

DMQA Open Seminar

---

# Introduction to Analysis for Sound data

---

2020 . 01 . 08

Data Mining & Quality Analytics Lab.

발표자 : 정 기원

dia517@korea.ac.kr

# 목차

---

1. Introduction
2. Sound data definition
3. Feature Engineering for sound data
4. Deep learning for sound data
5. DBCASE Challenge
6. Conclusion

# Introduction

---

## ❖ 발표자 소개



- 정기원
- 고려대학교 산업경영공학 전공
- Data Mining & Quality Analytics Lab (김성범 교수님)
- 석사 과정 (2020.03 ~ )

## ✓ 관심 연구 분야

- Machine Learning / Deep Learning
- Deep learning for signal processing and time series analysis
- Abnormal respiratory sound classification using hierarchical attention networks model
- An optimization algorithm for Hyper-path problem under travel time uncertainty

# Introduction

## ❖ Introduction to Analysis for sound data

음성 인식  
Speech Recognition



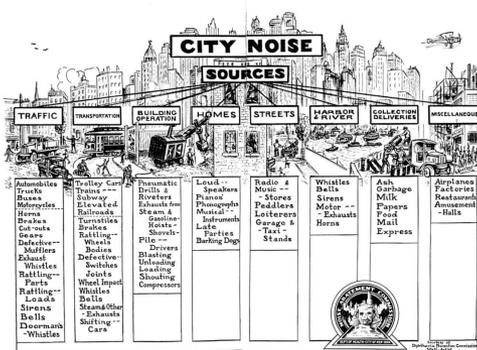
Siri

구글 '나우'

Hi, how can I help?

아마존 '알렉사'

오디오 분류  
Audio Classification



**CITY NOISE SOURCES**

TRAFFIC	TRANSPORTATION	BUILDING OPERATIONS	HOUSES	STREETS	MANUFACTURING	COLLECTION	MISCELLANEOUS
Automobiles Trucks Buses Motorcycles Horns Buses Cabs Scares Defective Machines Exhaust Whistles Rattling Pumps Loads Sirens Bells Doormat Whistles	Trolley Cars Trains Elevated Subways Turbines Cables Bells Defective Machines Exhaust Whistles Rattling Pumps Loads Sirens Bells Doormat Whistles	Drills Gubery Elevated Subways Turbines Cables Bells Defective Machines Exhaust Whistles Rattling Pumps Loads Sirens Bells Doormat Whistles	Drills Gubery Elevated Subways Turbines Cables Bells Defective Machines Exhaust Whistles Rattling Pumps Loads Sirens Bells Doormat Whistles	Drills Gubery Elevated Subways Turbines Cables Bells Defective Machines Exhaust Whistles Rattling Pumps Loads Sirens Bells Doormat Whistles	Drills Gubery Elevated Subways Turbines Cables Bells Defective Machines Exhaust Whistles Rattling Pumps Loads Sirens Bells Doormat Whistles	Drills Gubery Elevated Subways Turbines Cables Bells Defective Machines Exhaust Whistles Rattling Pumps Loads Sirens Bells Doormat Whistles	Drills Gubery Elevated Subways Turbines Cables Bells Defective Machines Exhaust Whistles Rattling Pumps Loads Sirens Bells Doormat Whistles

도시에서 발생하는 소음 종류

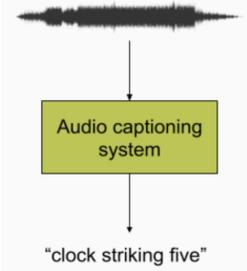
오디오 캡셔닝  
Audio Captioning



When's the fairy princess coming?

iMore

유튜브 자막 생성 기능



Audio captioning system

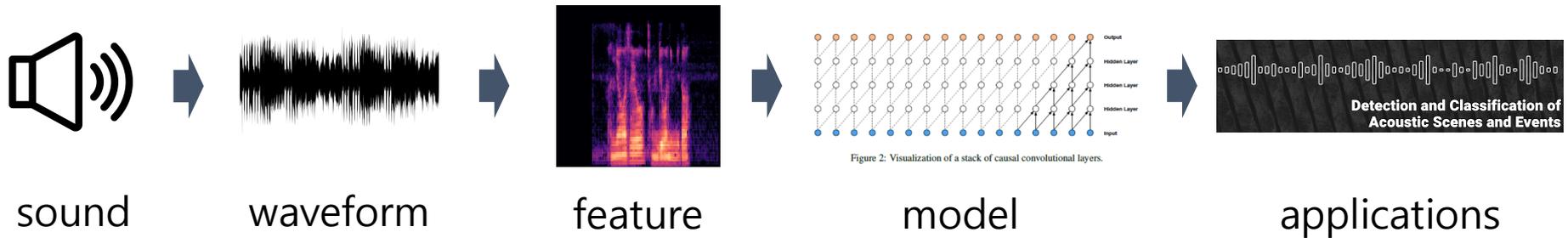
"clock striking five"

➡ 다양한 소리 분석 분야에서 인공지능 기술을 적용하는 사례가 증가하고 있다.

# Introduction

## ❖ Introduction to Analysis for sound data

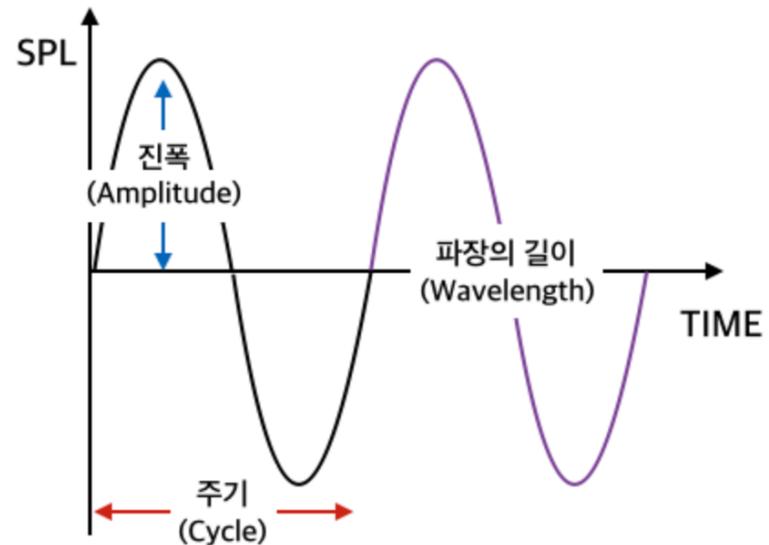
- 소리 데이터 개념 및 원리
- 데이터 전처리 및 특징 추출 기법
- 소리 데이터에 적합한 데이터 증강 기법 및 딥러닝 모델
- 소리 분석 활용 사례



# Sound data

## ❖ 소리 Sound

- 공기나 물 같은 매질의 진동을 통해 전달되는 종파
- 사람의 귀에 들려오는 소리는 공기 속을 전해오는 파동
- 소리의 3요소
  - 세기(소리의 크기) / 높낮이(소리의 높고 낮음) / 음색(소리의 색상)
- 소리의 주파수(Hz) 원리
  - 낮은 음일수록 주파수가 낮다.
  - 높은 음일수록 주파수가 높다.

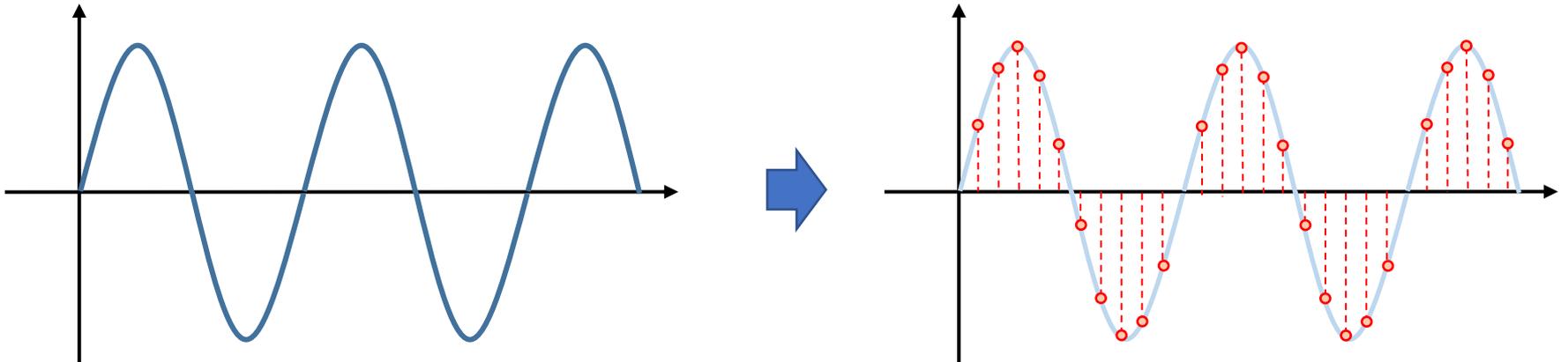


[ 주파수의 3요소 ]

# Sound data

## ❖ 소리의 샘플링 레이트(Sampling rate) 과정

- 소리를 컴퓨터에 입력시키기 위해 음파를 숫자로 표현할 필요가 있음
- **샘플링 레이트(Sampling Rate)**는 **우리가 듣는 소리**(아날로그 신호)을 **숫자화**(디지털로 변환) 시킨 것
- 표준 Sampling rate는 44.1khz
  - 소리로부터 1초당 44100개의 샘플을 추출했다는 것을 의미



# Sound data

## ❖ 소리의 샘플링 레이트(Sampling rate) 과정

- 소리를 컴퓨터에 입력시키기 위해 음파를 숫자로 표현할 필요가 있음
- **샘플링 레이트(Sampling Rate)**는 **우리가 듣는 소리**(아날로그 신호)을 **숫자화**(디지털로 변환) 시킨 것
- 표준 Sampling rate는 44.1khz
  - 소리로부터 1초당 44100개의 샘플을 추출했다는 것을 의미



1 Second

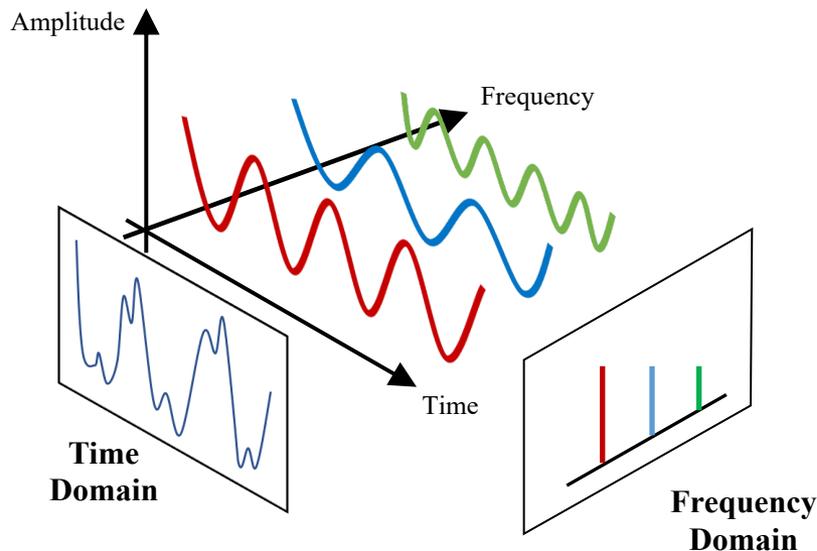


출처 : <https://deepmind.com/blog/article/wavenet-generative-model-raw-audio>

# Feature Engineering

## ❖ 특징 추출 (Feature extraction)

- 파동은 시간 영역(Time domain)과 주파수 영역(Frequency domain)로 구성
- 소리 특성 정보를 추출하기 위한 다양한 특징 추출 방법론이 존재함
  - 스펙트럼(Spectrum) / 스펙트로그램(Spectrogram) / 스칼로그램(Scalogram) / MFCC ...



소리의 시간 영역/주파수 영역

**Spectrum**

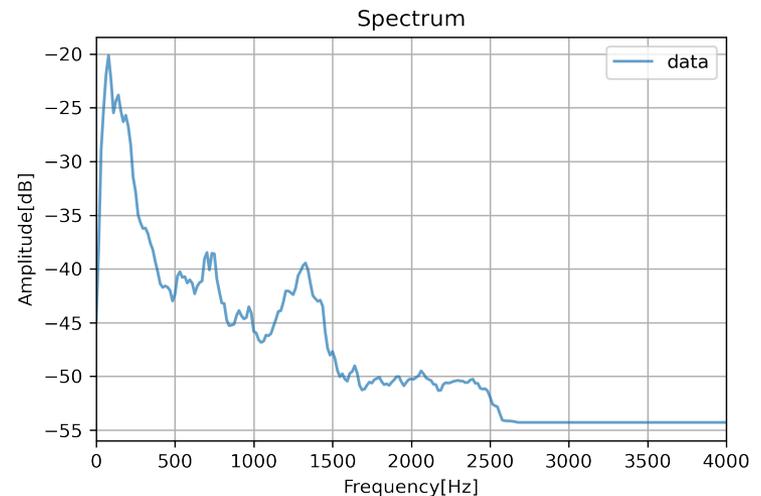
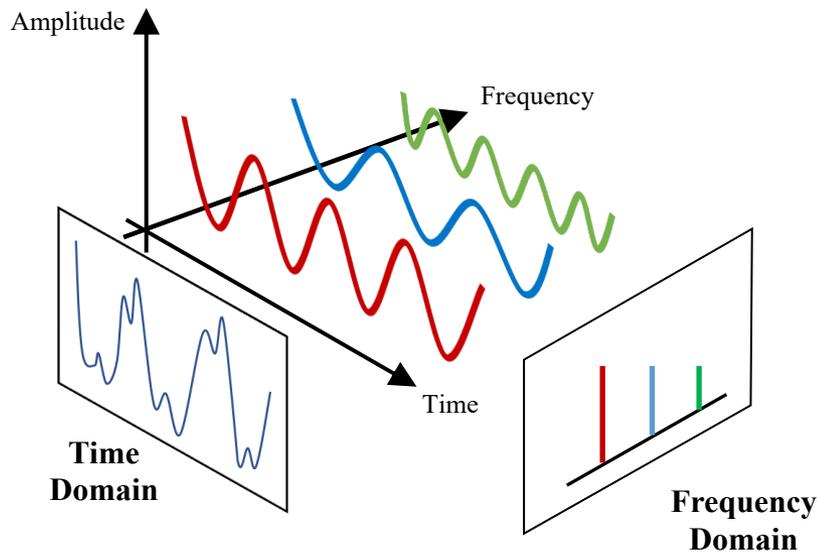
**Mel Spectrogram**

**MFCC**

# Feature Engineering

## ❖ 스펙트럼 (Spectrum)

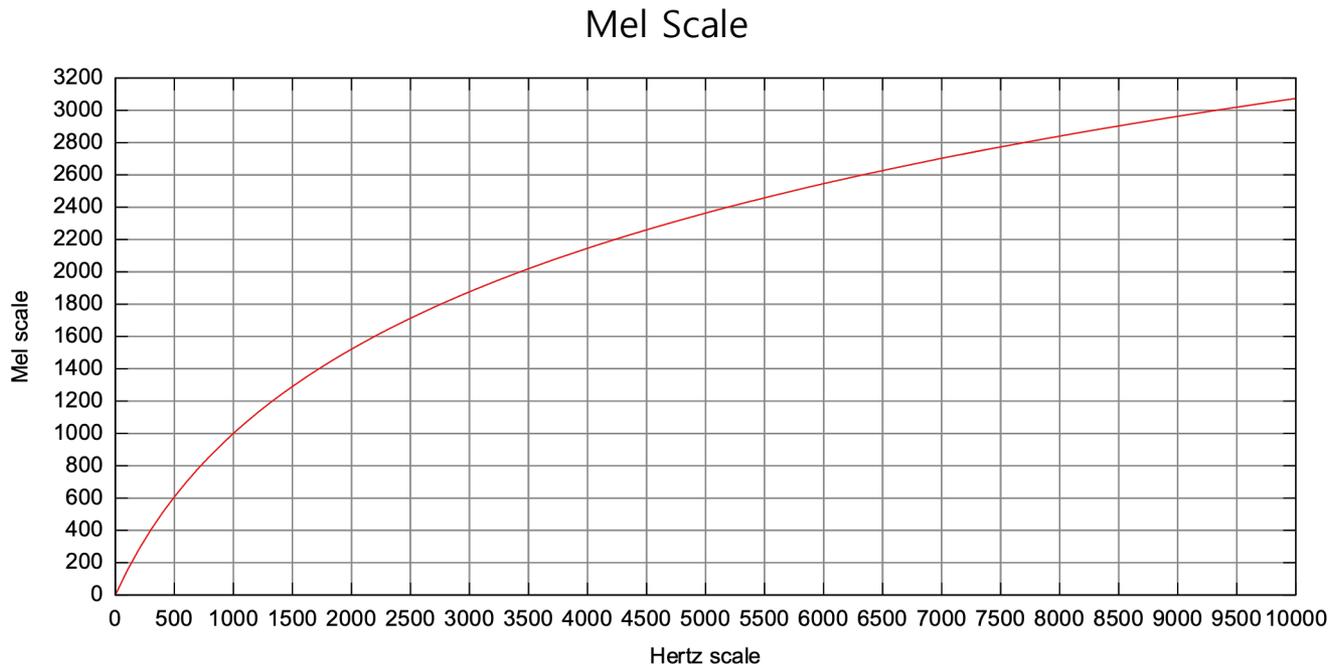
- 파동의 시간 영역(Time domain)을 주파수 영역(Frequency domain)로 변환
- 음향 신호를 주파수, 진폭으로 분석하여 보여줌
- 고속 푸리에 변환(Fast Fourier transform)을 적용
- x 축은 주파수(Frequency), y 축은 진폭(Amplitude)을 나타냄



