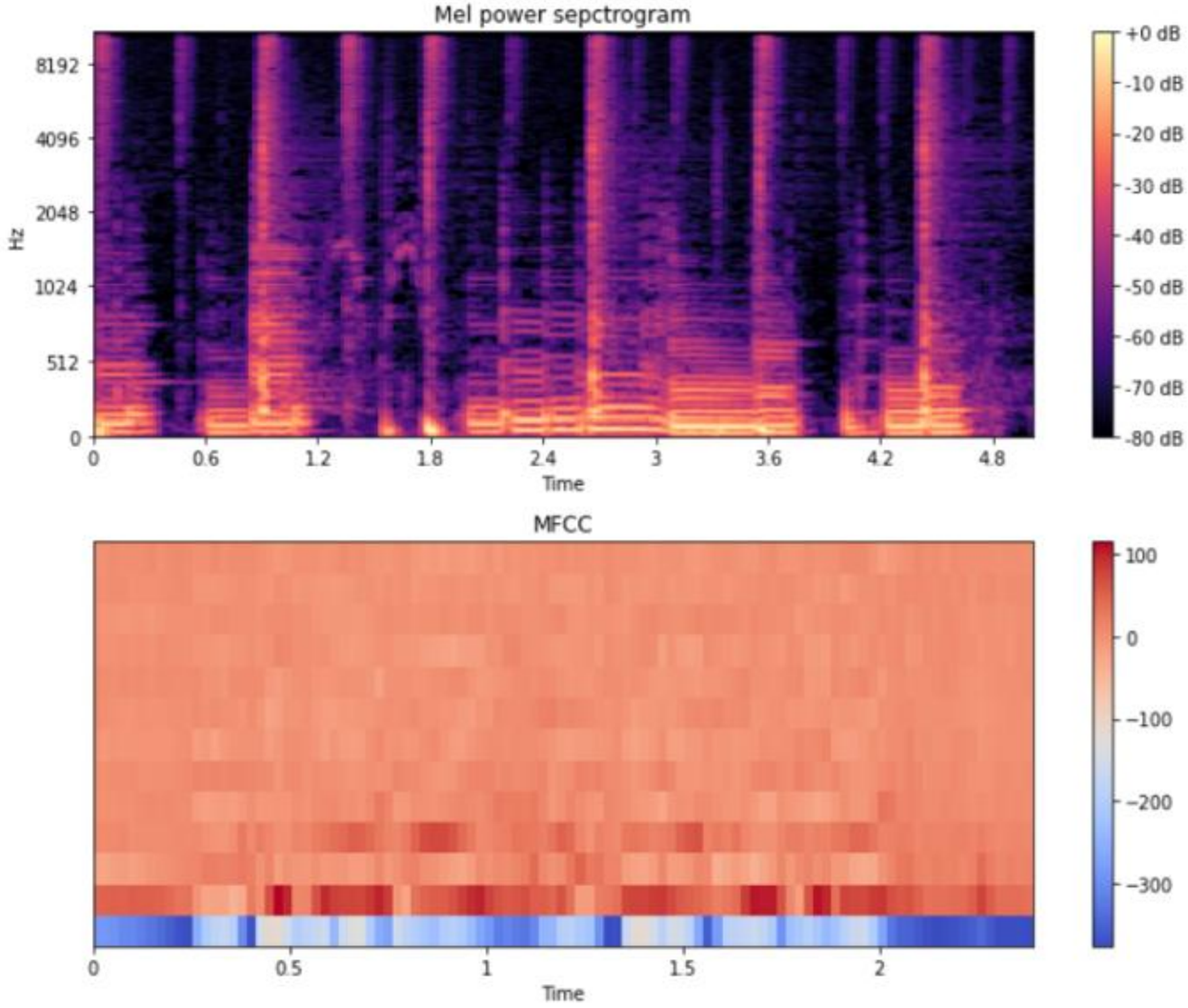
이외에도 MFCC (Mel-Frequency Cepstral Coefficients)가 존재하지만 구축 과정에서 버리는 정보가 지나치게 많아 최근 딥러닝 모델에서는 멜 스펙트럼 혹은 로그 멜 스펙트럼이 널리 쓰인다고 합니다.

하지만 Conv 1D에서는 1D 시퀀스 데이터를 입력으로 주로 사용하기 때문에 MFCC와 GTCC 사용을 고려할 수 있습니다.

이들은 음성의 멜 스케일 주파수 특성을 요약하여, 음성 신호에서 중요한 정보를 압축적으로 표현하는 1D 벡터를 생성합니다. 이러한 1D 특성은 시간에 따른 변화를 포착하는 1D 컨볼루션 레이어와 잘 맞기 때문에, 음성 인식에 주로 사용된다고 합니다.

(그러나, 멜스펙트럼 역시 Convolution 1D에서의 입력 데이터로 사용할 수 있습니다.)

08
MFCC


Mel power sepctrogram
8192
4096
2048
1024
512
0
Hz
0
0.6
1.2
1.8
2.4
3
3.6
4.2
4.8
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB
MFCC
0
0.5
1
1.5
2
Time
100
0
−100
−200
−300

MFCC는 음성의 멜 스케일 주파수 특성을 요약하여, 음성 신호에서 중요한 정보를 압축적으로 표현한 것입니다.

MFCC 그래프에서 가로축은 시간을 나타내고, 세로축은 MFCC 벡터의 각 성분(보통 "cepstral coefficients"라고 함)을 나타냅니다. 각 시간 프레임에 대해 계산된 이러한 계수들은 1D 벡터로, 시간에 따라 나열되어 시각화된 2D 이미지를 형성합니다.

하지만 MFCC 자체는 원래 1D 시퀀스 데이터입니다. 각 시간 단계에 대해 MFCC는 다수의 계수(벡터)를 생성하며, 이들 벡터는 음성 인식, 감정 분석, 화자 인식 등 다양한 음성 처리 작업에서 입력 특성으로 사용될 수 있습니다.

MFCC 이미지를 보면 각각의 시간 단계에 대해 여러 개의 계수가 나타나는데, 이것은 오디오 신호의 특정 시간 프레임에서 파생된 특성 벡터입니다. 그래서 MFCC는 기본적으로 1D 시퀀스 데이터로 간주되지만, 시간에 따라 나열되어 2D 이미지 형태로 시각화할 수 있습니다. 그러나 이 데이터를 1D 컨볼루션에 입력으로 사용할 때는, 각각의 시간 단계에서의 MFCC 벡터를 독립적인 1D 시퀀스로 간주하여 모델에 적용합니다.