

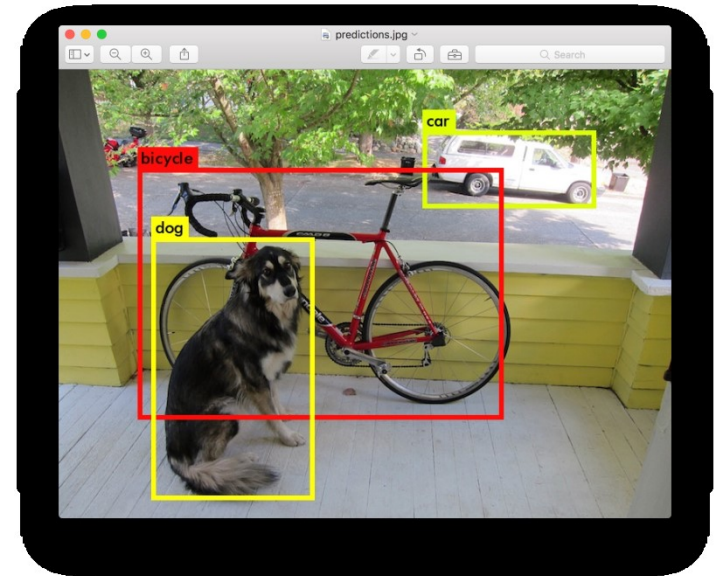
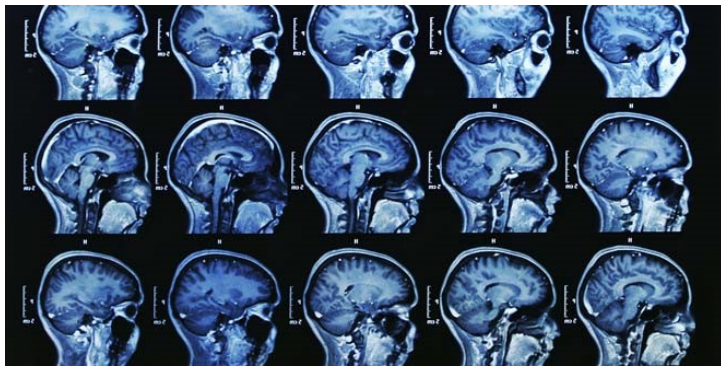
# 딥러닝으로 Sound Event Detection



J.MARPLE

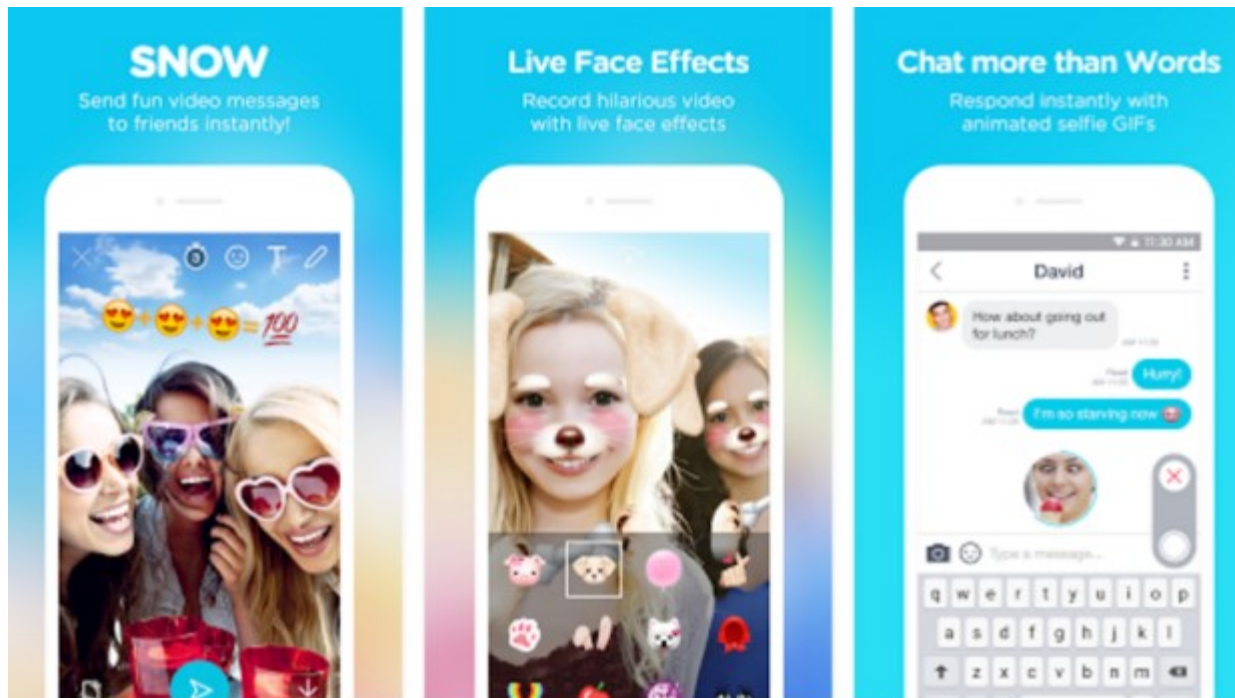
황승원

# 가장 '핫'한 기술 딥러닝



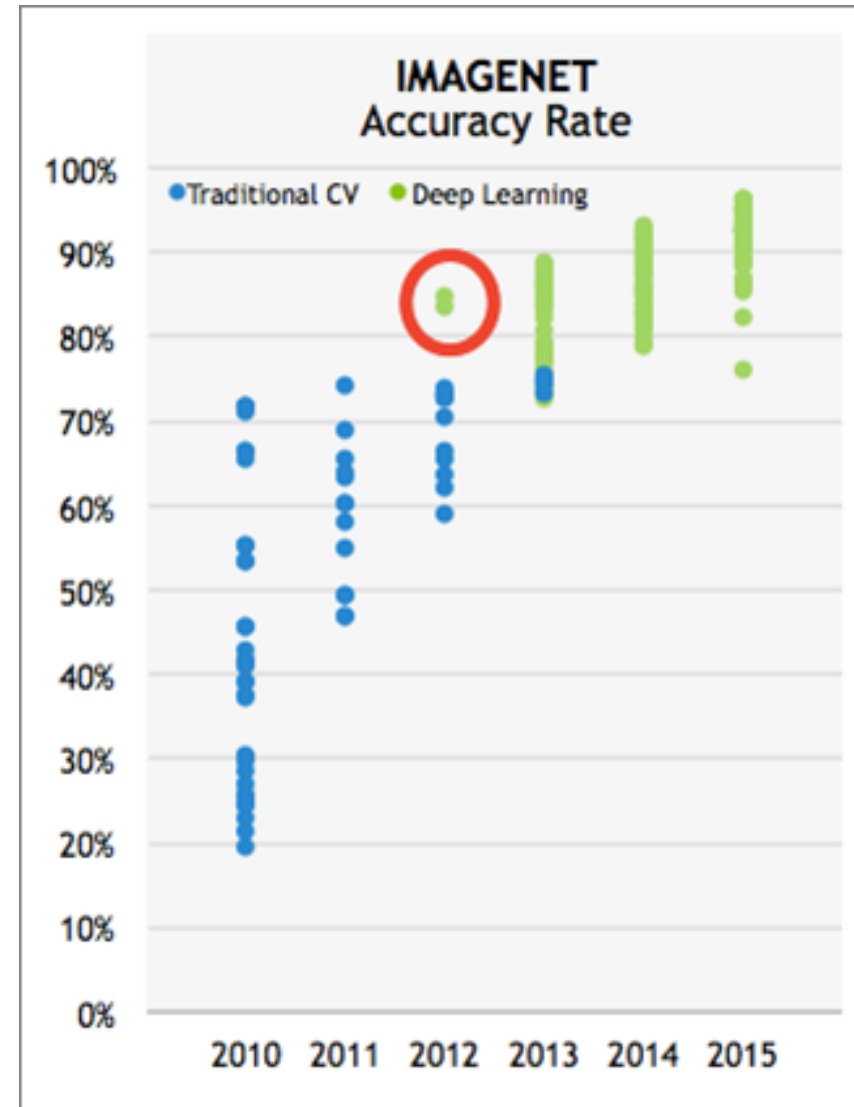
# CNN

- CNN (Convolution Neural Network)
  - Convolution 필터



# CNN

- **IMAGENET**
  - 이미지 인식 경연대회
  - 2011년까지는 인식률이 75%를 못 넘었었음
  - 2012년 CNN을 활용한 Alexnet 등장, 뛰어난 성능 발휘



# CNN



# DCASE

- **DCASE (Detection and Classification of Acoustic Scenes and Events)**
  - 소리 인식 경연대회
  - 2013년부터 시작, 2016년 2회 대회 개최, 2017년 3회째를 맞이하고 있음
  - NMF 등의 기법에서 딥러닝을 활용한 기법으로 바뀌어나가는 것을 확인할 수 있음

# DCASE



- **DCASE 2016**
  - Task 1 : Acoustic scene classification
  - **Task 2 : Sound event detection in synthetic audio**
    - Class : Clearing throat, Coughing, Door knock, Door slam, Drawer, Human laughter, Keyboard, Keys (put on table), Page turning, Phone ringing, Speech
  - Task 3 : Sound event detection in real life audio
  - Task 4 : Domestic audio tagging

# DCASE

- **Baseline**
  - Acoustic scene classification을 위한 basic approach
  - 2016년 Task 2의 Baseline code는 NMF(non-negative matrix factorization, 비음수행렬분해)로 되어 있음



# DCASE

Rank	Submission Information		Corresponding		Segment-based (overall)	
	Code	Name	Author	Affiliation	ER 	F1 
1	Komatsu	Komatsu	Tatsuya Komatsu	NEC Corporation, Japan	0.3307	80.2 %
2	Choi	Choi	Inkyu Choi	Seoul National University, South Korea	0.3660	78.7 %
3	Hayashi_1	BLSTM-PP	Tomoki Hayashi	Nagoya University, Japan	0.4082	78.1 %
4	Hayashi_2	BLSTM-HMM	Tomoki Hayashi	Nagoya University, Japan	0.4958	76.0 %
5	Phan	Phan	Huy Phan	University of Lubeck, Germany	0.5901	64.8 %
6	Giannoulis	Giannoulis	Panagiotis Giannoulis	National Technical University of Athens, Athena Research and Innovation Center, Greece	0.6774	55.8 %
7	Pikrakis	Pikrakis	Aggelos Pikrakis	University of Piraeus, Greece	0.7499	37.4 %
8	DCASE	DCASE2016_Baseline	Emmanouil Benetos	Queen Mary University of London, United Kingdom	0.8933	37.0 %
9	Vu	Vu	Toan H. Vu	National Central University, Taiwan	0.8979	52.8 %
10	Gutierrez	Gutierrez	J.M. Gutierrez-Arriola	Universidad Polit�cnica de Madrid, Spain	2.0870	25.0 %
11	Kong	Kong	Qiuqiang Kong	University of Surrey, United Kingdom	3.5464	12.6 %

# DCASE

- **DCASE 2017**
  - Task 1 : Acoustic scene classification
  - **Task 2 : Detection of rare sound events**
    - Class : Baby crying, Glass breaking, Gunshot
  - Task 3 : Sound event detection in real life audio
  - Task 4 : Large-scale weakly supervised sound event detection for smart cars













# DCASE

- 2017 DCASE Task 2 baseline
  - MLP (multi-layer perceptron) = Deep Learning

Event-based overall metrics (onset only,  $t_{\text{collar}}=500\text{ms}$ )

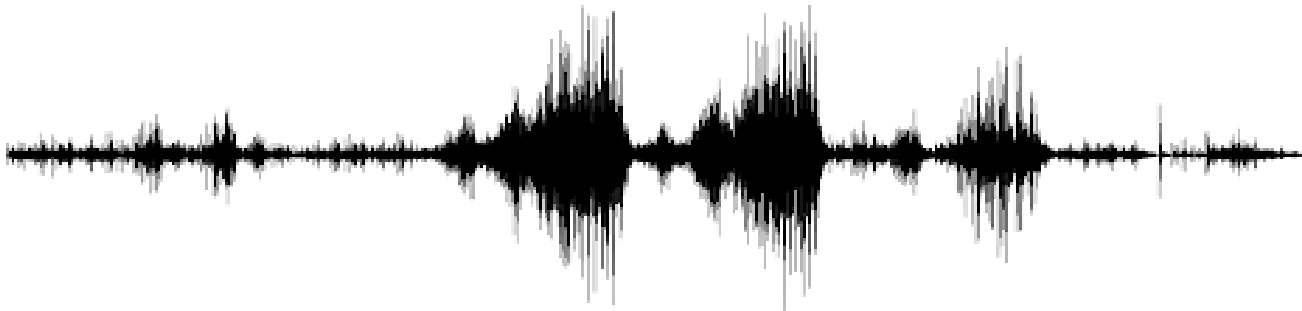
	ER	F-score
Baby cry	0.67	72.0 %
Glass break	0.22	88.5 %
Gun shot	0.69	57.4 %
Average	0.53	72.7 %

# DCASE

Rank	Submission Information		Technical Report	Event-based (overall / evaluation dataset)	
	Code	Name		ER 	F1 
1	Lim_COCAI_task2_1	1dCRNN1		0.1307	93.1 %
2	Lim_COCAI_task2_2	1dCRNN2		0.1347	93.0 %
3	Lim_COCAI_task2_3	1dCRNN3		0.1520	92.2 %
4	Lim_COCAI_task2_4	1dCRNN4		0.1720	91.4 %
5	Cakir_TUT_task2_2	CRNN-2		0.1733	91.0 %
6	Cakir_TUT_task2_1	CRNN-1		0.1813	91.0 %
7	Cakir_TUT_task2_4	CRNN-4		0.1867	90.3 %
8	Phan_UniLuebeck_task2_1	AED-Net		0.2773	85.3 %
9	Cakir_TUT_task2_3	CRNN-3		0.2920	86.0 %
10	Zhou_XJTU_task2_1	SLR-NMF		0.3133	84.2 %

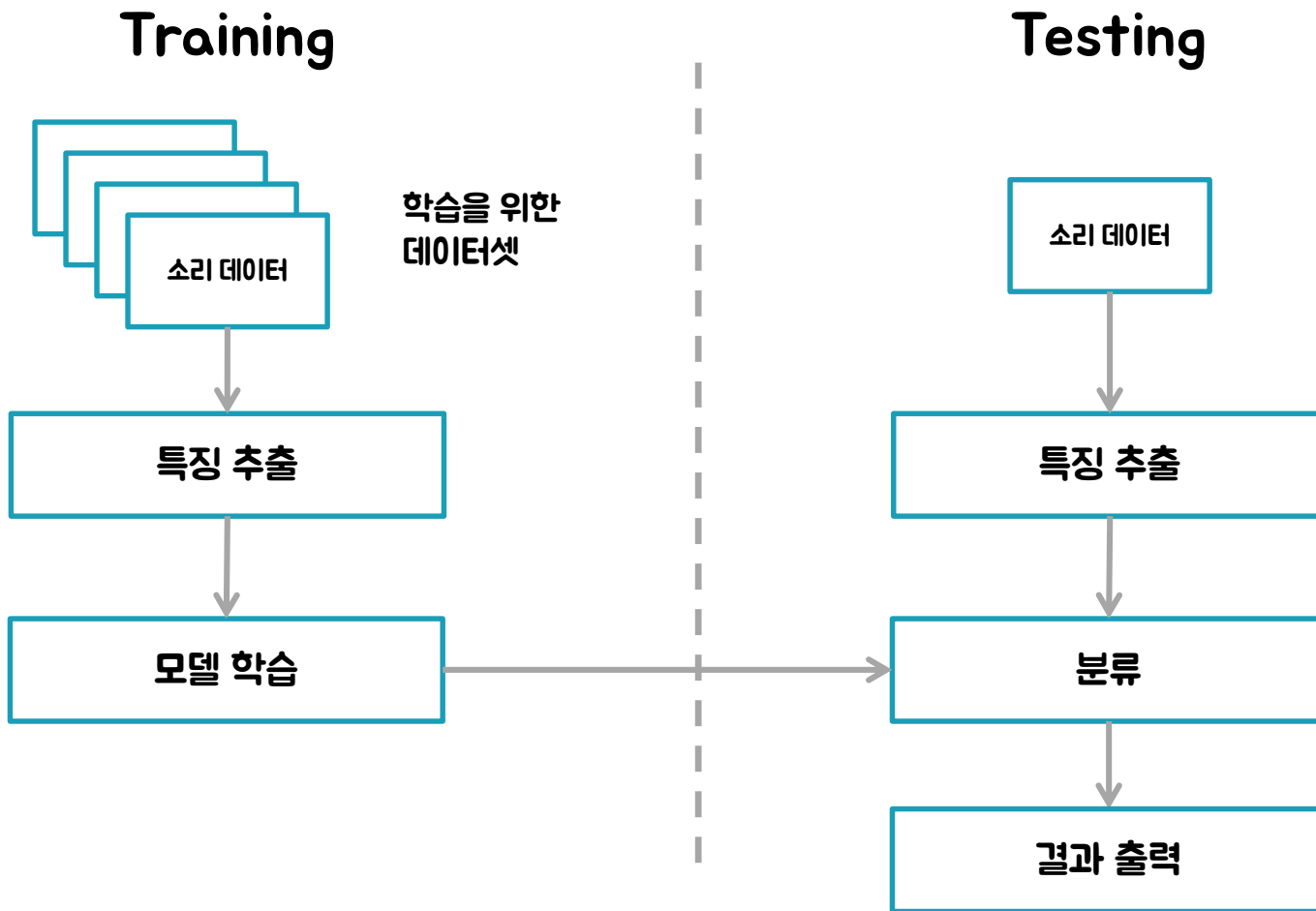
# 어떻게 가능한가?

- 소리 데이터 = 시계열 데이터

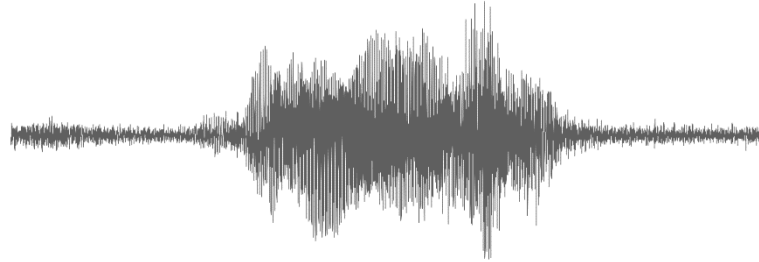


- 어떻게 딥러닝을 적용시킨걸까??

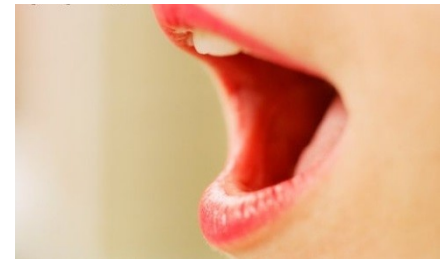
# 소리 인식의 개념



# 소리 데이터의 특징



?



# 소리 데이터의 특징

- **시계열 데이터 상태에서는**
  - 어떤 소리가 들어있는지 알기 어려움
  - 그 소리가 어떤 특징을 가지고 있는지 알기 어려움
- **신호 처리를 통해 소리의 특징을 분석할 필요가 있음**
  - 주파수 성분을 보면 대략적으로 파악 가능



# 기본적인 신호처리

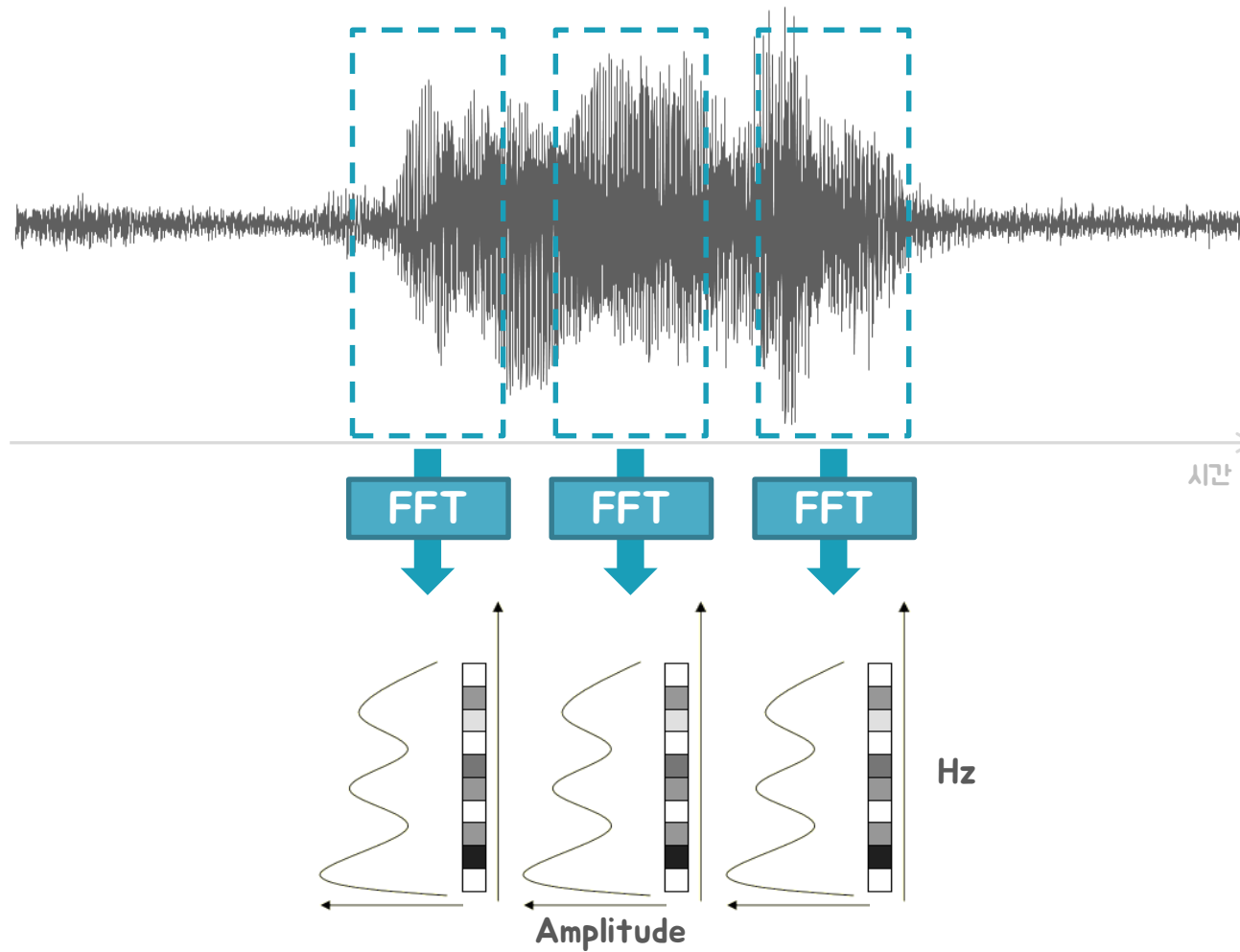
- FFT
  - 주파수 성분을 알 수 있음



- But, 시간축을 잃어버림

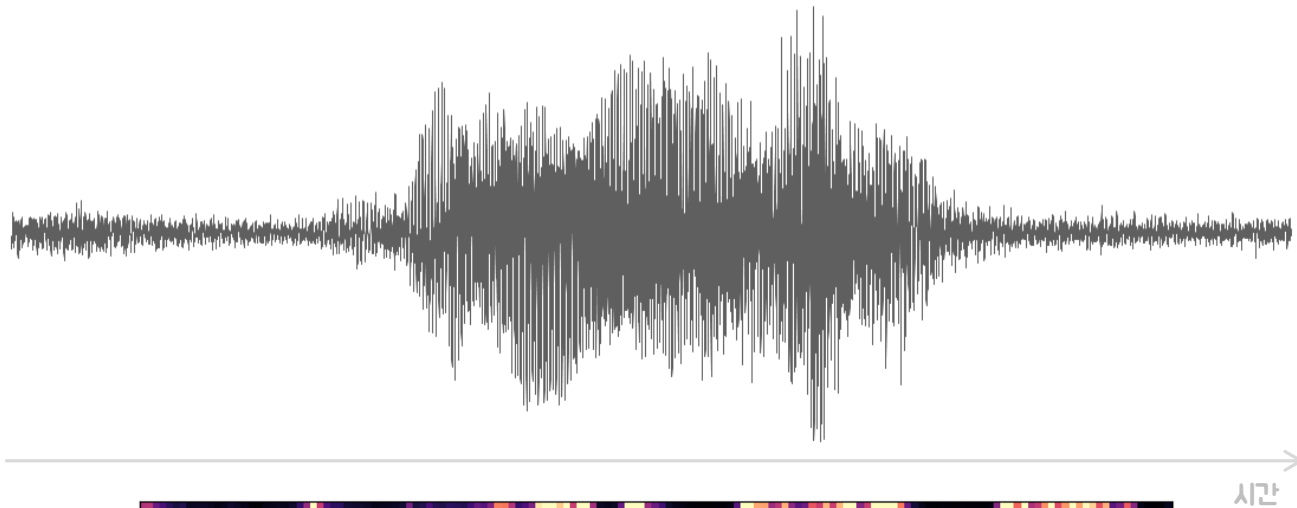
# 기본적인 신호처리

- STFT와 Spectrum



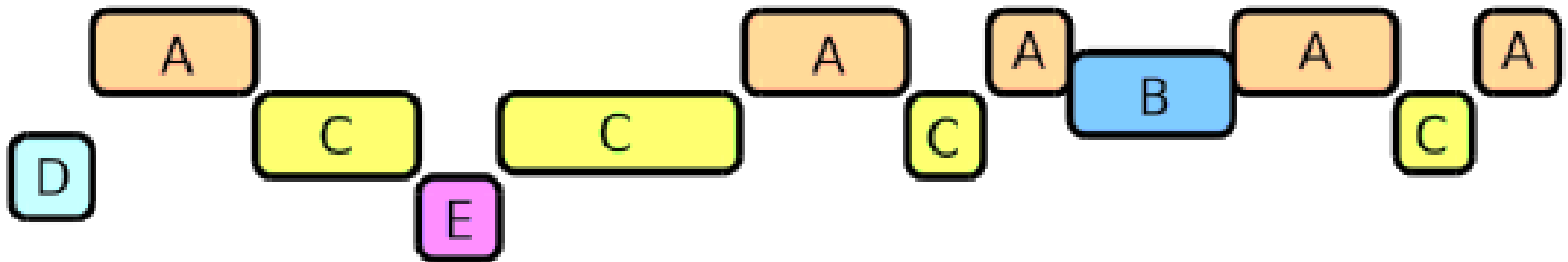
# 기본적인 신호처리

- Spectrogram



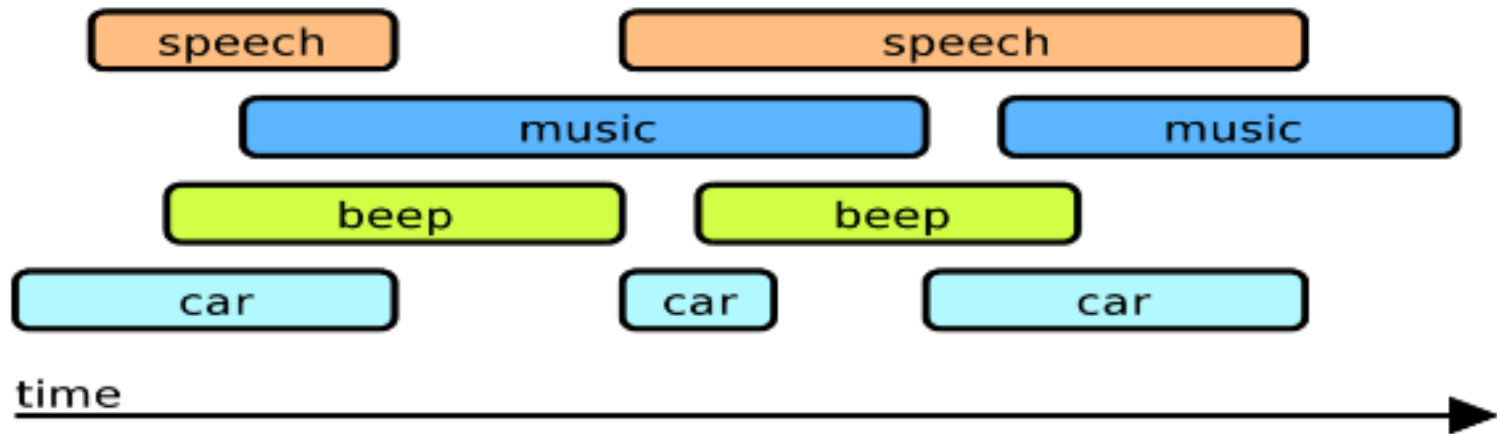
# 특징 추출의 어려움

- 우리의 바람



# 특징 추출의 어려움

- 현실

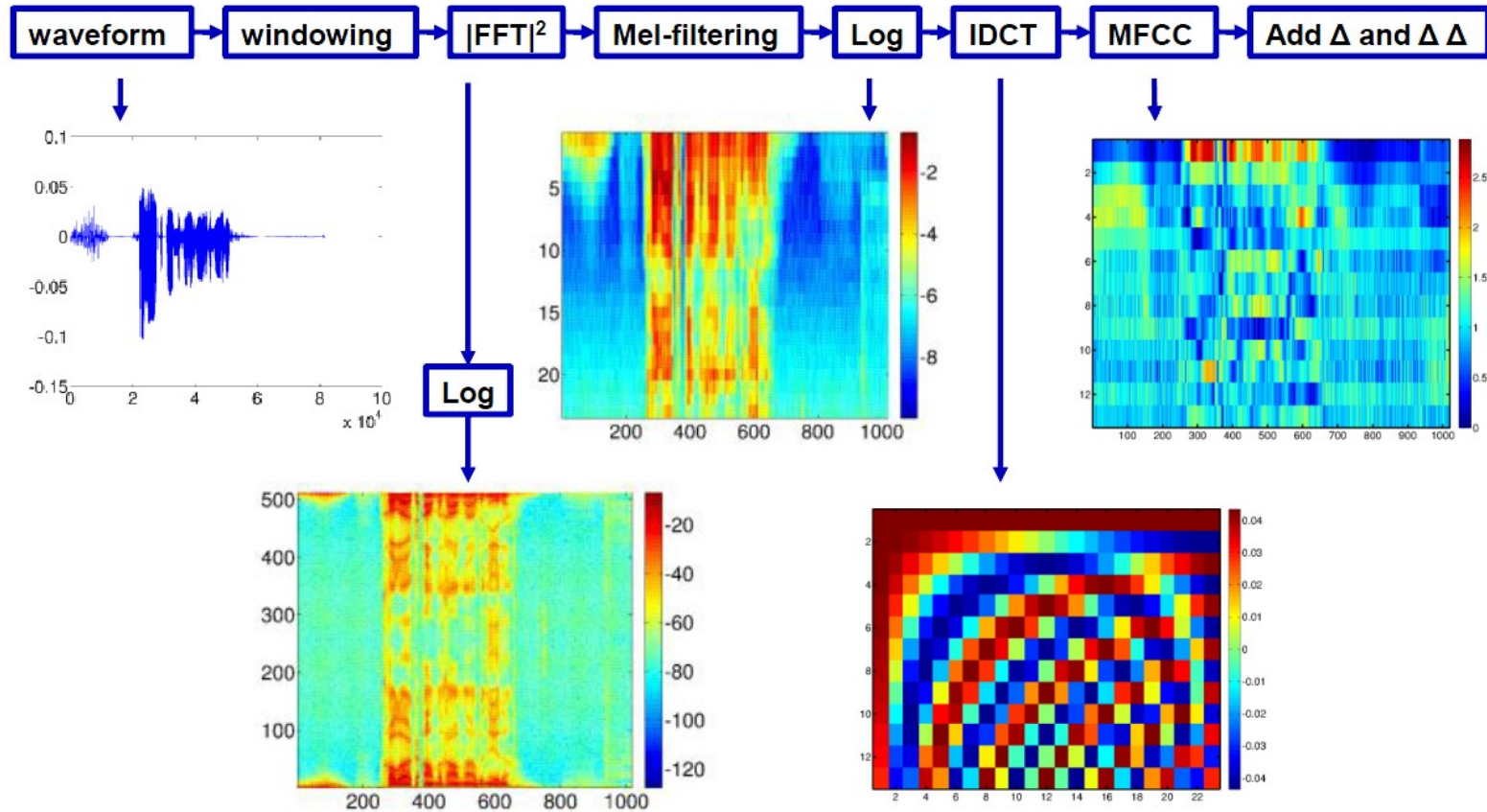


- 겹쳐있는 각 소리 성분들의 특징이 잘 드러나게 특징 추출을 해야 분류를 잘 할 수 있음

# 특징 추출의 어려움

- **Spectro-temporal representation**
  - Mel-spectral, MFCC, LPC, gammatone, subband autocorrelation 등
- **Summary statistics**
  - Zero crossing-rate, spectral bandwidth 등
- **Dimensionality reduction**
  - PCA, Feature selection 등

# 특징 추출의 어려움



# 특징 추출의 어려움

- 특징 추출을 위해 다양한 기법 활용해야 함
- Hyper parameter 튜닝을 잘해야 함
- ▶ Feature Engineering에 필요한 노력이 큼
- CNN을 활용하여 Feature Engineering에 대한 부담 줄이려함
  - CNN을 활용하려면 소리를 이미지로 변환해야 함

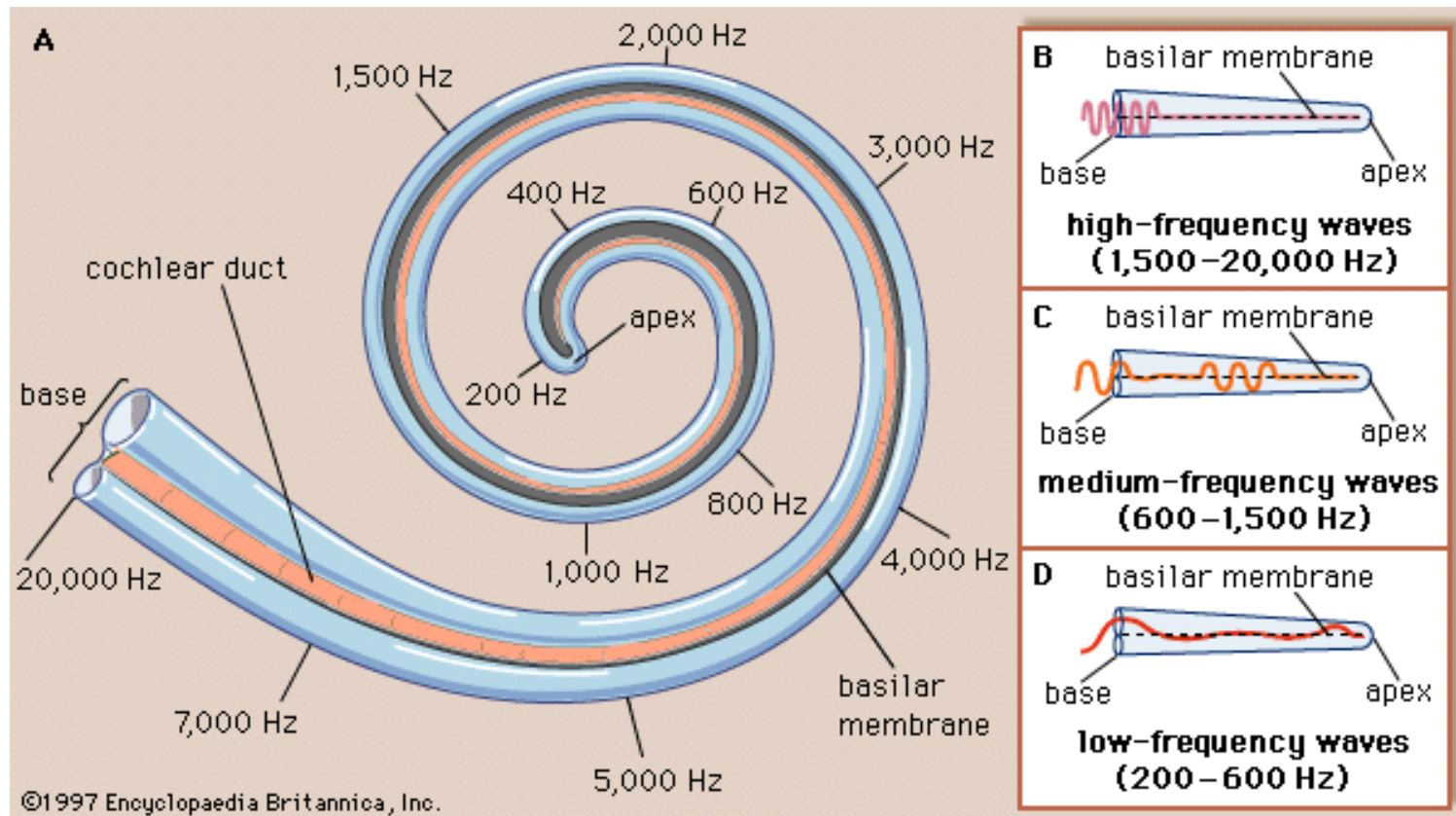


# 딥러닝용 특징 추출법

- Log-amplitude Mel-spectrogram + Convolution
  - Feature Engineering에 대한 부담이 크게 줄어듦
  - 특징 추출에 사람의 개입이 적음
  - 사람의 청각 특성이 반영되어 있음 (비선형성)
    - 주파수 축 : Mel-scale을 반영
    - Amplitude : Log를 취해줌

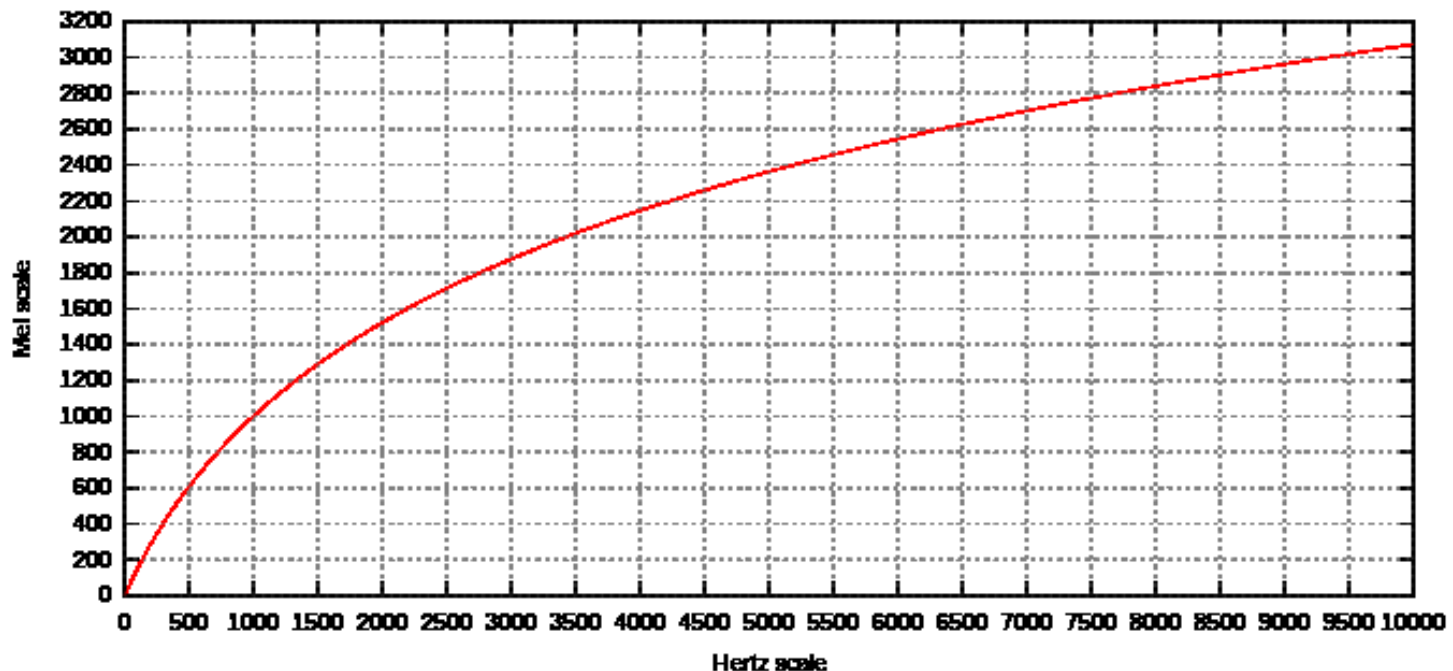
# Log-amplitude Mel-spectrogram

- 사람의 귀는 소리를 비선형적으로 받아들임



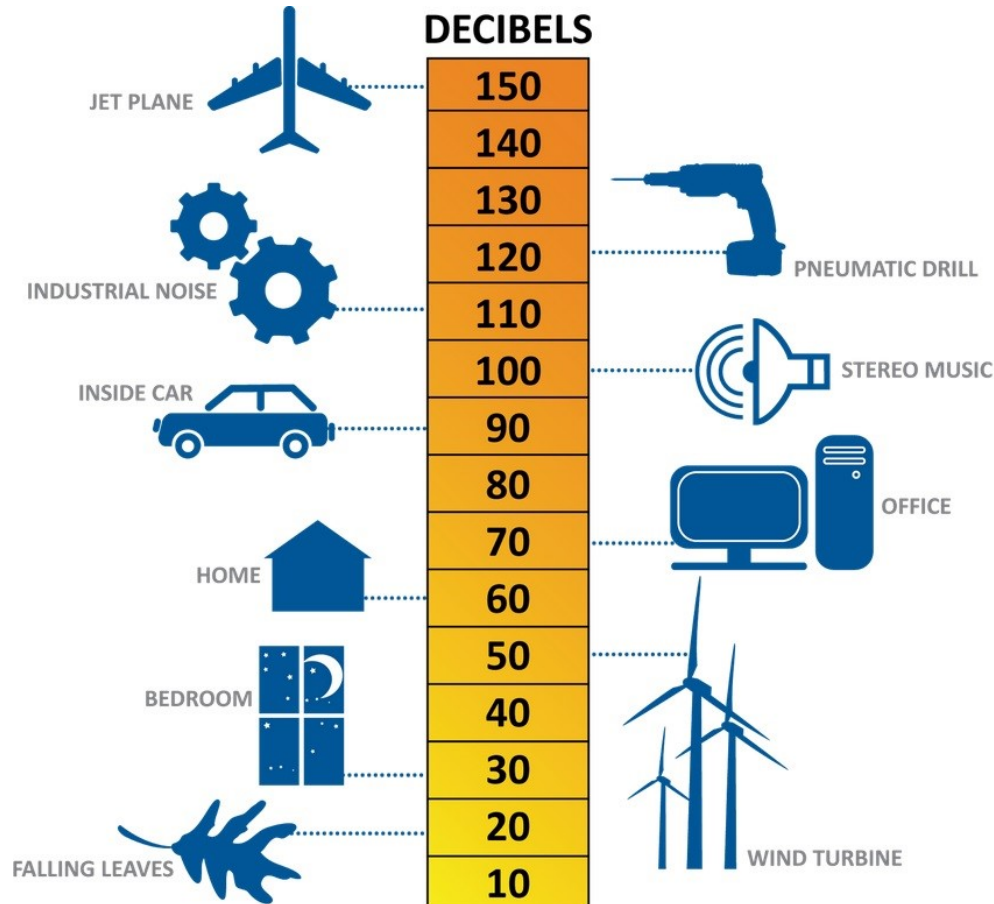
# Log-amplitude Mel-spectrogram

- Mel 곡선은 사람의 귀 속 달팽이관의 특성을 반영함
- STFT 이후 얻어진 spectrogram의 주파수 성분을 Mel 곡선에 따라 압축



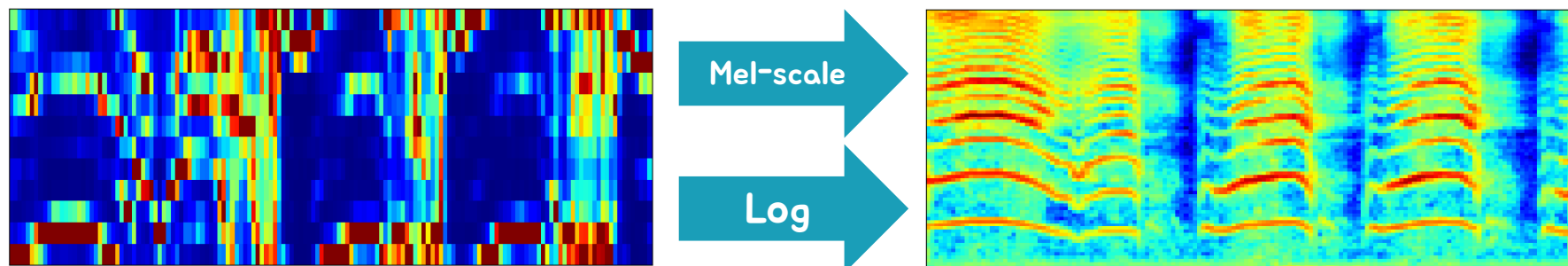
# Log-amplitude Mel-spectrogram

- dB



# Log-amplitude Mel-spectrogram

- 소리 크기도 비선형성을 나타내기 위해  
log를 취해줌 (amplitude squared to dB units)



# Log-amplitude Mel-spectrogram

- 가장 대표적인 하이퍼 파라미터

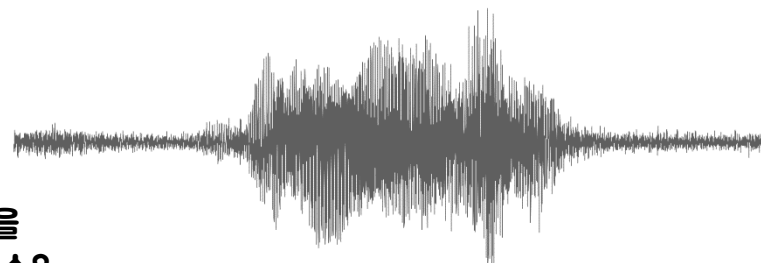
- Sampling rate

- Mel-band

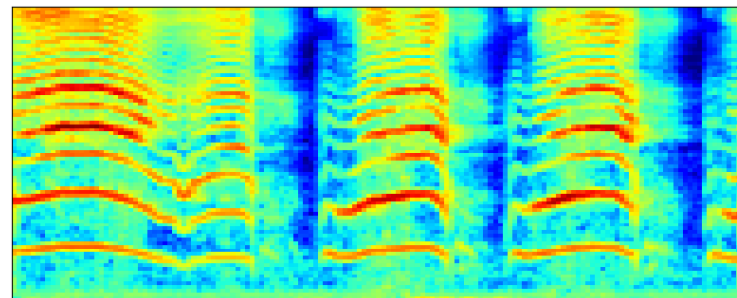
- Window size

- Hop-length

1초에 몇 개의 data?



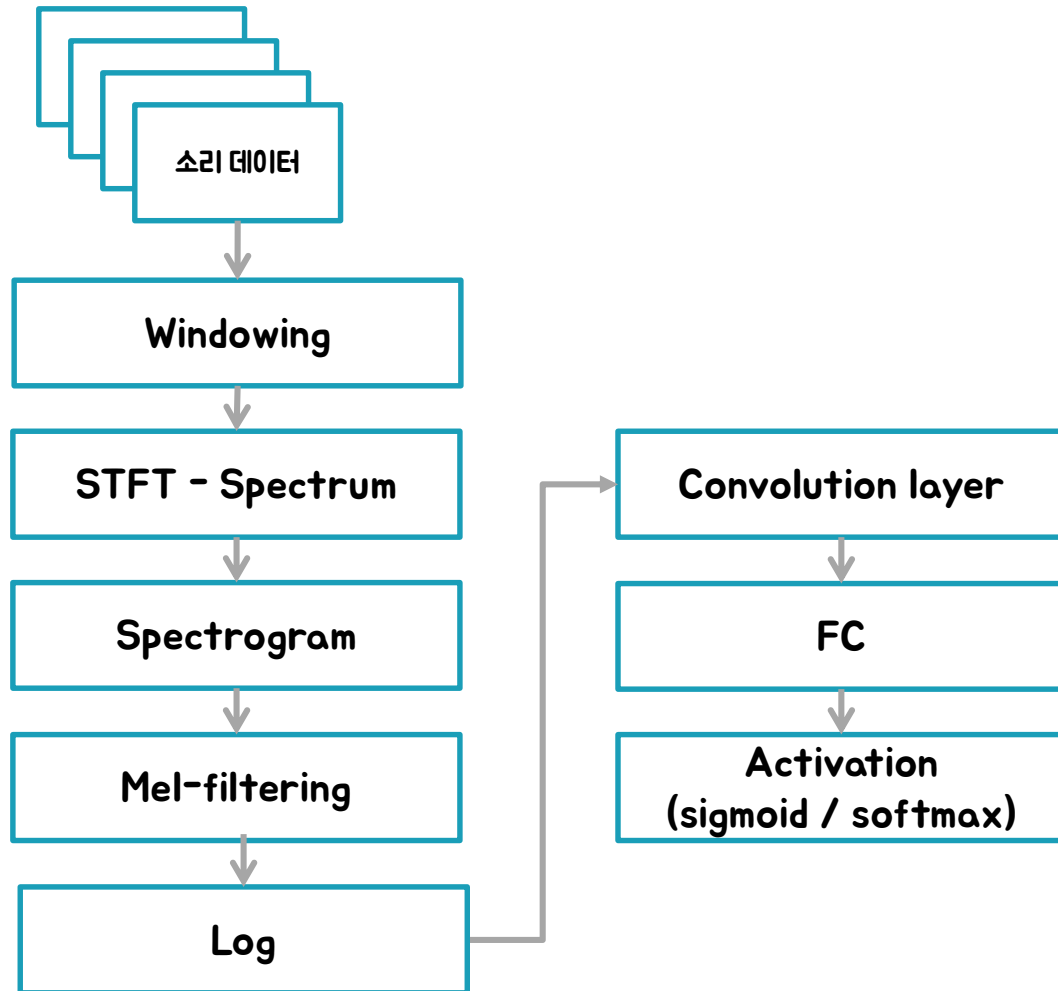
주파수 축을  
몇 개로 축소?



몇 초 단위로 STFT?

몇 초 범위로 STFT?

# 중간 정리



# DCASE 사례 분석

- DCASE 2017 Task 2
- RARE SOUND EVENT DETECTION USING 1D CONVOLUTIONAL RECURRENT NEURAL NETWORKS,  
Hyungui Lim, Jeongsoo Park,  
Yoonchang Han

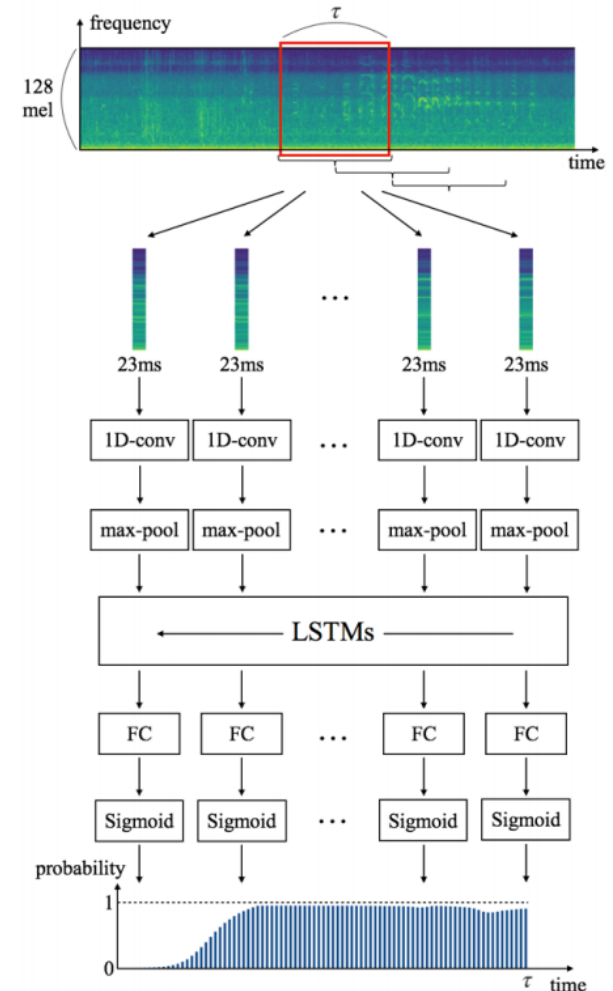


Figure 1: Overall framework of the proposed method.

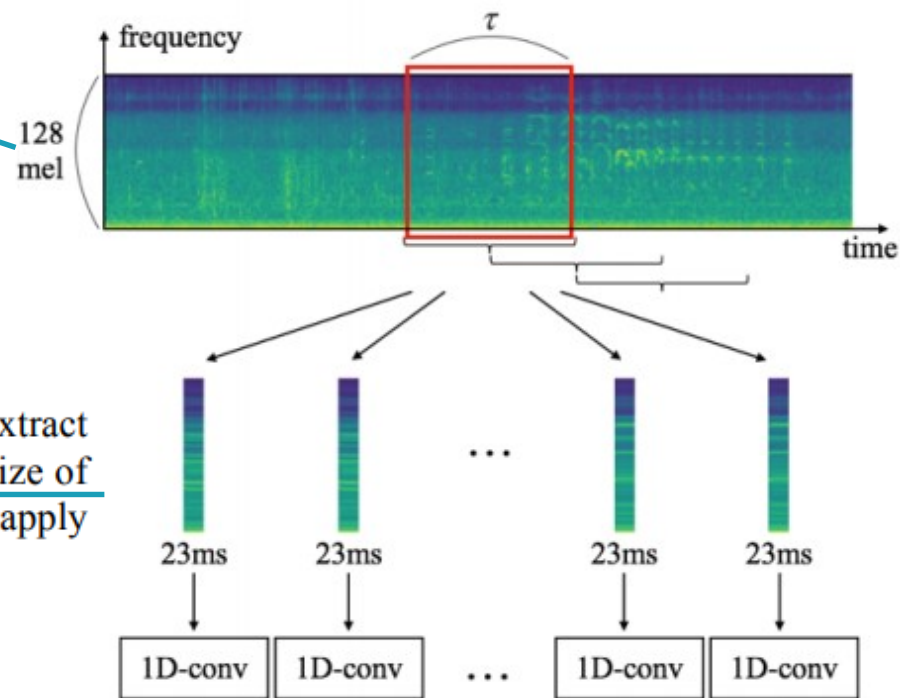


# DCASE 사례 분석

- Mel-band : 128
- Window size : 46ms
- Hop-length : 23ms

also use it as the input feature of our proposed method. To extract this feature, a window is applied to an audio signal with a size of 46 ms and overlapped with half size of the window. We also apply

- Mel-spectrum을  
1D-conv로 특징 추출



# 소리 인식의 활용 방안

- Multimedia information retrieval:
  - 환호성 등 특정 소리만 인식해 알려줌



- 비디오 auto tagging 등

# 소리 인식의 활용 방안

- 모니터링 / 감시:
  - 총소리 / 비명소리 감지
  - 차량 인식
  - 충돌 감지

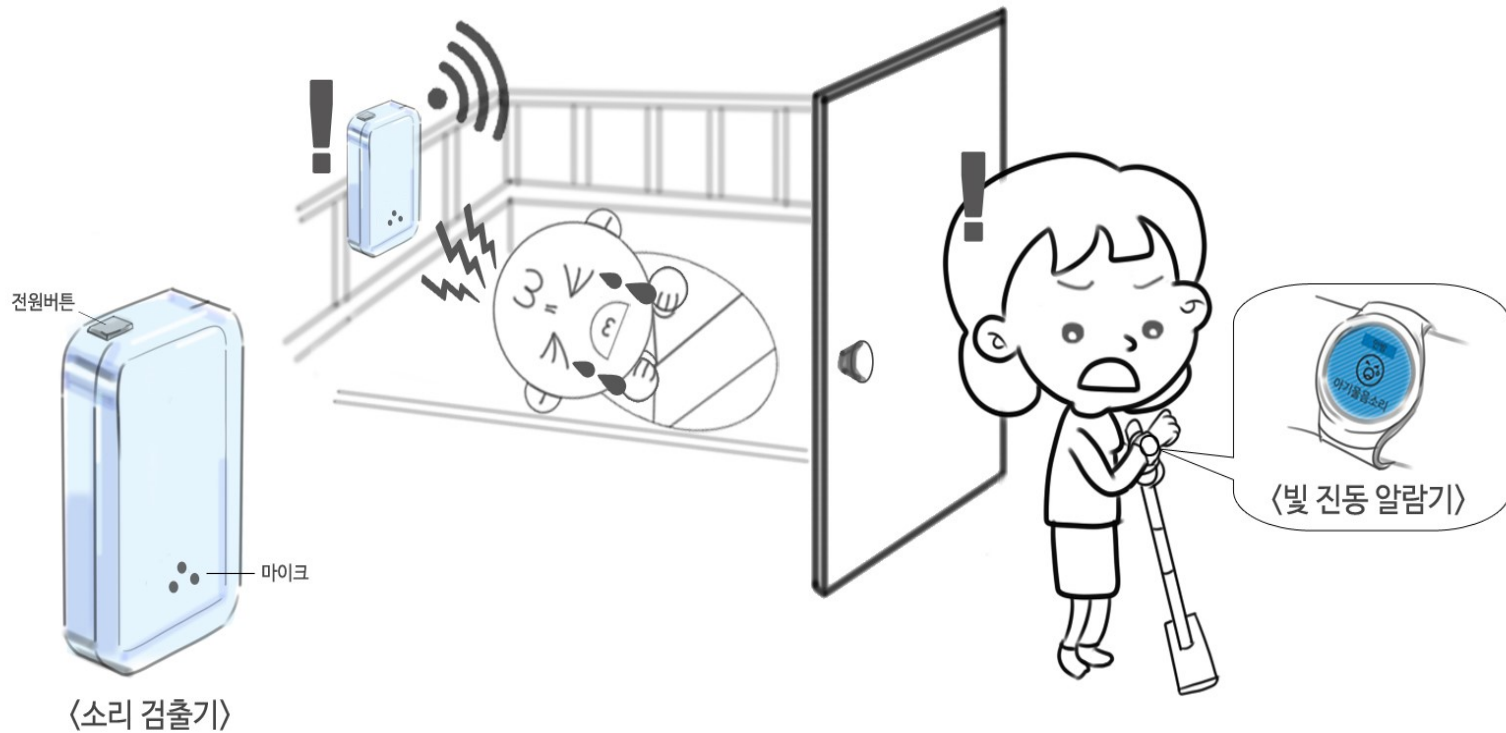


# 소리 인식의 활용 방안

- 보조 기술(Assistive technologies) :
  - 청각장애인을 위한 Sound visualization
  - 낙상 감지 (Acoustic fall detection)
  - 일상생활 모니터링 (Lifestyle monitoring)

# 제이마플의 활용 방안

- 청각장애인용 소리 인식기



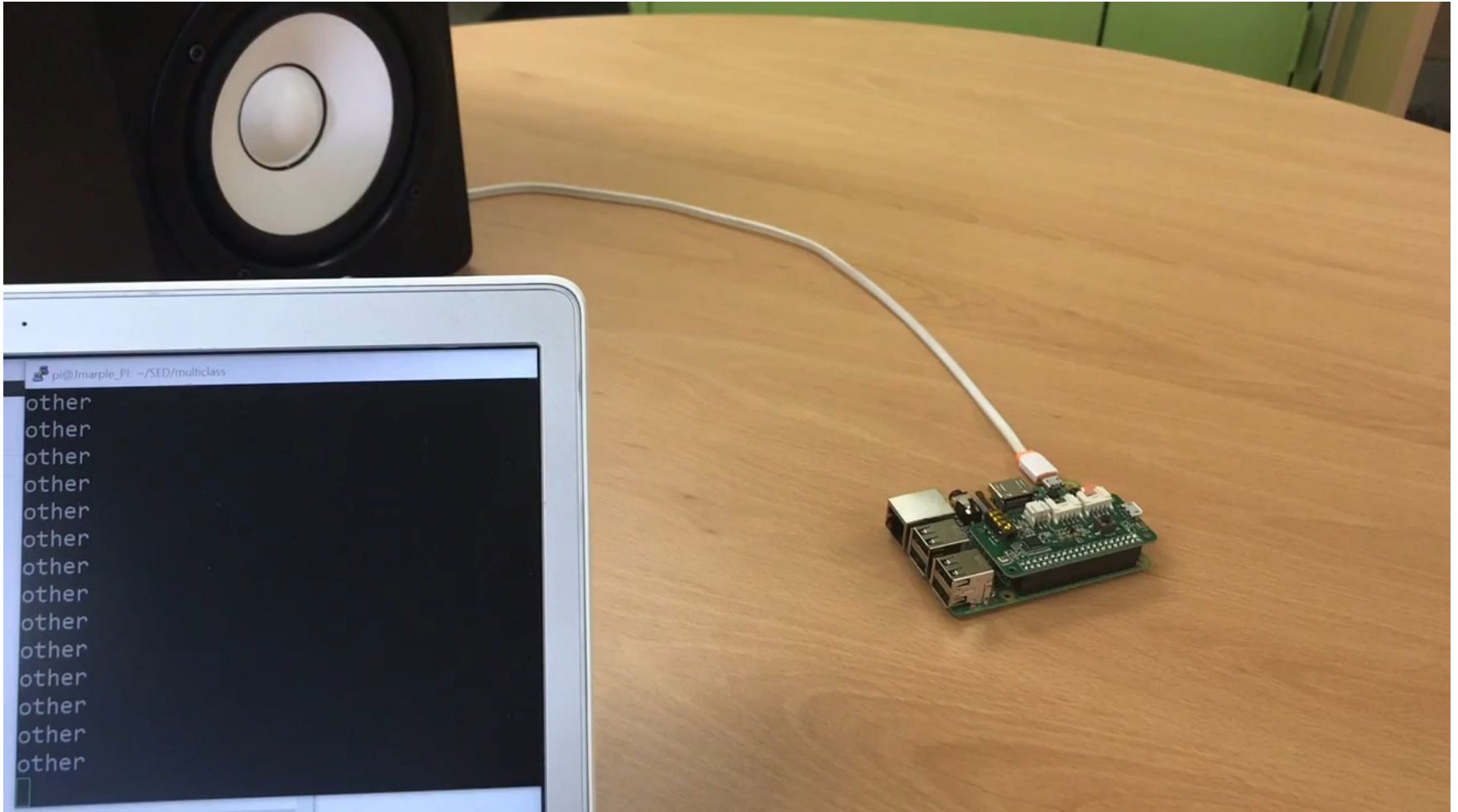
# 제이마플의 활용 방안



# 제이마플의 활용 방안

- **아기울음 감지기**
  - 아기울음을 감지해서 사용자에게 알려줌
  - 기존 제품은 소리 크기로 감지
  - 딥러닝 기술을 활용해 오탐을 크게 줄임

# 제이마플의 활용 방안

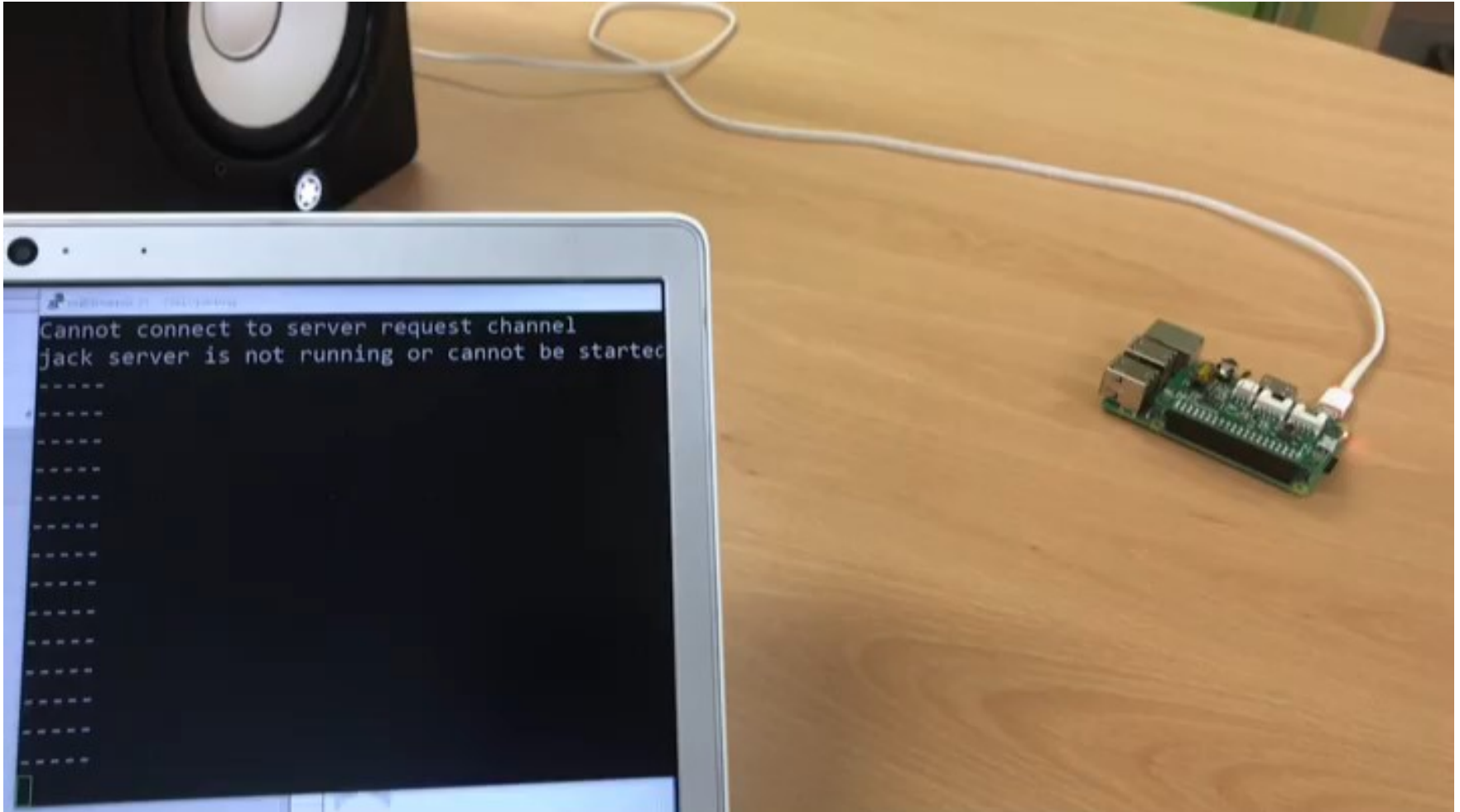




# 제이마플의 활용 방안

- **보안/안전 소리 감지 모듈**
  - 지하주차장 이상소리 감지 모듈
  - 지하주차장에서 발생하는 비명소리 등 이상소리를 감지하여 알려줌

# 제이마플의 활용 방안



# 제이마플의 활용 방안

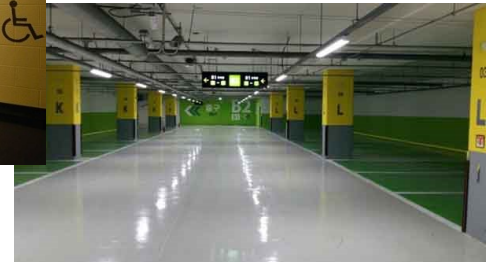


# 제이마플의 활용 방안

- 청각장애인 소리 인식기
  - 시제품 제작 중
  - 2018년 제품화 예정
- 아기울음 감지기
  - 시제품 제작 중
- 지하주차장 이상소리 감지 모듈
  - 현장 테스트 예정
- 추후, 고장 감지 등의 분야로 확대 적용해나갈 예정

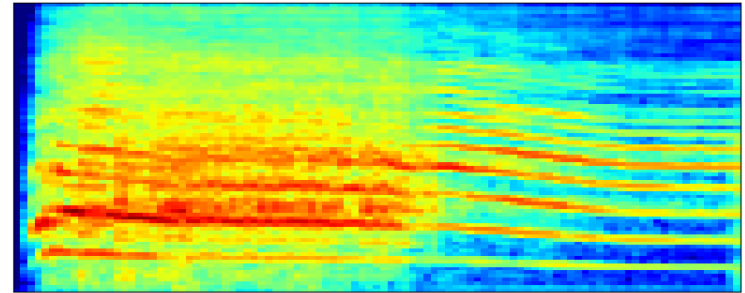
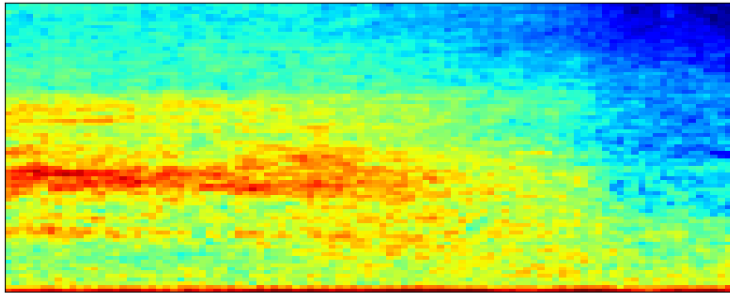
# 소리 인식의 어려운 점

- 알고리즘만큼이나 마이크도 중요
  - SNR(signal-to-noise ratio)
  - AGC(auto gain control)
- 어떤 환경에서 사용하는지 중요
  - 실내 or 실외?
  - 작은 공간 or 큰 공간?

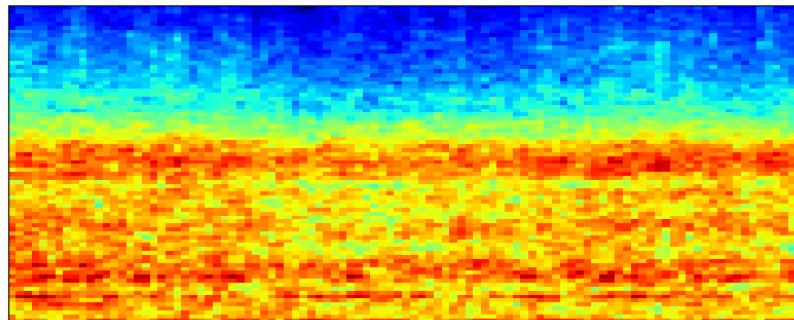


# 소리 인식의 어려운 점

- Spectrogram으로 비슷한 소리는 구분이 어려움

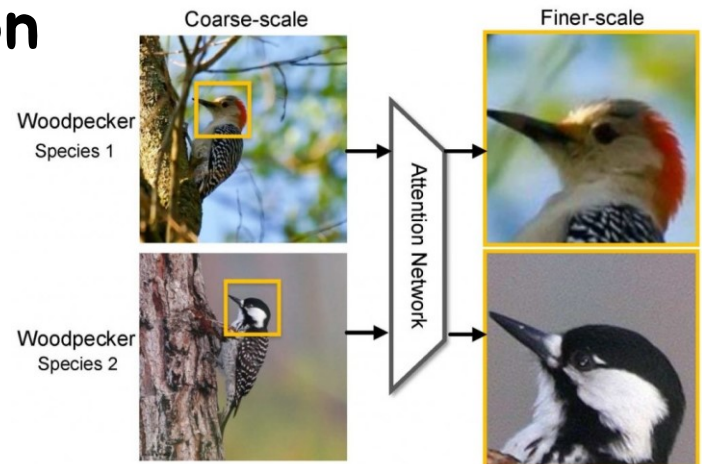


- 큰 소리 발생했을 때, 작은 소리를 감지하기 어려움



# 한계 극복 방안

- **Continual Learning**
  - 학습된 모델에 추가 학습
- **Attention**
  - 가장 특징적인 부분만 attention
  - 연산량 감소 / 정확도 향상



# 참고자료

- **DCASE 2016** (<http://www.cs.tut.fi/sgn/arg/dcase2016/index>)
- **DCASE 2017** (<http://www.cs.tut.fi/sgn/arg/dcase2017/index>)
- **non-speech acoustic event detection and classification, Tuomas Virtanen and Jort F. Gemmeke**  
(<https://sites.google.com/site/amadana0001/Home//tutorials>)
- **RARE SOUND EVENT DETECTION USING 1D CONVOLUTIONAL RECURRENT NEURAL NETWORKS, Hyungui Lim, Jeongsoo Park, Yoonchang Han**