# Feature Engineering

## ❖ 멜 스펙트로그램 (Mel Spectrogram)

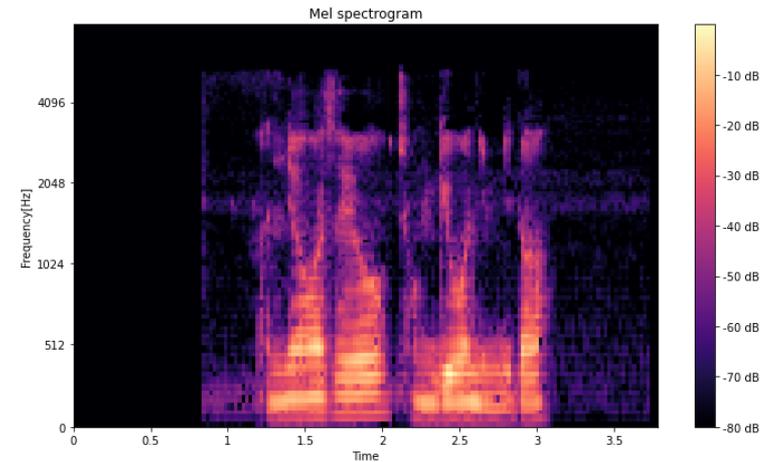
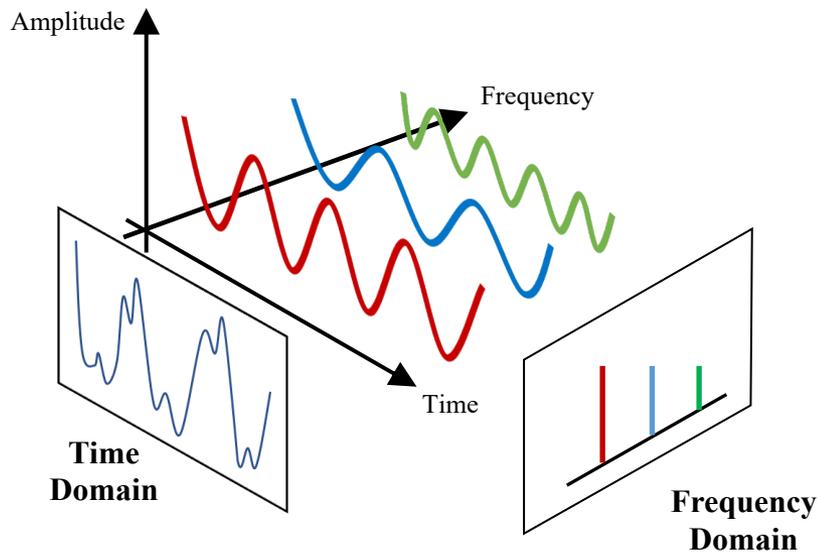
- 주파수 특성이 시간에 따라 달라지는 오디오를 분석하기 위한 특징 추출 기법
- 인간의 청각 영역을 반영한 mel scale을 적용함
  - 보통 고주파로 갈수록 사람이 구분하는 주파수 간격이 넓어지는데 이를 반영해줌



# Feature Engineering

## ❖ 멜 스펙트로그램 (Mel Spectrogram)

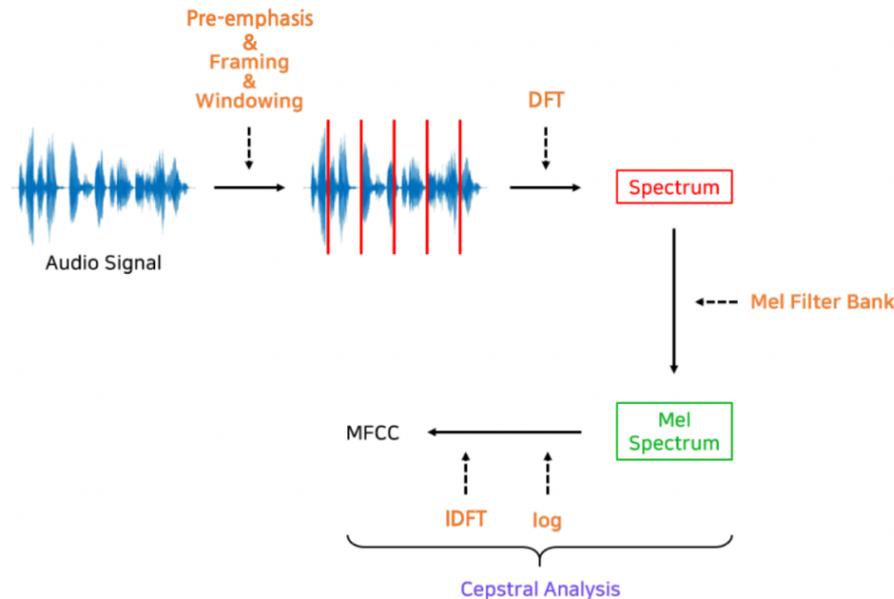
- 주파수 특성이 시간에 따라 달라지는 오디오를 분석하기 위한 특징 추출 기법
- 인간의 청각 영역을 반영한 mel scale을 적용함
  - 보통 고주파로 갈수록 사람이 구분하는 주파수 간격이 넓어지는데 이를 반영해줌
- x 축은 시간(Time), y 축은 주파수(Frequency), z 축은 진폭(Amplitude)을 나타냄



# Feature Engineering

## ❖ MFCC (Mel-Frequency Cepstral Coefficient)

- 멜 스펙트럼(Mel Spectrum)에서 켈스트럴(Cepstral)분석을 통해 추출된 값
  - Cepstral은 스펙트럼에서 배음 구조를 유추할 수 있도록 도와주는 분석
- 로그 멜 스펙트럼에 역푸리에변환(Inverse Fourier Transform)을 적용
  - 주파수 정보의 상관 관계가 높은 문제를 해소

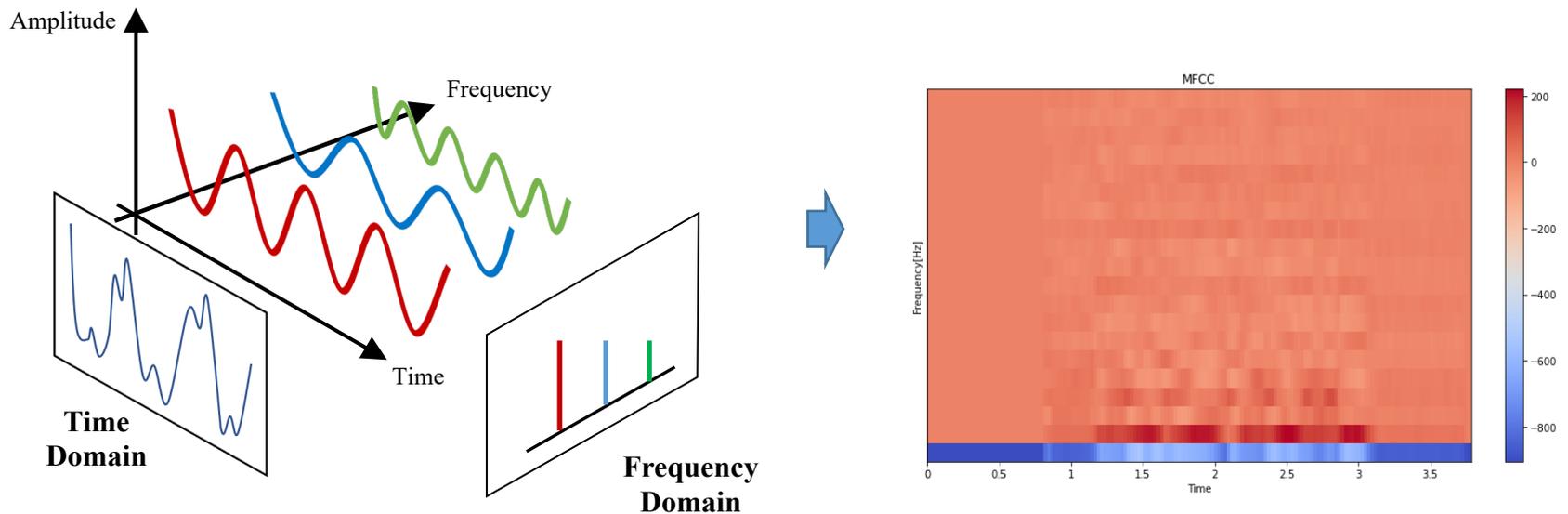


[ MFCC Process ]

# Feature Engineering

## ❖ MFCC (Mel-Frequency Cepstral Coefficient)

- 멜 스펙트럼(Mel Spectrum)에서 켈스트럴(Cepstral)분석을 통해 추출된 값
  - Cepstral은 스펙트럼에서 배음 구조를 유추할 수 있도록 도와주는 분석
- 로그 멜 스펙트럼에 역푸리에변환(Inverse Fourier Transform)을 적용
  - 주파수 정보의 상관 관계가 높은 문제를 해소

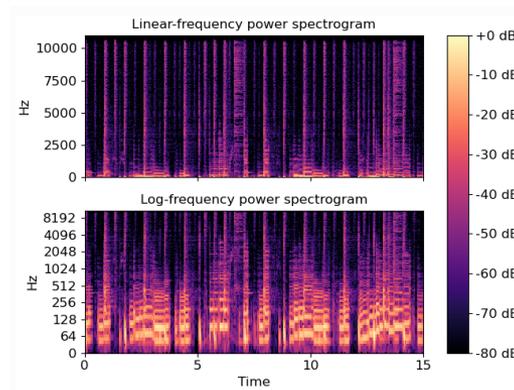
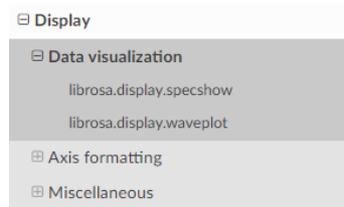


# Feature Engineering

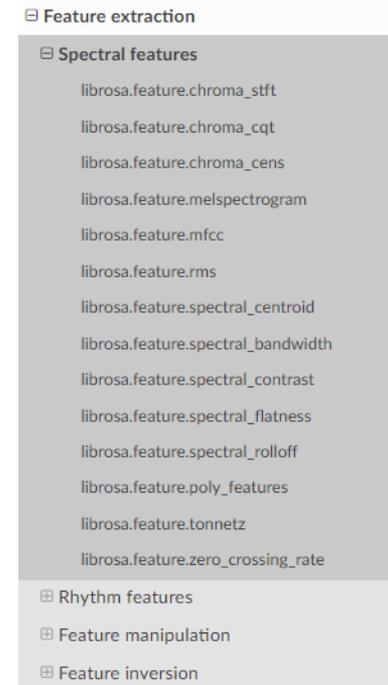
## ❖ librosa 패키지

- 오디오 신호를 분석하는 Python 모듈
- 오디오 데이터 입.출력 및 다양한 특징 추출 방법론을 쉽게 사용할 수 있음
- 오디오 데이터로 추출된 특징의 시각화 기능 제공

[ Visualization ]



[ Feature extraction ]

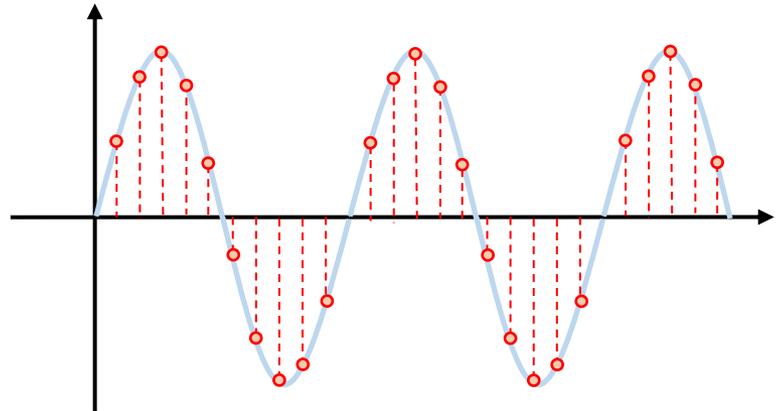


Librosa module document : <https://librosa.org/doc/latest/index.html>

# Feature Engineering

## ❖ 추출된 특징 사용

- 고차원 특징이기에 머신러닝 모델 적용을 위해 기술통계량을 사용
  - 평균, 절사평균, 중위수, 최빈수, 최솟값, 최댓값, 분산, 표준편차, 절대편차 ...
- 최근 딥러닝 모델에는 특징 고유의 값이나 히트맵을 이미지로 변환하여 사용



“ 안녕하세요, 정기원 입니다. ”

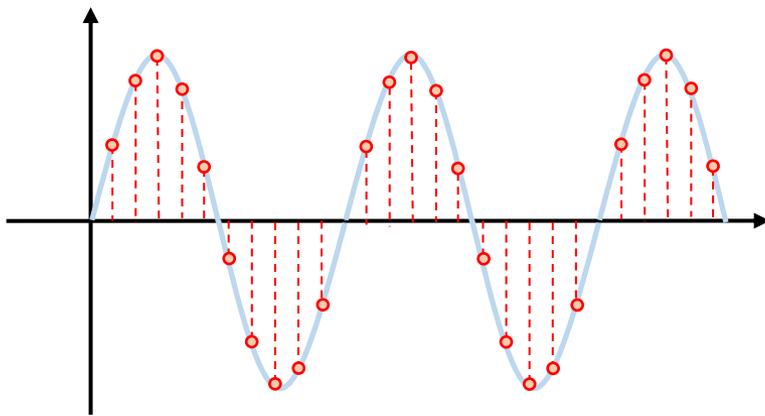
약 3초 길이의 오디오 파일

크기 : (83173, )

# Feature Engineering

## ❖ 추출된 특징 사용

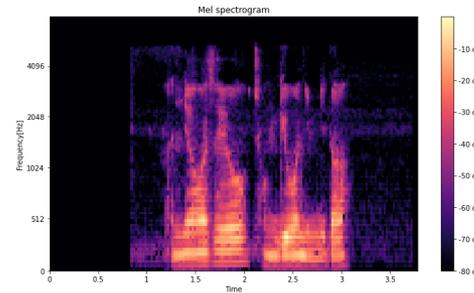
- 고차원 특징이기에 머신러닝 모델 적용을 위해 기술통계량을 사용
  - 평균, 절사평균, 중위수, 최빈수, 최솟값, 최댓값, 분산, 표준편차, 절대편차 ...
- 최근 딥러닝 모델에는 특징 고유의 값이나 히트맵을 이미지로 변환하여 사용



크기 : (83173, )

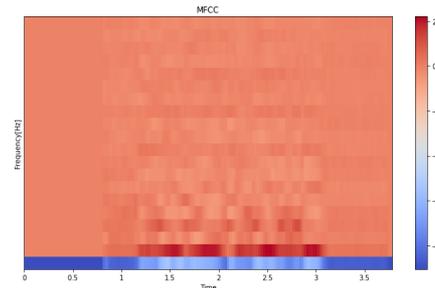


## Mel Spectrogram



크기 : (128,163)

## MFCC

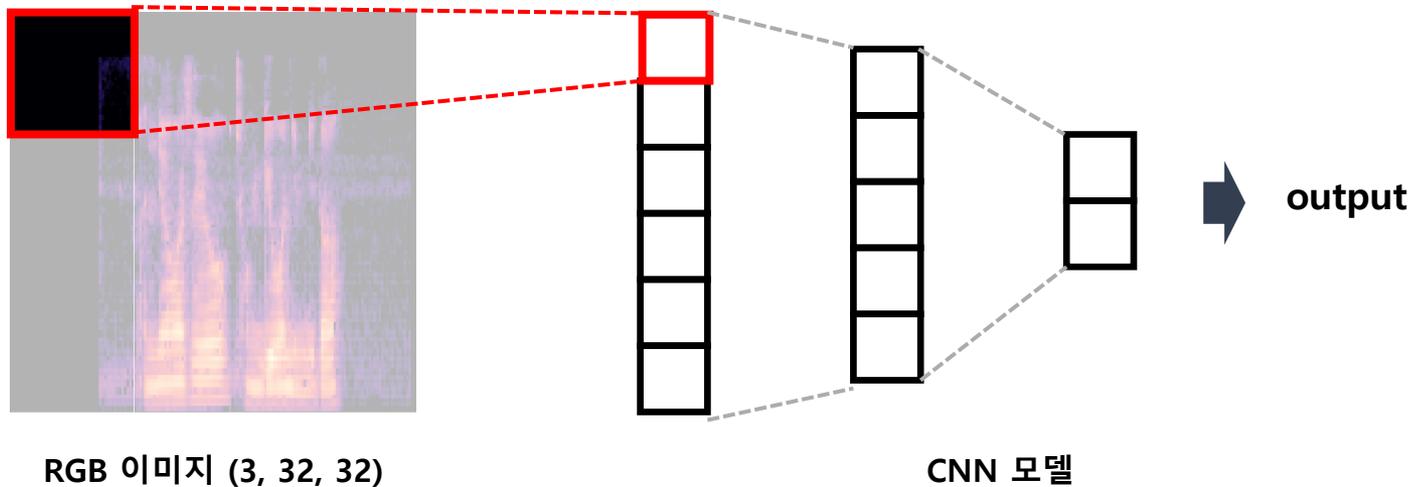


크기 : (20,163)

# Feature Engineering

## ❖ 추출된 특징 사용

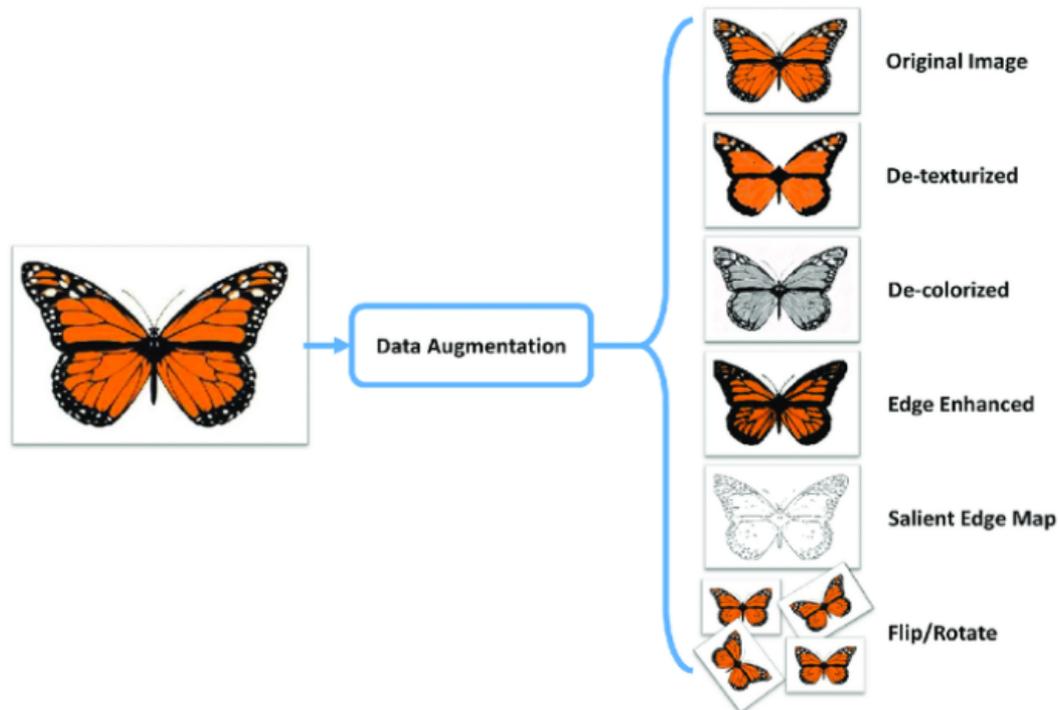
- 고차원 특징이기에 머신러닝 모델 적용을 위해 기술통계량을 사용
  - 평균, 절사평균, 중위수, 최빈수, 최솟값, 최댓값, 분산, 표준편차, 절대편차 ...
- 최근 딥러닝 모델에는 특징 고유의 값이나 히트맵을 이미지로 변환하여 사용



# Data Augmentation

## ❖ 데이터 증강 기법(Data Augmentation)

- 일반적으로 레이블이 존재하는 데이터에 변화를 주어 원본 데이터와 같은 레이블을 갖는 새로운 데이터를 만드는 기법
  - 좌우 반전 / 이미지 자르기 / 이미지 회전 / 밝기 조절 등



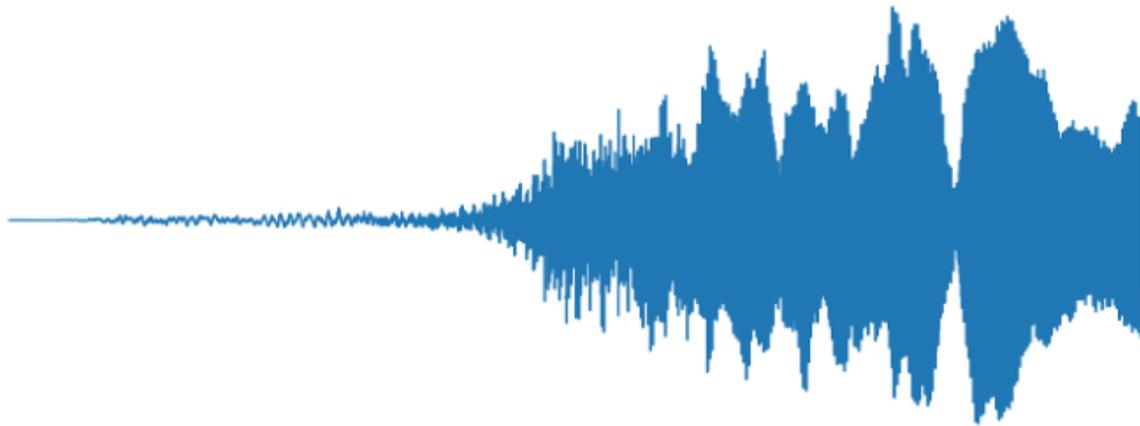
이미지 출처 : <https://medium.com/secure-and-private-ai-writing-challenge/data-augmentation-increases-accuracy-of-your-model-but-how-aa1913468722>

# Data Augmentation

---

## ❖ 소리 데이터에 적합한 데이터 증강 기법(Data Augmentation)

- Adding noise : white noise를 추가하여 원본 오디오에 잡음을 생성
- Shifting : 원본 오디오 데이터를 좌우로 이동
- Stretching : 원본 오디오 데이터의 빠르기를 조정



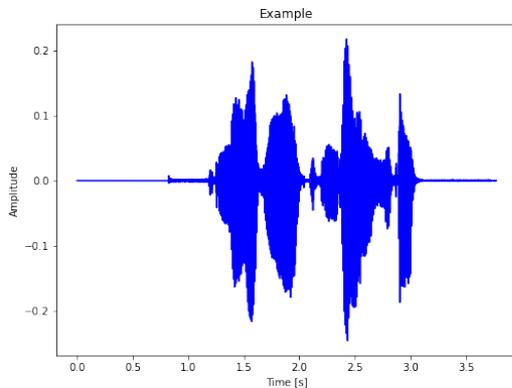
# Data Augmentation

## ❖ 소리 데이터에 적합한 데이터 증강 기법(Data Augmentation)

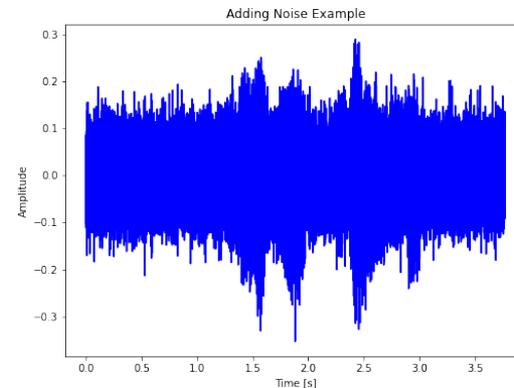
- **Adding noise : white noise**를 추가하여 원본 오디오에 잡음을 생성
- Shifting : 원본 오디오 데이터를 좌우로 이동
- Stretching : 원본 오디오 데이터의 빠르기를 조정



“ 안녕하세요, 정기원 입니다. ”



Adding noise



# Data Augmentation

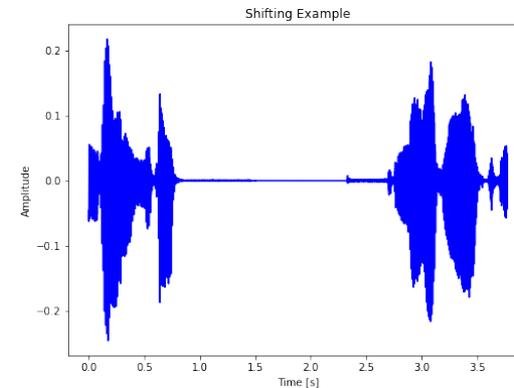
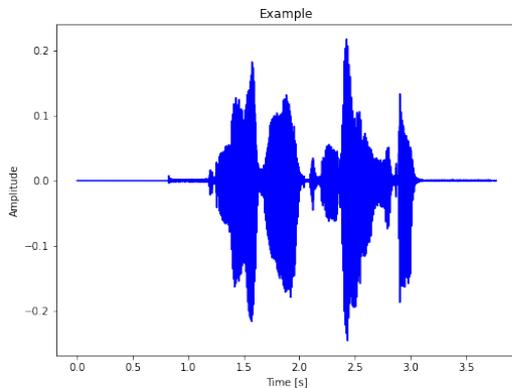
## ❖ 소리 데이터에 적합한 데이터 증강 기법(Data Augmentation)

- Adding noise : white noise를 추가하여 원본 오디오에 잡음을 생성
- **Shifting** : 원본 오디오 데이터를 좌우로 이동
- Stretching : 원본 오디오 데이터의 빠르기를 조정



“ 안녕하세요, 정기원 입니다. ”

Shifting



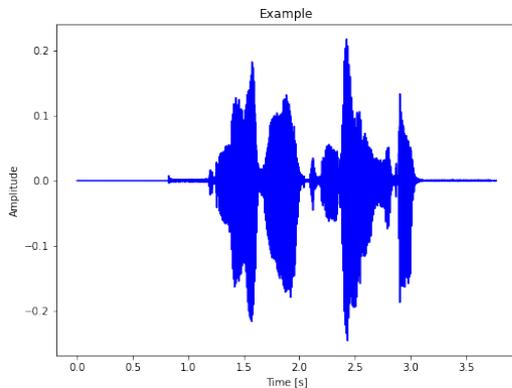
# Data Augmentation

## ❖ 소리 데이터에 적합한 데이터 증강 기법(Data Augmentation)

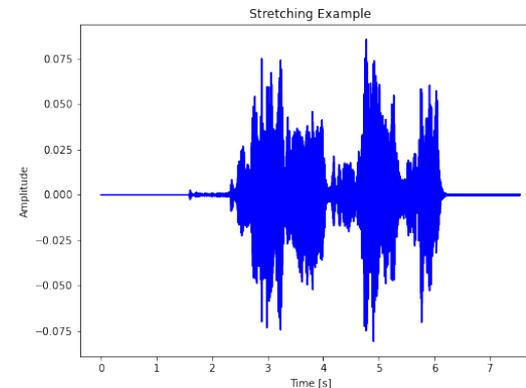
- Adding noise : white noise를 추가하여 원본 오디오에 잡음을 생성
- Shifting : 원본 오디오 데이터를 좌우로 이동
- **Stretching** : 원본 오디오 데이터의 빠르기를 조정



“ 안녕하세요, 정기원 입니다. ”



Stretching



# Deep Learning model

## ❖ 소리 데이터를 위한 딥러닝 모델

- 2016년 구글 딥마인드에서 오디오 생성 모델인 wavenet 공개
- 텍스트를 음성으로 변환하는 Text-to-Speech(TTS) 수행하기에 적합한 딥러닝 모델
  - 당시 TTS 모델은 많은 양의 데이터를 필요로 하고 부자연스러운 음성이 생성

### Wavenet: A generative model for raw audio

[A Oord](#), [S Dieleman](#), [H Zen](#), [K Simonyan](#)... - arXiv preprint arXiv ..., 2016 - arxiv.org

This paper introduces **WaveNet**, a deep neural network for generating raw audio waveforms. The model is fully probabilistic and autoregressive, with the predictive distribution for each audio sample conditioned on all previous ones; nonetheless we show that it can be ...

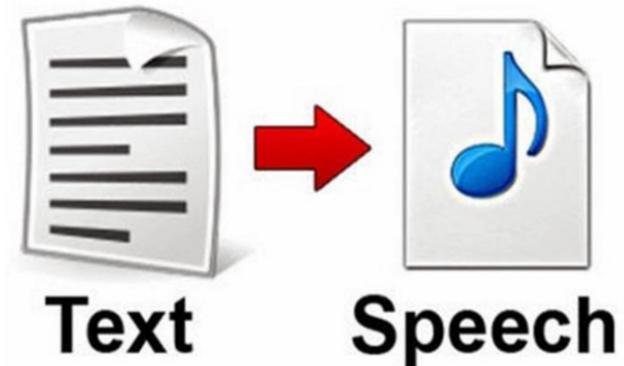
☆ 99 2462회 인용 관련 학술자료 전체 12개의 버전 88 52 code implementations

## WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

Aäron van den Oord	Sander Dieleman	Heiga Zen <sup>†</sup>
Karen Simonyan	Oriol Vinyals	Alex Graves
Nal Kalchbrenner	Andrew Senior	Koray Kavukcuoglu

{avdnoord, sedielem, heigazen, simonyan, vinyals, graves, nalk, andrewsenior, korayk}@google.com  
Google DeepMind, London, UK

<sup>†</sup> Google, London, UK



# Deep Learning model

## ❖ Wavenet

- 오디오의 파형 형태를 직접 사용해서 새로운 파형을 생성하는 확률론적 모델
- 30개의 residual block을 쌓은 형태의 구조를 보임
- 주요 특징 3가지
  - 음성 파형 학습을 위한 새로운 구조를 제시함
  - 조건부 모델링을 이용해 특징적인 음성을 생성할 수 있음
  - 오디오 파형만을 이용해 자연스럽게 새로운 음성 파형을 생성할 수 있음

[ Wavenet 모델링 ]

$$p(x) = \prod_{t=1}^T p(x_t | x_1, x_2, \dots, x_{t-1})$$

특정  $t$  시점에서 파형  $x_t$ 가 생성될 확률 분포

[ Wavenet 전체 구조 ]

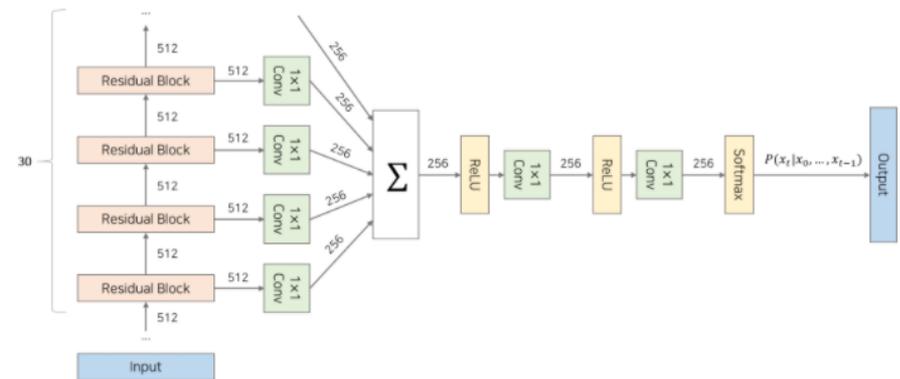
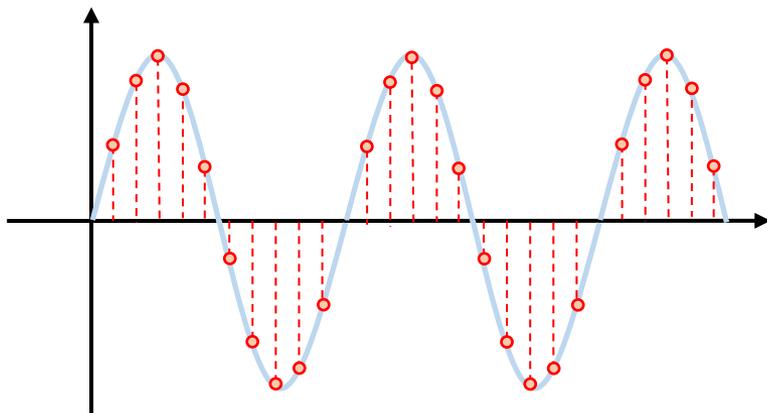


그림 출처 : <https://www.youtube.com/watch?v=nsrSrYtKkT8>

# Wavenet

## ❖ SoftMax 함수

- Wavenet 입/출력은 아날로그 음성 데이터를 변환한 디지털 데이터 값
- 오디오는 16 비트의 정수 값으로 저장해서 사용하기에 확률론적 모델링이 힘들
  - 매  $t$  시점마다  $2^{15} \sim 2^{15} + 1$  사이의 숫자가 나옴
  - $t$  시점에 특정 파형이 나올 확률을 계산하는데 총 65,536개의 확률을 고려
- $\mu - law$  companding 변환을 통해 작은 정수 범위로 총 256개의 확률을 고려



$2^{15} \sim 2^{15} + 1$  사이의 정수  
→ 65536개의 확률



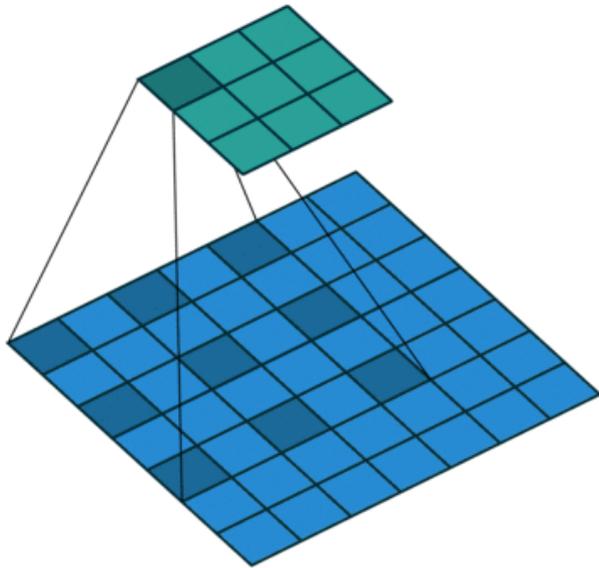
$$f(x_t) = \text{sign}(x_t) \frac{\ln(1 + \mu|x_t|)}{\ln(1 + \mu)}$$

-127 ~ 128 사이의 정수  
→ 256개의 확률

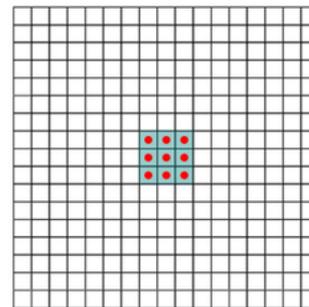
# Wavenet

## ❖ Dilated Causal Convolutions

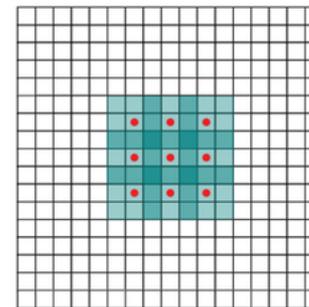
- Dilated convolution과 Causal convolution 개념을 결합한 컨볼루션 연산
- Dilated convolutions – 필터에 zero padding을 추가해 모델의 receptive field를 늘려줌
  - receptive field란 필터가 한 번에 볼 수 있는 데이터를 탐색할 수 있는 영역
  - 입력된 데이터의 특징을 잡아내기 위해서는 receptive field는 높으면 높을수록 좋음



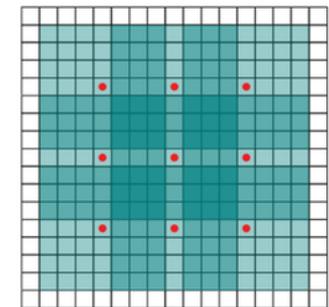
[ receptive field size ]



1-dilated convolution



2-dilated convolution



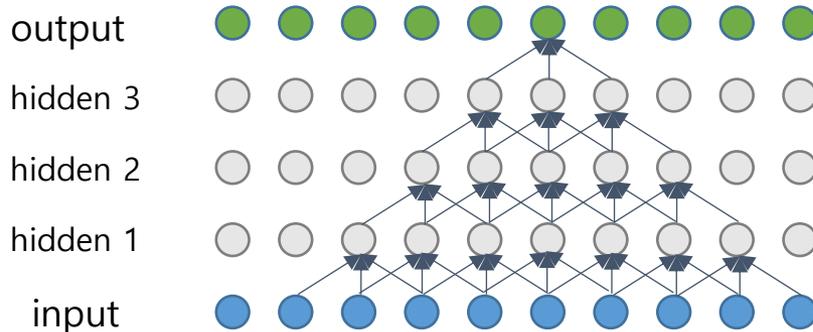
4-dilated convolution

# Wavenet

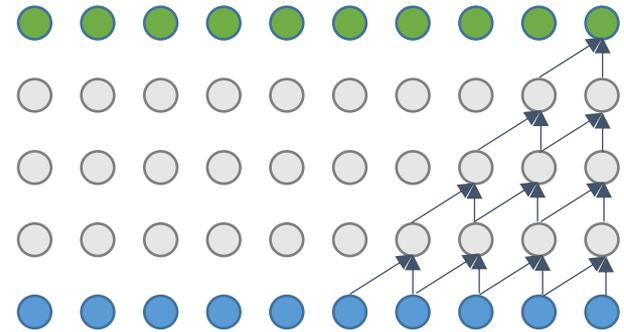
## ❖ Dilated Causal Convolutions

- Causal convolutions – 시간 순서를 고려하여 필터를 적용하는 컨볼루션 연산
  - RNN 계열의 모델처럼 시계열 데이터를 모델링할 수 있음
  - Receptive field를 넓히기 위해서 많은 양의 레이어를 쌓아야 한다는 단점이 존재

[ Standard Convolutions ]



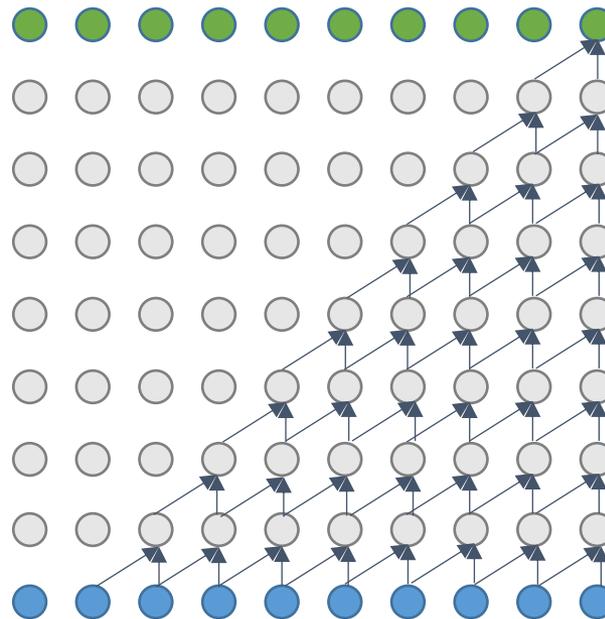
[ Causal Convolutions ]



## ❖ Dilated Causal Convolutions

- Causal convolutions – 시간 순서를 고려하여 필터를 적용하는 컨볼루션 연산
  - RNN 계열의 모델처럼 시계열 데이터를 모델링할 수 있음
  - Receptive field를 넓히기 위해서 많은 양의 레이어를 쌓아야 한다는 단점이 존재

[ Causal Convolutions ]

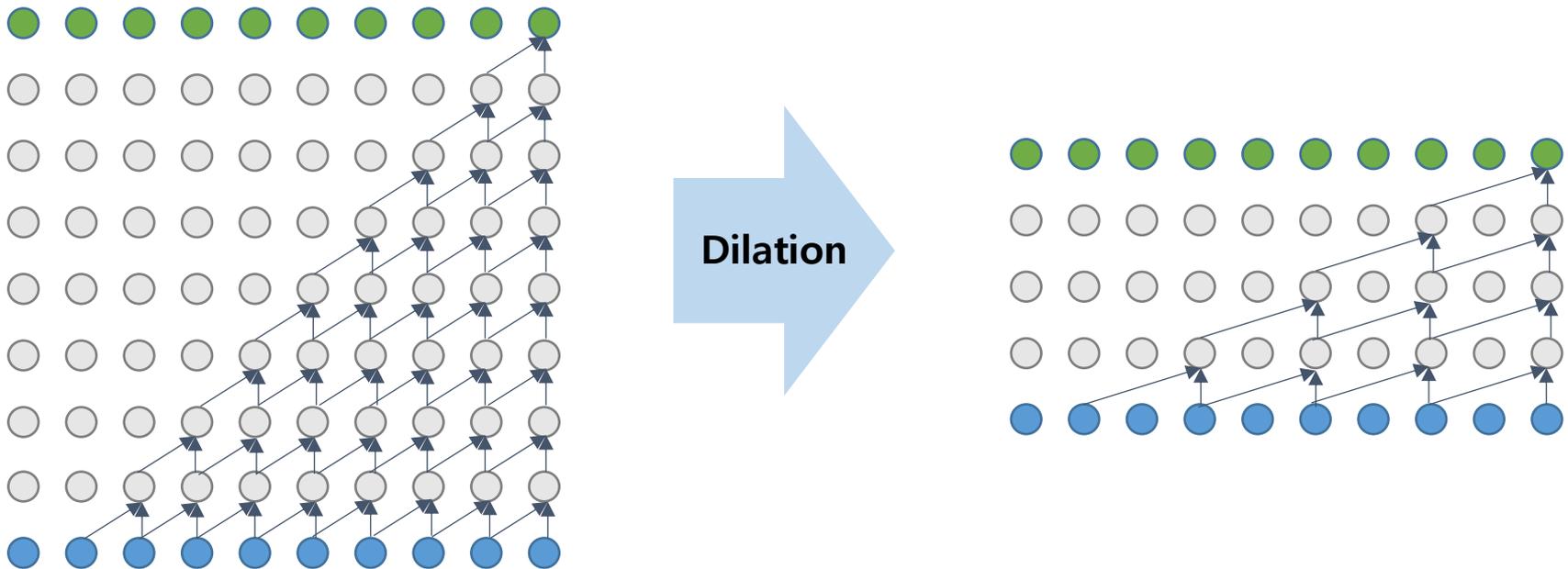


# Wavenet

## ❖ Dilated Causal Convolutions

- Dilation기법을 이용해 일정 스텝을 건너뛰며 필터를 적용하는 dilated causal convolution
- 적은 층의 레이어로 receptive field를 넓힐 수 있는 효과를 가져옴

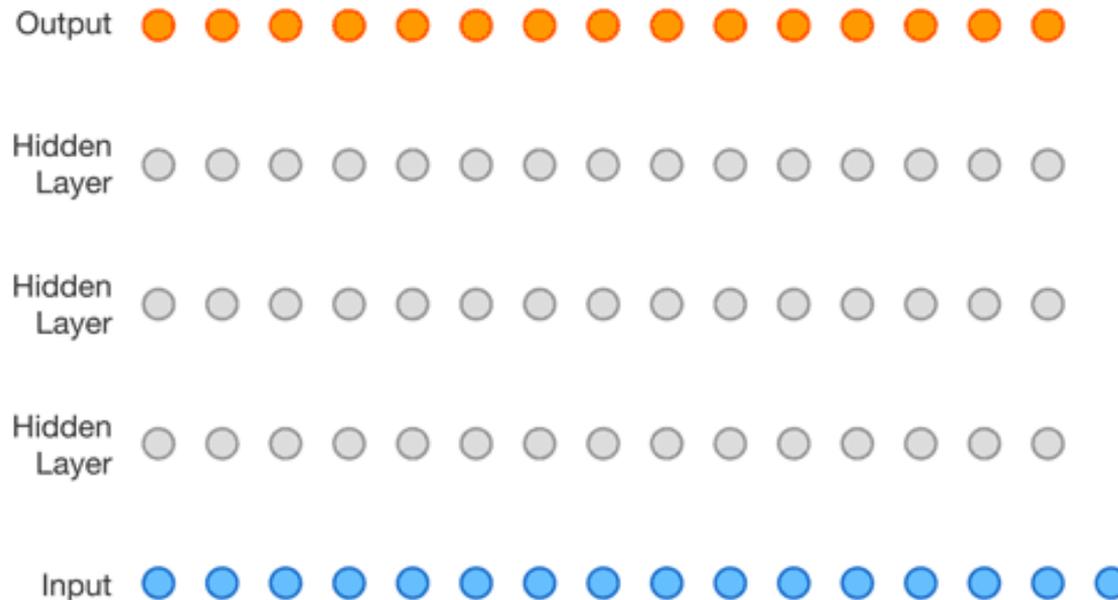
[ Causal Convolutions ]



# Wavenet

## ❖ Dilated Causal Convolutions

- Dilation기법을 이용해 일정 스텝을 건너뛰며 필터를 적용하는 dilated causal convolution
- 적은 층의 레이어로 receptive field를 넓힐 수 있는 효과를 가져옴
- 논문에서는 dilation 크기를 2배씩 증가시키고 이를 반복
  - 1, 2, 4, ..., 512, 1, 2, 4, ..., 512



출처 : <https://deepmind.com/blog/article/wavenet-generative-model-raw-audio>

# Wavenet

## ❖ Conditional wavenet

- 확률 모델에 조건 정보를 추가함으로써 특정한 성질을 가진 오디오를 생성할 수 있음
- 조건부 모델링 방법 2가지
  - 전역적 조건 : 시점 변화에 영향을 받지 않는 정보를 추가
  - 지역적 조건 : 시점 변화에 영향을 받는 정보를 추가

$$p(x) = \prod_{t=1}^T p(x_t | x_1, x_2, \dots, x_{t-1}, h)$$

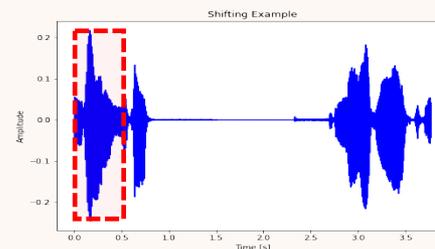
### 전역적 조건 (Global Conditioning)



" 안녕하세요~ "  
" 아이고 배고프다 ... "  
" 반갑습니다! "

ex ) 화자 고유의 음성을 생성

### 지역적 조건 (Local Conditioning)



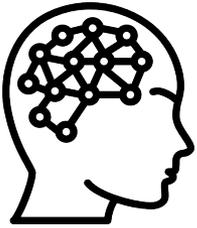
" **아이고** 배고프다 ... "

ex ) 특정 텍스트에 맞는 음성을 생성

# Wavenet

## ❖ Generated Audio by wavenet

- 기존 TTS 방법론의 부자연스러운 음성과 달리 자연스러운 음성을 생성
- 화자의 음성 정보를 조건부 정보로 활용하여 다양한 음색의 음성을 생성



parametric



concatenative



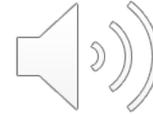
wavenet



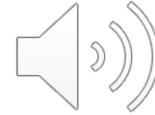
Speaker 1



Speaker 2



Speaker 3



Speaker 4

## ❖ WaveNet seminar in DMQA

- 링크 주소 : <http://dmqm.korea.ac.kr/activity/seminar/242>

### WaveNet: A Generative Model for Raw Audio

2018년 9월 16일 오후 3:50 / 조회수: 1582

#### REFERENCES

20181005\_WaveNet.pdf

#### INFORMATION

2018년 10월 5일 오후 1시 30분 ~ 고려대학교 신공학관 218호

발표자:  **곽민구**

#### TOPIC

WaveNet: A Generative Model for Raw Audio

2018-10-01 Seminar  
WaveNet: A Generative Model for Raw Audio

3 Experiments

### 03 | 2. Text-to-Speech

- 실험 세팅
  - 사용 언어: North American English & Mandarin Chinese
  - Single speaker
  - Locally conditioned WaveNet
- 평가 모델
  - Statistical parametric: LSTM-RNN
  - Concatenative: HMM-driven
  - WaveNet (L): linguistic feature
  - WaveNet (L+F): linguistic feature +  $\log F_0$ 
    - logarithmic fundamental frequency: linguistic feature로부터 값을 얻는 모델이 추가적으로 존재

2018-10-01 Seminar  
WaveNet: A Generative Model for Raw Audio

2 WaveNet

### 02 | WaveNet의 주요 포인트

- Autoregressive Model
  - Dilated causal convolutions
  - Output - categorical/softmax distribution
  - Gated activation units
  - Residual and skip connections
  - Conditional WaveNet
- WaveNet
- Input: audio → Output: audio
- Input: audio & text information → Output: audio (Text-to-Speech)

2018-10-01 Seminar  
WaveNet: A Generative Model for Raw Audio

3 Experiments

### 03 | 2. Text-to-Speech

- 평가 결과
  - MOS test

Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
<b>WaveNet (L+F)</b>	<b>4.21 ± 0.081</b>	<b>4.08 ± 0.085</b>
Natural (8-bit $\mu$ -law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

- WaveNet이 다른 모델들에 비해 높은 점수를 얻었으며, 실제 음성과 비교해서도 좋은 점수를 얻었다

# Deep Learning model

## ❖ Feature & CNN Model for audio classification

- Feature – **Mel spectrogram** / **MFCC** 를 사용할 것을 권장
- CNN Model – **Resnet** 모델을 사용할 것을 권장

### [ CNN 모델 별 오디오 분류 성능 비교 ]

#### CNN architectures for large-scale audio classification

[S Hershey](#), [S Chaudhuri](#), [DPW Ellis](#)... - ... on acoustics, speech ..., 2017 - [ieeexplore.ieee.org](#)

Convolutional Neural Networks (CNNs) have proven very effective in image classification and show promise for audio. We use various CNN architectures to classify the soundtracks of a dataset of 70M training videos (5.24 million hours) with 30,871 video-level labels. We examine fully connected Deep Neural Networks (DNNs), AlexNet [1], VGG [2], Inception [3], and ResNet [4]. We investigate varying the size of both training set and label vocabulary, finding that analogs of the CNNs used in image classification do well on our audio ...

☆ 99 934회 인용 관련 학술자료 전체 8개의 버전 6 code implementations

**Table 2:** Comparison of performance of several DNN architectures trained on 70M videos, each tagged with labels from a set of 3K. The last row contains results for a model that was trained much longer than the others, with a reduction in learning rate after 13 million steps.

Architectures	Steps	Time	AUC	d-prime	mAP
Fully Connected	5M	35h	0.851	1.471	0.058
AlexNet	5M	82h	0.894	1.764	0.115
VGG	5M	184h	0.911	1.909	0.161
Inception V3	5M	137h	<b>0.918</b>	<b>1.969</b>	0.181
ResNet-50	5M	119h	0.916	1.952	<b>0.182</b>
ResNet-50	17M	356h	<b>0.926</b>	<b>2.041</b>	<b>0.212</b>

### [ Feature 별 호흡음 분류 성능 비교 ]

#### [PDF] [Crackle and wheeze detection in lung sound signals using convolutional neural networks](#)

PS Faustino - 2019 - [repositorio-aberto.up.pt](#)

MOTIVATION: Respiratory disease is among the leading causes of death in the world. Most of these deaths occur in poorer countries where pollution is more prominent and medical care is less accessible. Prevention and early detection are essential steps in managing respiratory disease. Auscultation is an essential part of clinical examination as it is an inexpensive, noninvasive, safe, easy-to-perform, and one of the oldest diagnostic techniques used by the physician to diagnose various pulmonary diseases. The drawbacks ...

☆ 99 관련 학술자료 00 no code implementation

Table 10 - Five-fold cross validation mean test metrics for each method. The best results for each of the metrics are highlighted in bold.

Input features	Num. param.	Mean Test accuracy	Mean AS	Mean HS	Mean SP	Mean SE	Mean 'Normal' recall	Mean 'Crackle' recall	Mean 'Wheeze' recall	Mean 'Both' recall
Raw audio	63,428	37%	0.37	0.36	0.41	0.33	0.41	0.45	0.22	0.09
PSD	74,756	40%	0.40	0.39	0.37	0.42	0.37	0.52	0.33	0.28
MS	74,756	<b>43%</b>	<b>0.43</b>	<b>0.42</b>	0.36	<b>0.51</b>	0.36	<b>0.62</b>	<b>0.37</b>	<b>0.34</b>
MFCC	74,756	43%	0.42	0.42	<b>0.42</b>	0.42	<b>0.42</b>	0.55	0.26	0.26

# DCASE Challenge

## ❖ 인공지능 기반 음향 이벤트 및 장면 인식 경진대회 DCASE Challenge

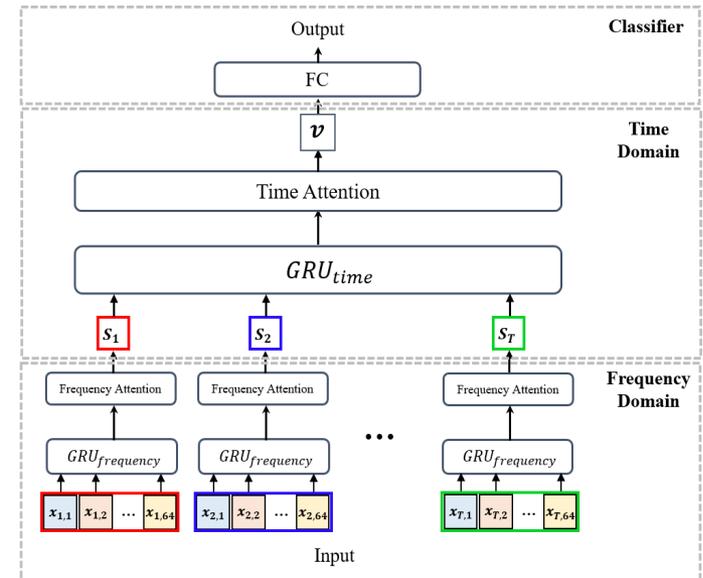
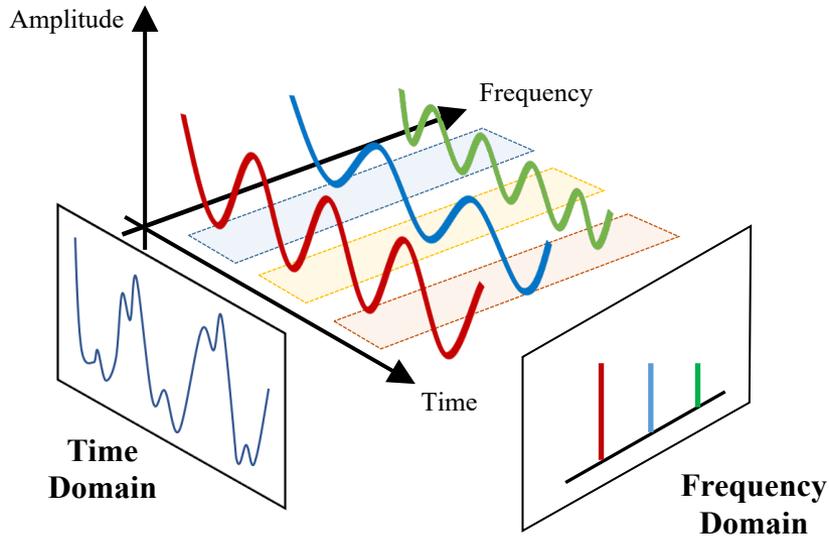
- IEEE AASP가 주관하는 음향 기술 관련 대회
- 첨단 AI 기술을 활용해 오디오 신호를 이해하고 해석하는 것을 목표로 하는 대회



- ✓ **Acoustic Scene Classification**
- ✓ **Unsupervised Anomalous Sound Detection for Machine condition Monitoring**
- ✓ **Sound Event Localization and Detection with directional inference**
- ✓ **Sound Event Detection and Separation in Domestic Environments**
- ✓ **Few-shot Bio-acoustic Event Detection**
- ✓ **Automated Audio Captioning**

# Conclusion

- 소리의 개념과 소리 데이터에 대한 적절한 전처리 및 특징 추출 기법 소개
- Wavenet과 같은 소리 데이터에 적합한 다양한 딥러닝 모델이 존재
- 최근 다양한 영역의 알고리즘 기법이 소리 데이터에 적용되고 있는 추세



[ model architecture ]

# Reference

---

- [1] · Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
  
- [2] · Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., ... & Bengio, