음성인식과 음성합성 기술학습과정

2020. 10.

정 준 수 Ph.D.

과정 목표: 소리(Sound) 정보처리 과정 학습

- 음성인식기술, 응용 분야에 필요한 소리의 정보처리 및 모델에 필요한 기초 지식의 이해
- 그렇다면 인간은 이러한 소리를 어떻게 인지할까요?
- Computer가 소리를 이해하는 과정을 살펴 보는 시간

음성인식기술, 응용 분야 확대 추세

음성인식 기술이 최근 인공지능(AI) 개인비서나 스마트홈 가전제품, 자율주행차를 필두로 한 스마트카 등 다양한 산업 분야에 적용되는 추세

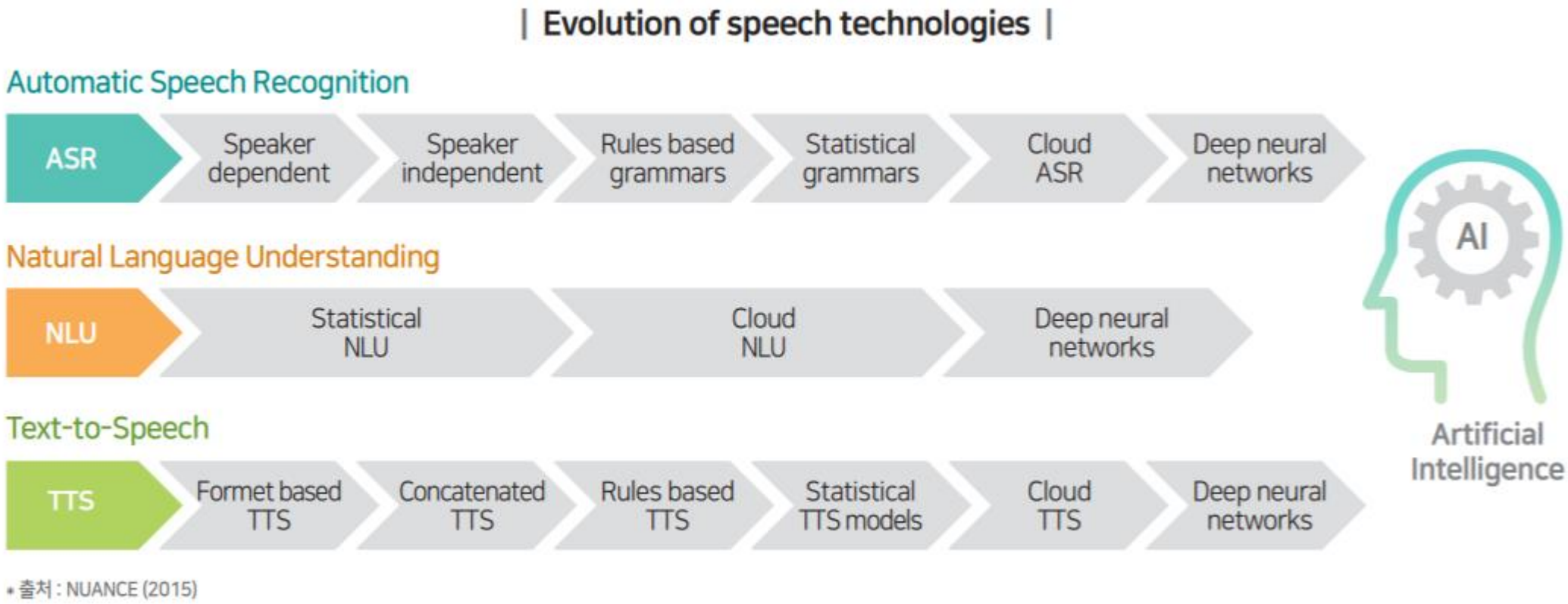
음성인식이란 소리 센서를 통해 얻은 음향학적 신호(acoustic speech signal)를 컴퓨터가 해석, 그 내용을 문자 데이터로 전환 처리하는 기술이다. 음성인식 기술은 화자의 고유 정보를 바탕으로 개인 식별이 가능하고 입력 속도가 빠르다는 장점을 갖고 있음

- Sound
  - Speech Classification & Auto-tagging (Acoustic Scene / Event Identification)
- Speech
  - Speech Recognition (STT)
  - Speech Synthesis (TTS)
  - Speech Style Transfer (STS)

[출처] https://www.kakaobrain.com/blog/16

관련 기술 동향

음성인식 기술은 1980년대 소개된 IBM에서 제안한 통계적인 방식에서 클라우드 방식으로 발전하고 있으며, 궁극적으로 심층신경망(Deep Neural Network, DNN)을 적용하는 방식으로 발전 예상



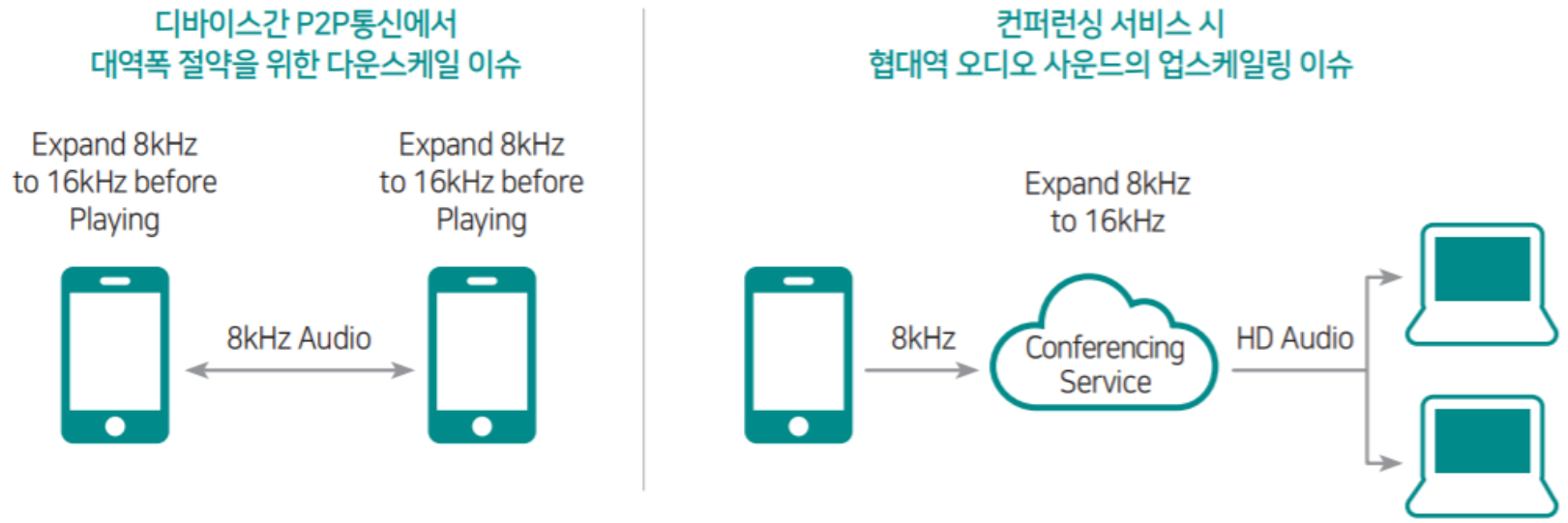
* 출처 : NUANCE (2015)

기술 동향: 샘플링 주파수

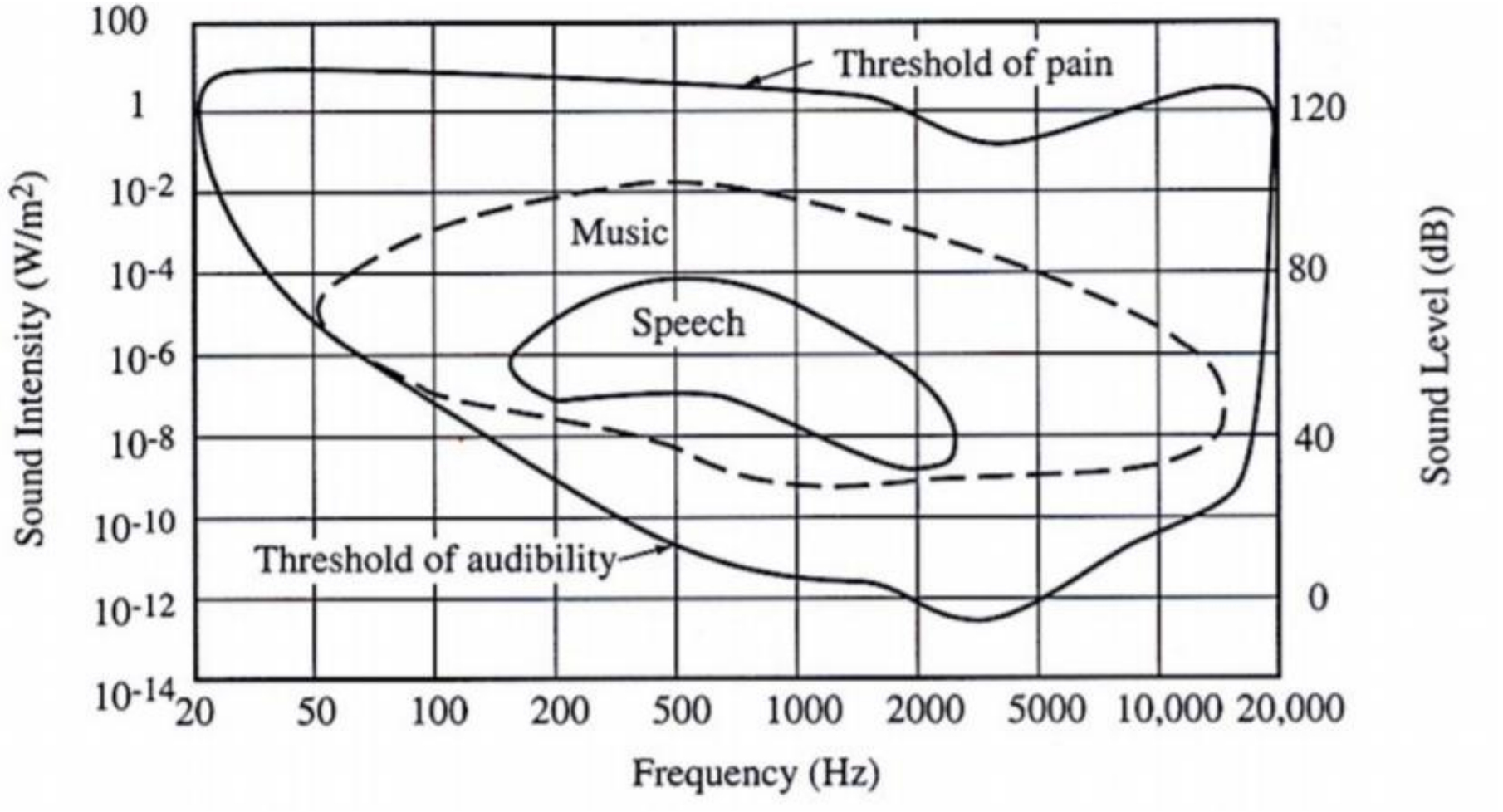
IoT 서비스 확대로 디바이스간 P2P 통신에서 대역폭을 절약하기 위한 다운스케일 이슈발생

최근 IBM등 선도기업에서는 음성-문자변환(STT)관련 서비스를 전화 회의 등으로 확장

콜센터 서비스를 제외한 다양한 STT 서비스 분야에서 협대역(8kHz) 및 광대역(16kHz) 샘플링 주파수 기반 음성인식을 동시에 활용하는 사례 증가-> 협대역 오디오 사운드의 광대역 업스케일링 이슈 발생



그렇다면 어떻게 Sampling Rate를 설정할 것인가?



음성학

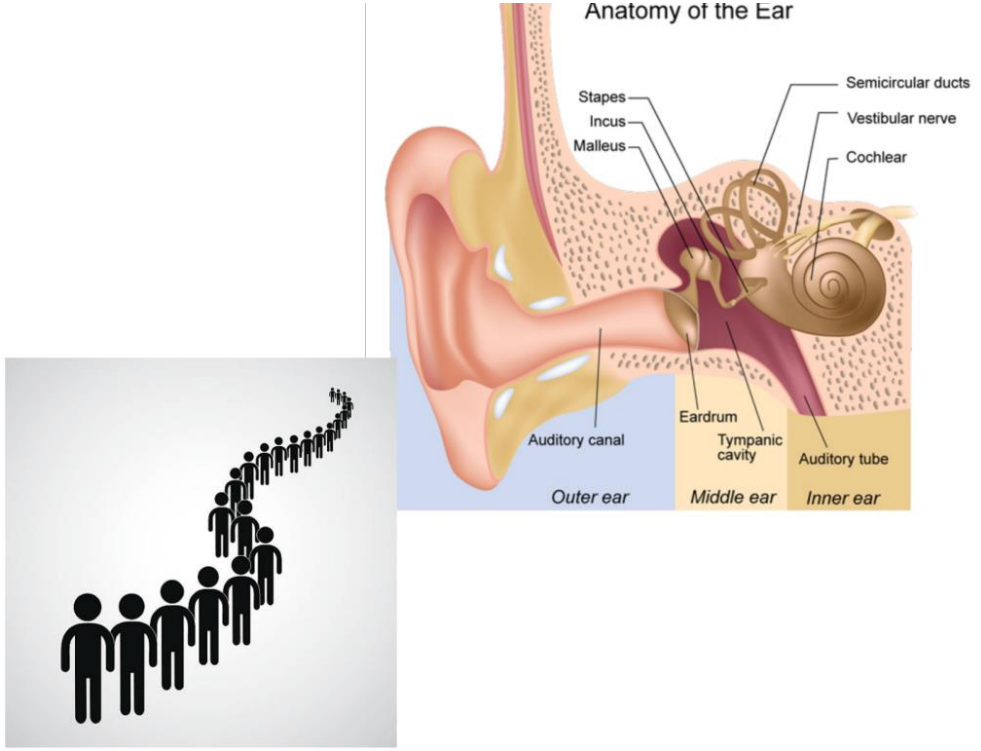
음성학은 말소리의 생성과 인지를 다루는 학문이다.
전통 음성학 에서는 말소리의 조음과 청취 인상에 바탕을 둔 연구가 주로 이루어 졌고,
현대 음성학에서는 말소리의 조음 및 청취적 특성뿐만 아니라,
말소리의 음향적 특성에 바탕을 둔 연구로 그 범위가 확대 되었다.

전통 음성학은 말소리의 조음과 청취에 대한 주관적인 연구가 주를 이루었으나
현대 음성학은 과학적 연구 방법에 바탕을 둔 객관적인 연구가 이루어지고 있다.

이런 차이에도 불구하고 전통 음성학과 현대 음성학은 배타적인 것이 아니라 상호 보완적인 관계에 있다.
전자는 말소리에 관한 인상과 영감을 주고 후자는 말소리에 관한 과학적 증거를 제공한다.

소리의 지각

소리와 음향 파형 : 압력의 변화가 고막에 영향을 미칠 때 생성된다. 음향 파형은 소리가 생성되는 압력 변화(pressure fluctuatuion)를 시간에 따라 기록한 것이다.



소리의 전파

소리 전파는 중간에 틈(압력의 변화)이 이동하여 마지막에 줄 서 있는 사람도 맨 앞에 서 있는 사람의 영향을 받는 것과 비슷함, 반동에 의한 전파라는 점이 줄 서기와 다름 인접한 사람들 사이의 공백은 음의 공기 압력 즉 희박(rarefaxtion)에 해당하고 충돌은 양의 공가 즉 압축(compression)에 해당한다.

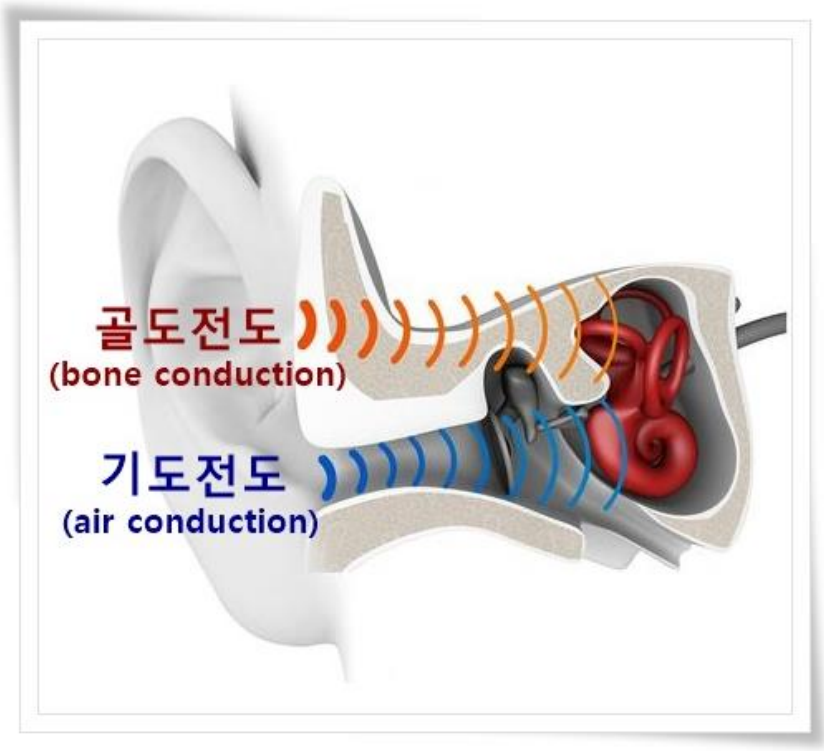
소리를 듣는 원리를 이해하려면 먼저 음향이란 공기를 통해 전달되는 보이지 않는 진동이라는 것을 알아야 합니다. 누군가 말을 하고, 나뭇잎이 바스락거리고, 전화벨이 울리거나 무언가 '음향'을 만들어 낼 때 진동이 공기를 통해 모든 방향으로 전달됩니다. 이것을 음파라 부릅니다

소리의 전달 경로

초기의 소리 전달 경로는 에너지 형태의 변환이 일어나기 전의 과정에서 **기도 전도(air conduction)**와 **골도 전도(bone conduction)**로 나누어집니다. 기도 전도는 소리를 듣는 일반적인 방법으로 소리가 공기를 통해 외이-중이-내이를 거쳐 전달되는 경로인 반면, 골도 전도는 소리가 고막을 거치지 않고 뼈를 통해 내이로 직접 전달되는 방법을 말합니다.

기도 전도 (air conduction): 외이 → 중이 → 내이 → 청신경 → 대뇌

골도 전도 (bone conduction): 뼈 → 내이 → 청신경 → 대뇌



기도 전도와 골도 전도

이때, 기도 및 골도에서 전달되는 주파수 성분은 지나는 경로에 따라 다를 수 있습니다.

골도 전도에는 세 가지 성분이 있습니다.

① 와우의 골 구조가 진동되어 생기는 성분 (**distortional component**)

② 이소골과 내이 액체의 질량에 골도의 진동이 영향을 미쳐 발생하는 관성 성분 (**inertial component**)

③ 골도의 진동이 외이도를 통해 고막으로 전달되어 생기는 반응 (**osseo-tympanic component**)

이들 성분은 임상적으로 난청의 원인에 따라 상호 작용을 하여 청력검사 결과에 영향을 미치기도 합니다.

디지털 신호처리

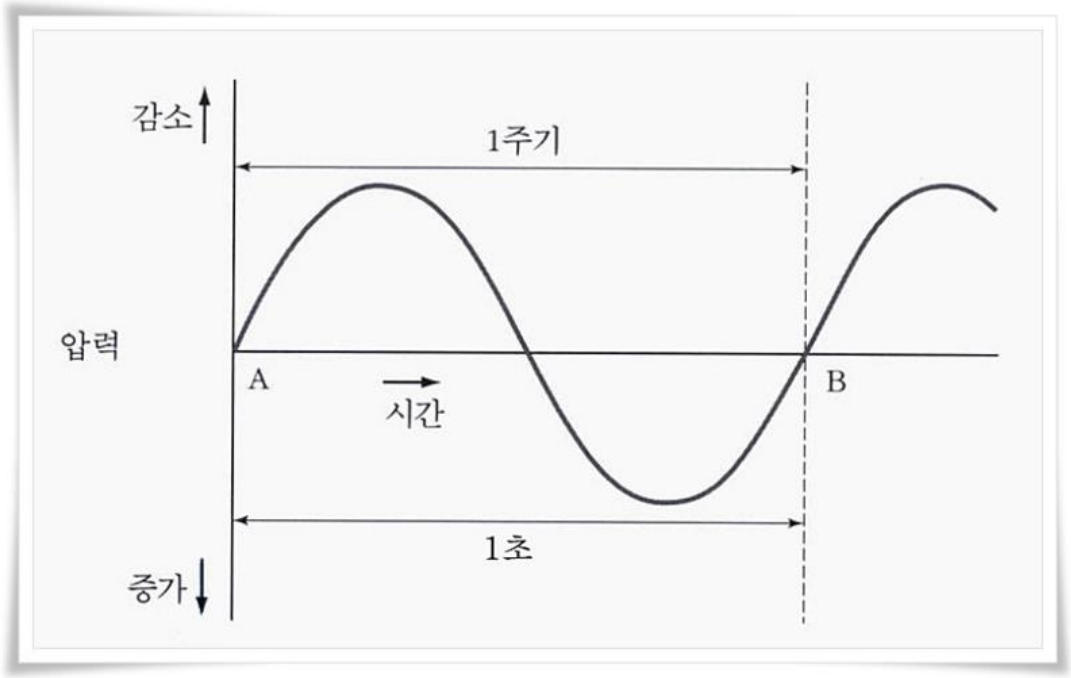
디지털 신호 처리(digital signal processing), 즉 컴퓨터로 음향 신호를 다루는 방법

- 단순음
  - 복합음
  - 주기적인 소리
- 비주기적인 소리

주파수 (Frequency)

주파수(frequency)는 공기 입자의 진동, 즉 압축과 희박을 겪는 속도를 말하는 것으로서, **하나의 연속적 압축과 희박을 1주기(cycle)**라고 합니다. **1초에 일어나는 주기의 수가 주파수**이며 이는 **Hertz(Hz)**로 나타냅니다.

한 예로 1초에 100여개의 연속적인 압축과 희박이 있으면 1초에 100 cycle, 즉 100Hz입니다.



주파수는 1초의 시간 동안 생기는 완전한 압축과 희박의 수

진폭(amplitude)압력 변화가 보통의 대기압으로부터 벗어난 정도를 나타낸다. 소리의 압력을 나타내는 파형에서 진폭은 수직 축에 나타난 다.

위상(phase) : 어떤 기준 시점에 대하여 파형이 가지는 상대적 시간. 한 원 안에 들어가는 직각 삼각형으로부터 진폭을 취하여 사인파를 그리면 원 둘레 한 바퀴는 종이 위에서 하나의 사인파와 같다.

강도(intensity)와 **진폭(amplitude)**은 정해진 시간동안 소리에 전해지는 에너지를 나타내는 데 쓰이는 용어입니다. 더 자세하게는 힘, 음압 혹은 에너지로 표현될 수 있습니다.

데시벨(dB)은 소리의 에너지를 나타내는 단위로 사용됩니다. 소리의 크기를 비교할 때 'A는 B보다 100배 크다'라고 표현합니다. 음의 크기를 측정할 때도 이러한 '~배'를 쓸 수 있지만, 우리가 듣는 소리 크기의 범위는 상당히 넓기 때문에 그 범위를 압축한 단위를 쓰게 됩니다. 즉, dB은 근본적으로 '배'와 같은 개념입니다. A가 B보다 몇 배 더 크다고 하면 그 기준을 알아야 하는 것처럼, dB에서도 그 기준이 무엇인지 알아야 합니다. 즉, dB은 두 소리 간 또는 기준에 대한 두 소리의 강도 혹은 음압의 비율을 나타내는 것입니다.

단순 주기파(사인파)

추운동과 같은 단순 조화 운동의 결과로 생김

- 어린 아이의 성대 진동은 사인파에 가깝다. 여자들의 성대 진동은 남자들의 성대진동보다 사인파에 가깝다.

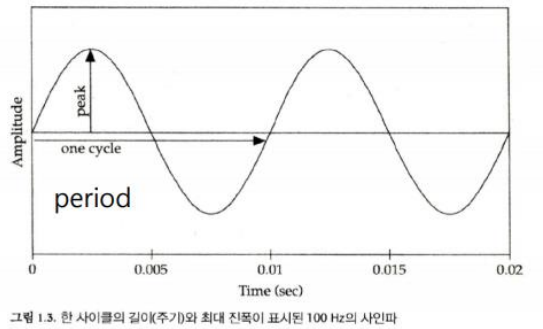


그림 1.3. 한 사이클의 길이(주기)와 최대 진폭이 표시된 100 Hz의 사인파

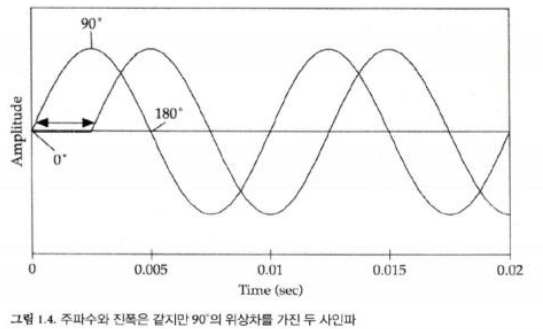


그림 1.4. 주파수와 진폭은 같지만 90°의 위상차를 가진 두 사인파

주파수(frequency): 수평축에서 단위 시간 당 사인파 패턴이 반복되는 횟수

사이클(cycle) : 반복되는 패턴

주기(period) : 한 사이클이 완성되는데 걸리는 시간

주파수는 초당 사이클 수를 나타내면 헤르츠(Hz)로 나타낸다.

복합 주기파

전체 파형의 한 사이클 안에 추가된 10개의 잔물결(작은 마루)를 셀 수 있다. 복합파의 패턴이 반복되는 빈도를 기본 주파수(fundamental frequency, F0)라고 한다.

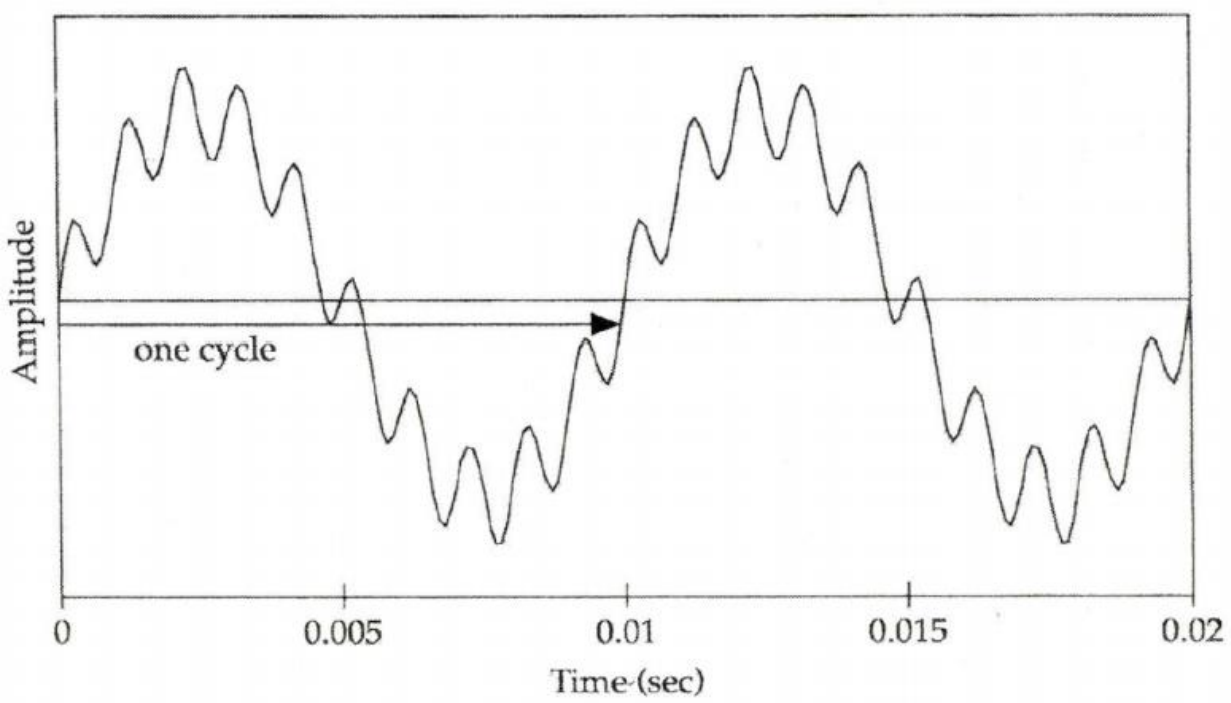
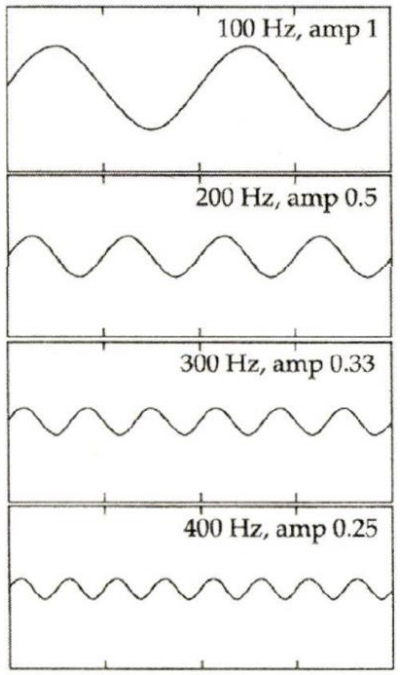
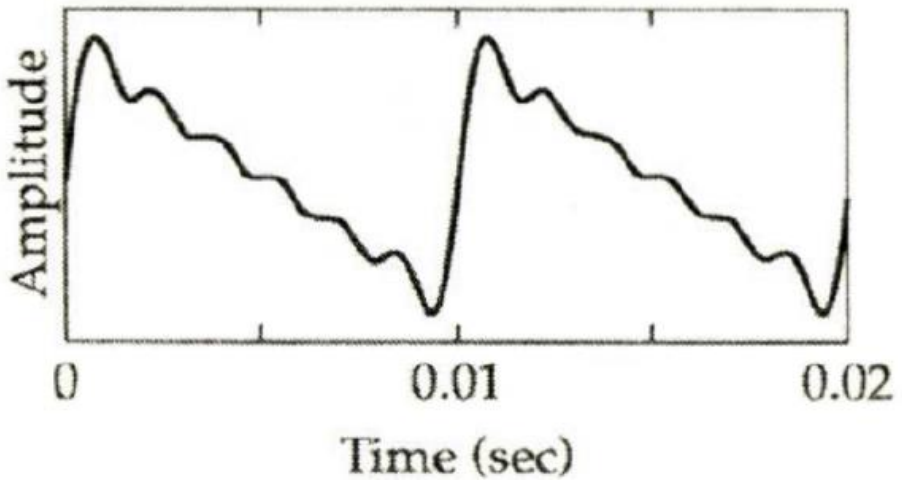


그림 1.5. 100 Hz의 사인파와 1,000 Hz의 사인파를 합성한 복합 주기파. 기본 주파수(F_0)의 한 사이클이 표시되어 있다.

보통 많은 주파수 구성 요소로 이루어져 있기 때 문에 표로 제시하는 것은 비현실적이다. 복합파 의 구성 요소인 단순 사인파를 진폭 대 주파수로 보여주는 그림을 파워 스펙트럼이라 한다.

주파수 (Hz)	100	200	300	400	500
진폭	1	0.5	0.33	0.25	0.2

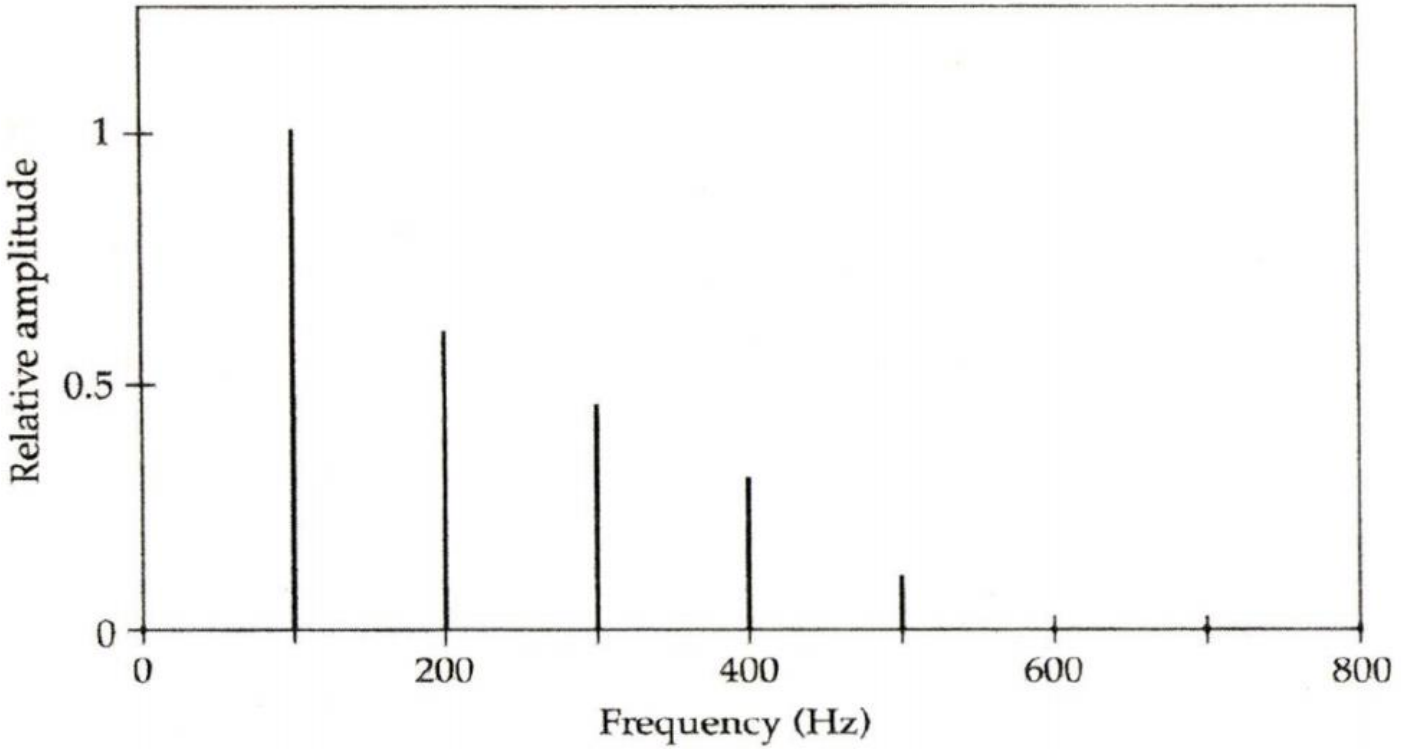




"톱니" 파 모양과 유사한 복합 주기파, 그리고 복합파를 만들기 위해 합성된 사인파들중 주파수가 가장 낮은 사인파 4개

파위 스펙트럼

어떤 복합파라도 주파수 , 진폭 , 위상이 각각 다른 일단의 사인파로 해체할 수 있다. 음파의 이러한 속성은 이 사실을 발견한 17 세기의 수학자 푸리에 (Fourier)의 이름을 따 퓨리에 변환 (Fourier transform)이라 한다.



이전 17 Page에 나타난 복합파의 구성 용인인 단순 주기파의 주파수 진폭을 나타낸 그래프

특정한 **순음은 주파수(frequency), 진폭(amplitude), 위상(phase) 등으로 결정**됩니다. **주파수는 소리의 고저**를, **진폭은 소리의 크기**를 결정 짓습니다. 위상이 순음의 인식에 미치는 영향은 거의 없으나 여러 순음이 합쳐진 복합음(complex sound)의 경우에 각 성분의 위상이 음 인식에 미치는 영향은 다양하다고 알려져 있습니다.

$i^0 = 1$

$i^1 = i$

$i^2 = -1$

$i^3 = (i^2)(i^1) = (-1)(i) = -i$

$i^4 = (i^2)(i^2) = (-1)(-1) = 1$

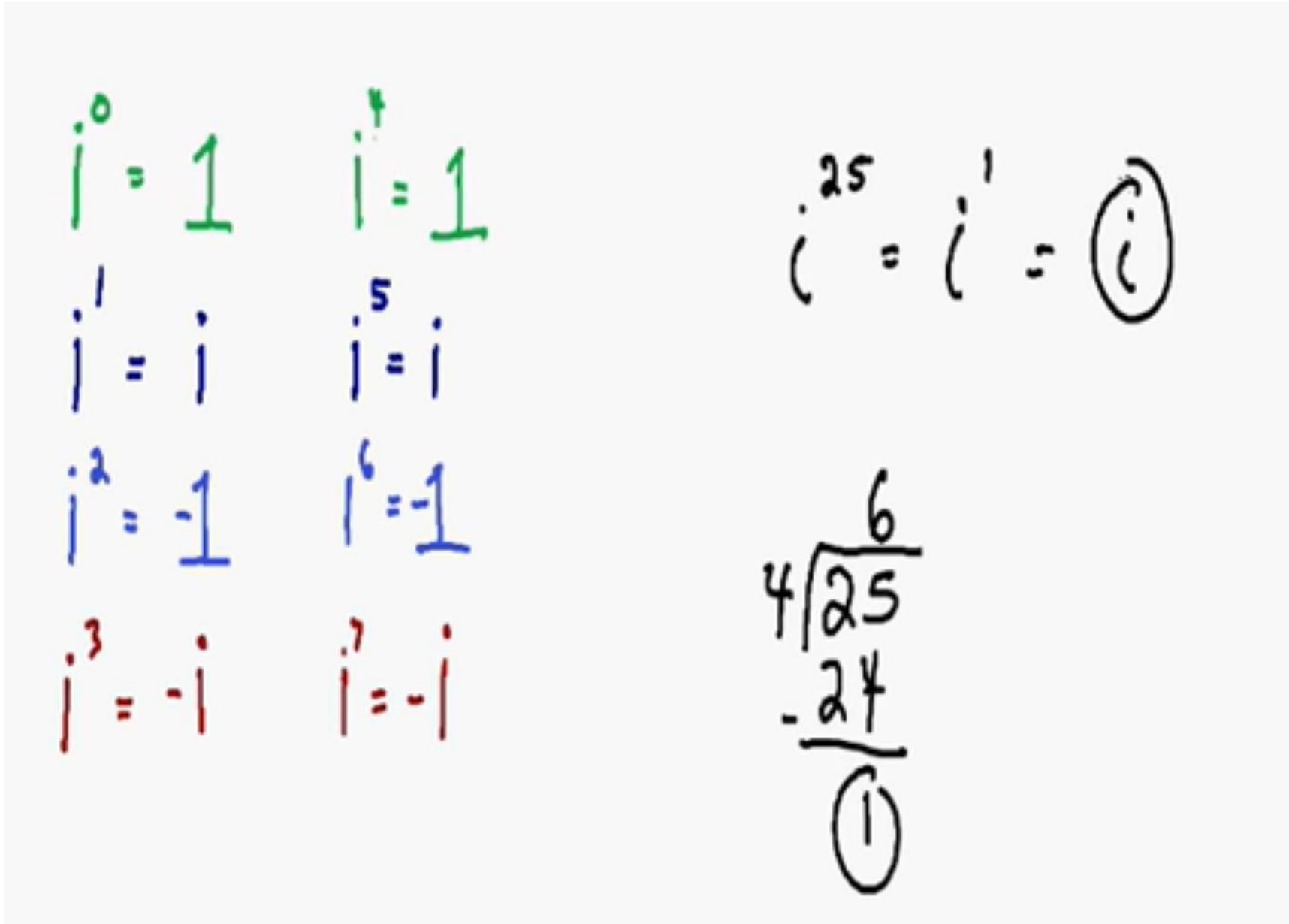
$i^2 = \sqrt{-1}^2$

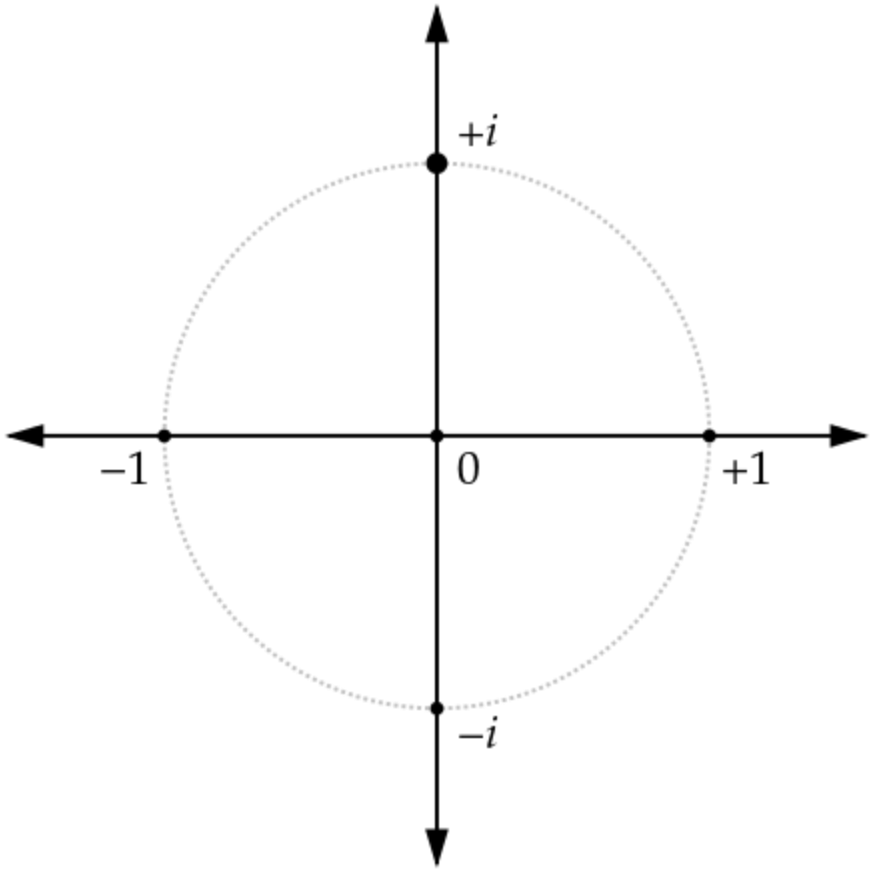
$i^2 = -1$

$i^5 = (i^4)(i^1) = (1)(i) = i$

$i^6 = -1$

$i^7 = -i$


$i^0 = 1$
$i^4 = 1$
$i^1 = i$
$i^5 = i$
$i^2 = -1$
$i^6 = -1$
$i^3 = -i$
$i^7 = -i$
$i^{25} = i^1 = i$
6
4 | 25
-24
1



복소 평면에서의 . 실수는 수평선에 놓이고, 허수는 수직선 위에 위치한다.

Fourier 변환 실습

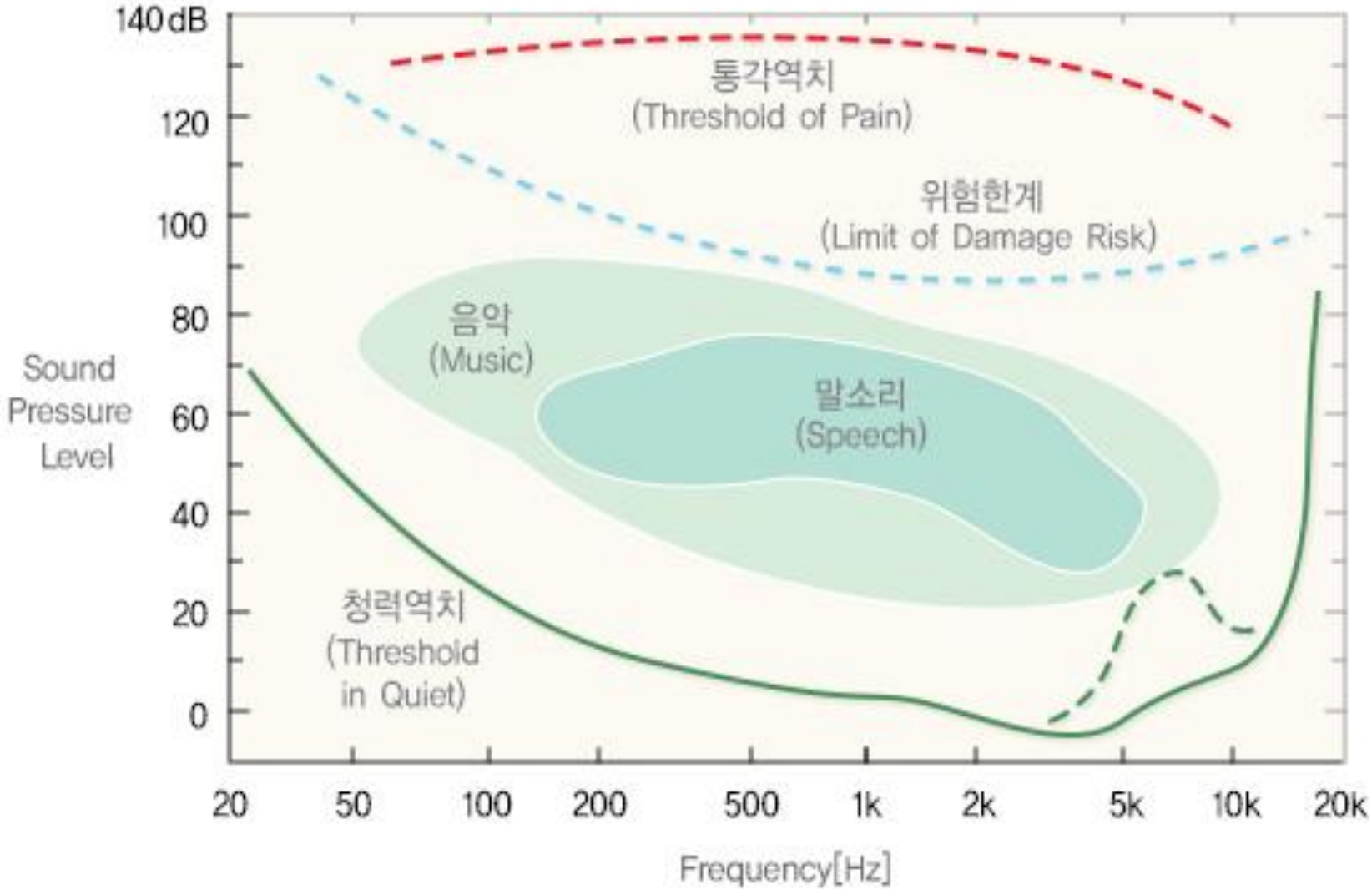
가청역치 (Threshold of Audibility)

최소 가청역치(absolute hearing threshold)는 **검사음의 50% 이상을 인지**할 수 있는, 정적과 겨우 구별할 수 있는 정도의 **최소 강도의 소리**를 말합니다. 최소 가청역치는 환자와 다양한 역치 측정 방법에 따라 차이가 납니다.

청력검사기기에서 사용되는 **청력검사 상의 0(audiometric zero)**이란 청력검사 상의 기준 0점을 말하며, **정상 청력을 가진 사람들의 평균 가청역치**입니다.

큰 소리에 대한 불쾌감은 **불쾌역치**(threshold of discomfort; **TD**, loudness discomfort level; **LDL**, uncomfortable loudness level; **UCL**)라고 부릅니다. 이는 **소리가 불쾌하게 들리기 시작하는 강도**를 말하는 것으로 **120~140dB SPL에서 간지러움 혹은 통증**을 경험합니다. 이것은 촉각이라 추정되며 귓바퀴, 외이도, 고막, 또 다른 중이내 구조의 신경 말단과 관련이 있습니다.

〈그림 인간의 가청역과 소리의 음압 분포〉

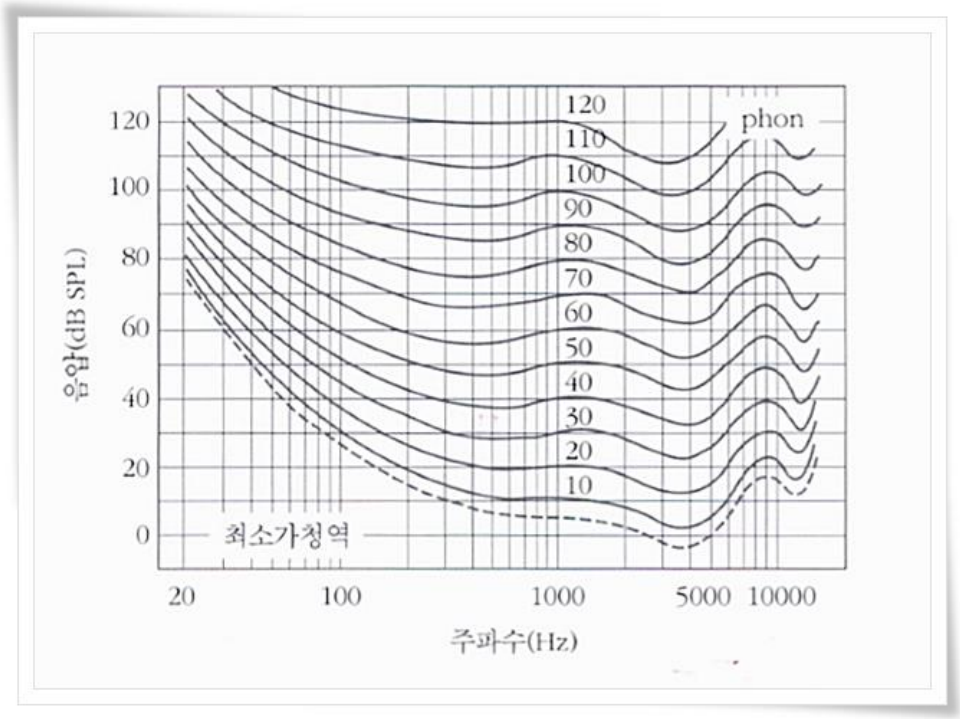


주관적 소리의 크기와 음조

소리의 세 가지 물리적 특성인 주파수, 강도, 주파수역은 인간의 청력 없이도 전기적 기구를 사용하는 물리적 방법으로 쉽고 일관되게 측정될 수 있습니다. 그러나 **인간의 청력만으로만 측정**하여 **주관적 특성을 나타내는 값인 음량(loudness), 음조(pitch)**를 아는 것 역시 중요합니다. 특히 이런 특성들은 주관적 청각과 청감각적 신호에 대한 대뇌 반응을 연구하는 **심리음향학(psychoacoustics)**의 기본적 개념이 됩니다

음량 (Loudness)

음압의 진폭이 커지면 소리가 크게 들리지만, 이들 사이의 관계는 정비례하지 않습니다. 우리가 **주관적으로 느끼는 소리의 강약**을 **음량(loudness)**이라고 합니다. 물리적 음압이 같더라도 주파수가 다르면 음의 강약이 다르게 느껴집니다. 같은 음량 정도를 연결한 것이 **Fletcher-Munson curve**라고도 불리는 **등청감곡선(equal loudness curve)**입니다



이 음량을 정량화한 단위는 phon과 sone입니다. **phon**은 **1000Hz를 기준으로 하여 같은 크기로 들리는 다른 주파수의 SPL 값들을 연결**한 것으로 **1000Hz에서의 dB SPL값**으로 정의합니다. 예를 들어, 1000Hz의 40dB은 40phon으로, 100Hz의 52dB, 10000Hz의 50dB와 같은 크기로 들린다는 것입니다. phon은 실제로 몇 배 크게 들리는가보다는 심리량의 단위로서, 음향 환경에 대하여 쉽게 정량화하는 데에는 사용하기가 어렵습니다. 이를 위해 만든 것이 **sone**으로, **40phon의 1000Hz 순음을 기준으로 만든 음량 척도(loudness scale)가 1 sone**이고, 여기서 소리의 크기가 반 정도로 느껴지면 0.5이라고 할 수 있습니다. sone phon과 sone의 관계는 아래와 같습니다.

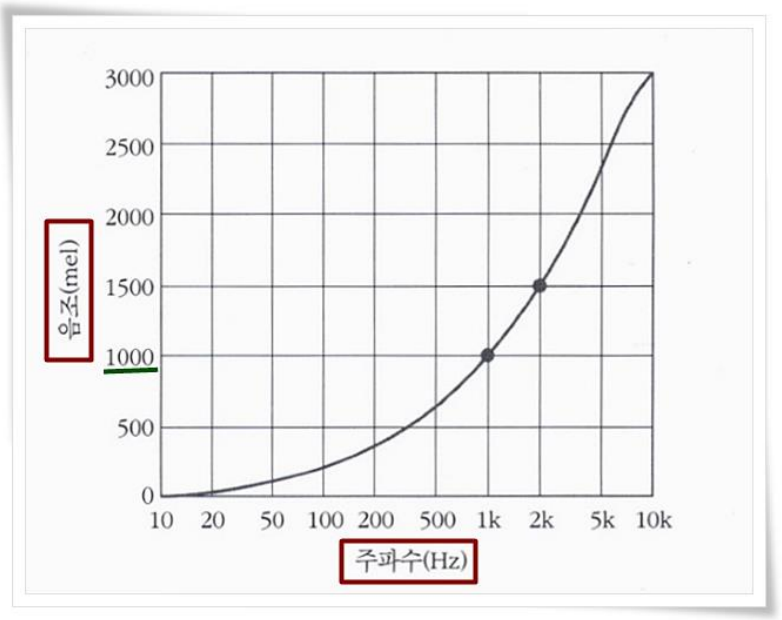
S = 2(P-40)/10

(**S**: sone, **P**: phon)

같은 주파수에서 감지할 수 있는 조그만 음의 크기 변화를 **음의 강도차 판별역치(intensity discrimination limen)**라고 합니다. 이는 자극음에서 인지할 수 있는 가장 작은 변화를 의미하며, 보통 50%의 정확도 확률을 보이는 강도를 역치로 결정합니다.

음조(Pitch)

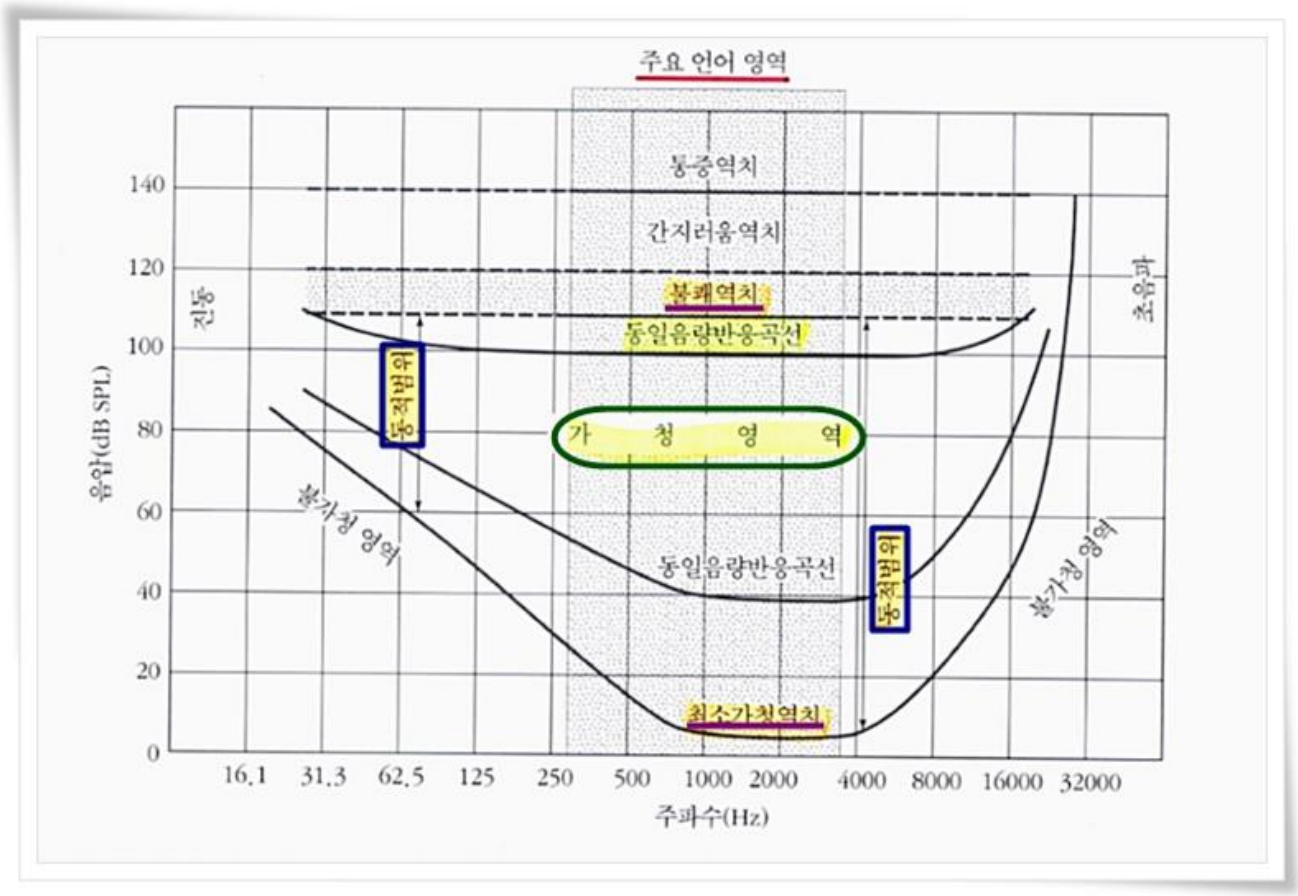
주관적인 음의 고저 감각을 **음조(pitch)**라고 합니다. 음조는 한쪽 또는 양쪽 청력과 밀접한 관계가 있습니다. 음의 주파수가 2배가 된다고 해서 음조도 2배가 되지는 않습니다. 이 음조를 정량화하기 위해서 만든 단위가 **mel**입니다. **1000Hz의 순음을 40dB SPL에서 고정시켰을 때를 1000mel**로 정의합니다. 따라서 2배 높게 들리는 음은 2000mel이 되어 2배의 음조를 느끼게 됩니다.



주파수에 따른 음조의 측정 단위(mel) 그래프 - 음강도는 40dB SPL로 고정

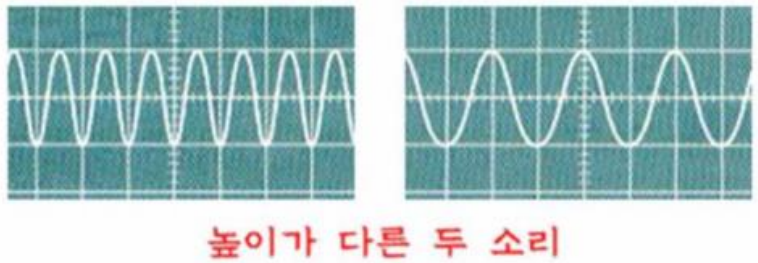
동적 범위 (Auditory Dynamic Range; DR)

가청 영역의 **동적 범위(auditory dynamic range; DR)**는 말을 들었을 때 **불쾌한 소리의 정도와 가청역치의 차이**를 말합니다. 이는 주파수에 따라서 다른데, 중간 주파수 대역보다는 저주파수나 고주파수에서 더 좁습니다.

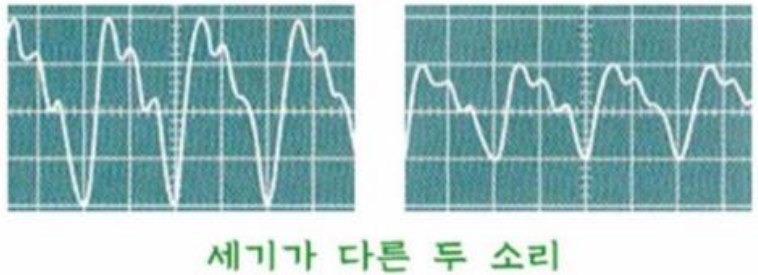


소리에서 얻을 수 있는 물리량

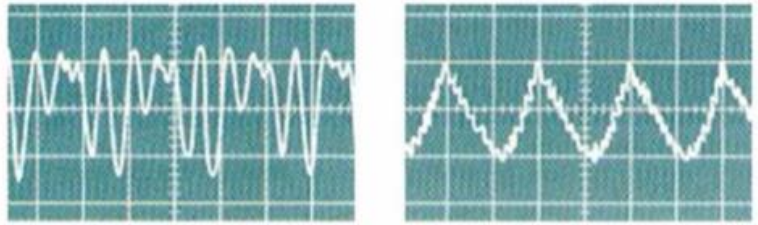
- Amplitude(Intensity) : **진폭**
- Frequency : **주파수**
- Phase(Degress of displacement) : **위상**



높이가 다른 두 소리



세기가 다른 두 소리



맵시가 다른 두 소리

물리 음향

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

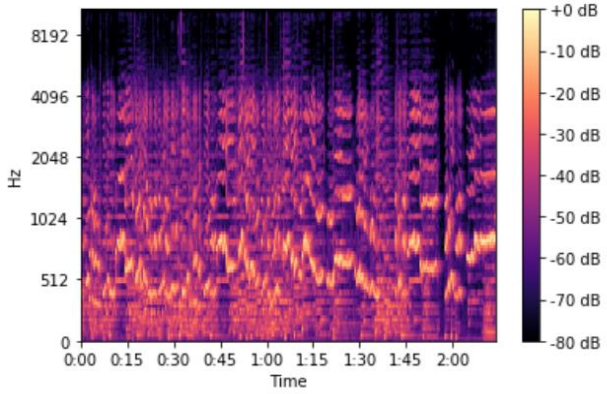
심리 음향

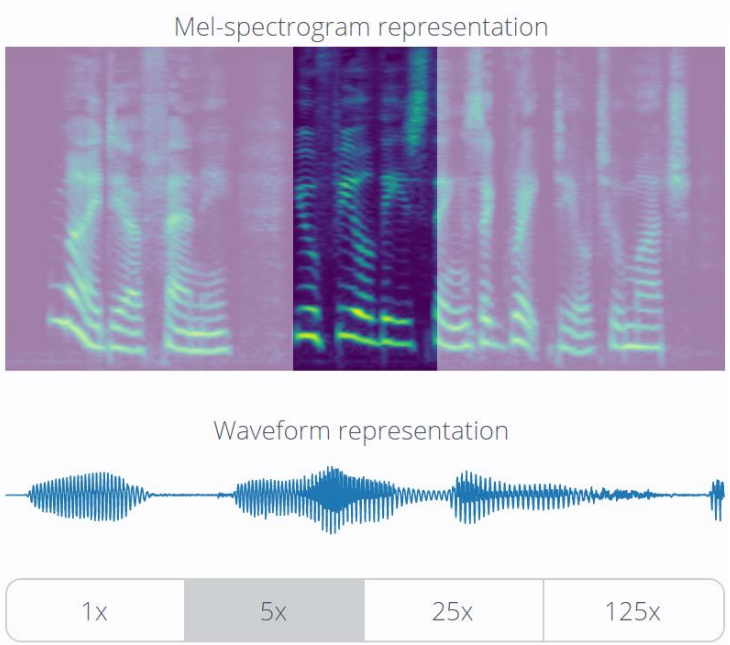
- Loudness : 소리 크기
- Pitch : 음정, 소리의 높낮이 / 진동수
- Timbre : 음색, 소리 감각

https://www.youtube.com/watch?v=s9PQ7qPkluM

The Mel Spectrogram

[9]:
```
S = librosa.feature.melspectrogram(whale_song, sr=sr, n_fft=n_fft,
                                   hop_length=hop_length,
                                   n_mels=n_mels)
S_DB = librosa.power_to_db(S, ref=np.max)
librosa.display.specshow(S_DB, sr=sr, hop_length=hop_length,
                         x_axis='time', y_axis='mel');
plt.colorbar(format='%+2.0f dB');
```





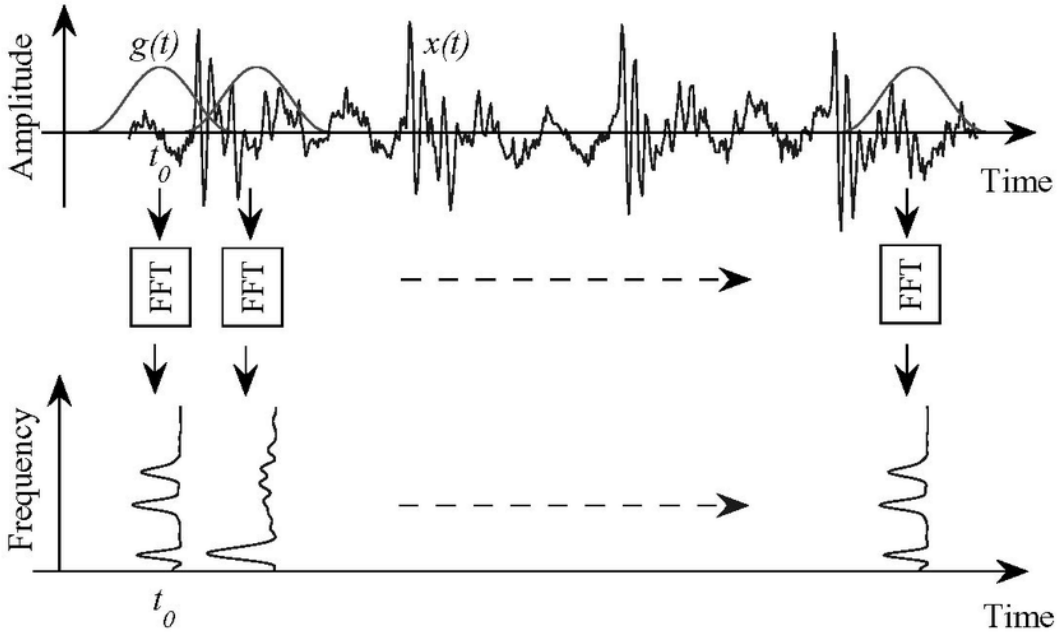
First rows

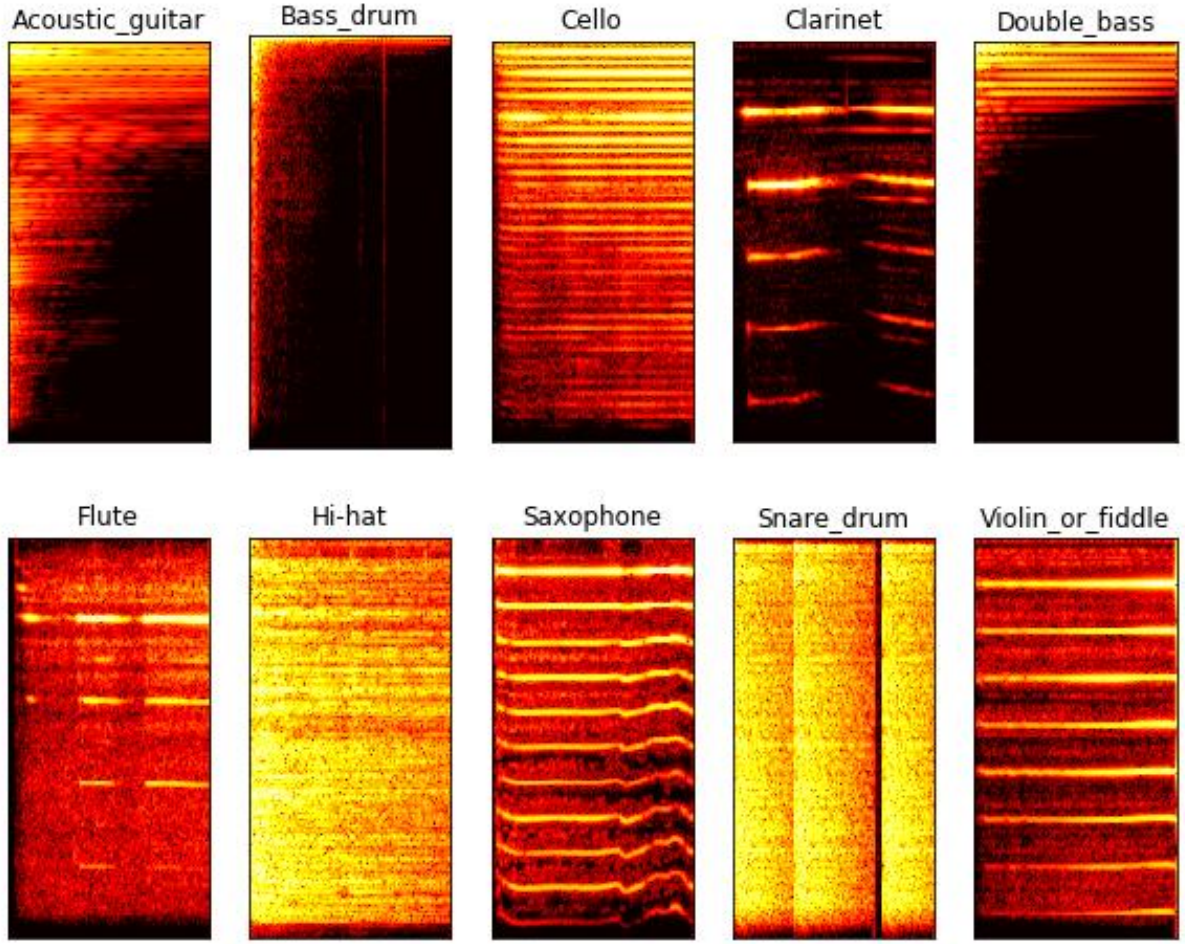
	0	1	10	100	101	102	103	104	105
0	0.071488	0.112931	0.093710	0.003623	0.002465	0.000812	0.001944	0.004218	0.004891
1	0.327453	0.696546	1.089640	0.010338	0.006793	0.005620	0.006476	0.013521	0.017855
2	0.310112	0.253482	0.210603	0.009782	0.007759	0.017301	0.011055	0.008583	0.010366
3	0.010208	0.109415	0.129118	0.020580	0.003772	0.006510	0.011296	0.012814	0.016841
4	0.053919	0.066492	0.029974	0.007042	0.011837	0.012552	0.007069	0.012861	0.014812
5	0.254680	0.086527	0.016534	0.011019	0.005026	0.008153	0.011811	0.008895	0.008429
6	0.208996	0.060032	0.019137	0.016377	0.009610	0.007126	0.014477	0.010154	0.013721
7	0.392639	0.159828	0.190023	0.123125	0.096298	0.053775	0.071976	0.059147	0.069463
8	0.320323	0.092376	0.031002	0.028073	0.022313	0.012524	0.011378	0.007347	0.061672
9	0.108713	0.042663	0.011409	0.142361	0.072162	0.047233	0.037520	0.051558	0.181629

MelNet combines various representational and modelling improvements to yield a highly expressive, broadly applicable, and fully end-to-end generative model of audio.

Short-Time Fourier Transform

Sampling Rate = 16 kHz
Window Length = 25 ms = 400 samples
Step Size = 10 ms = 160 samples
N FFT = 512 samples

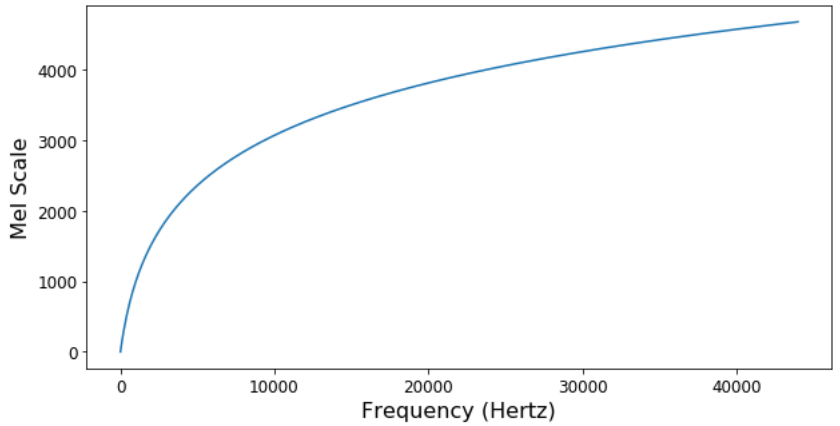


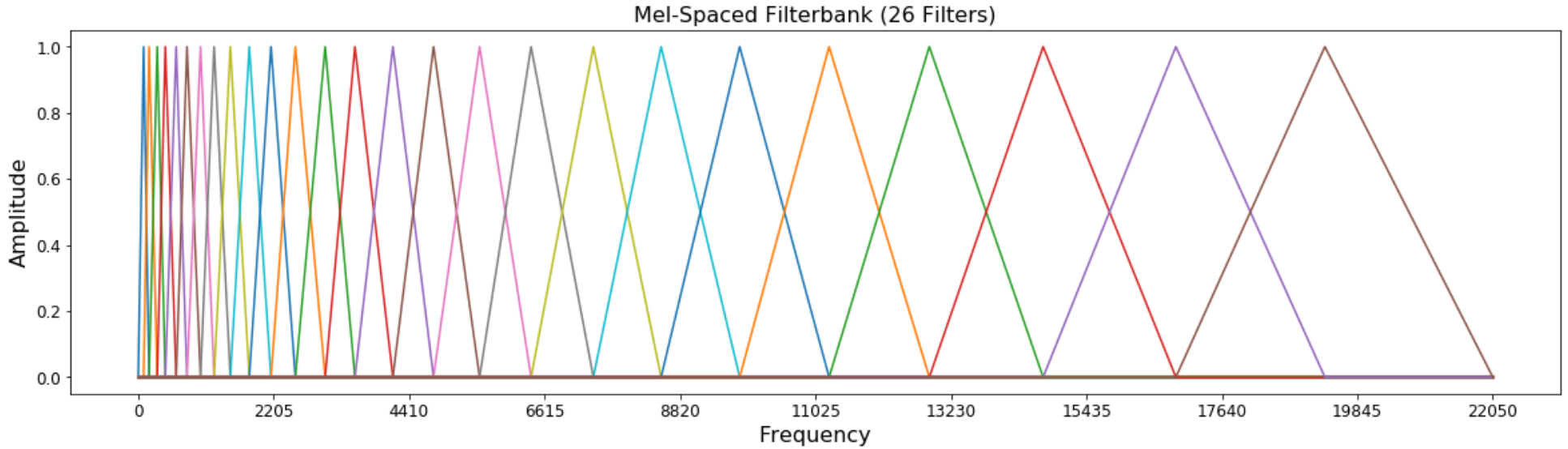


Mel Filterbank

$$M(f) = 1125 \ln(1 + f/700)$$

$$M^{-1}(m) = 700(\exp(m/1125) - 1)$$





실습 파일

https://github.com/JSJeong-me/Sound

정 준 수 Ph.D.

jsjeong@hansung.ac.kr

소프트웨어를 아는 자가 미래를 연다!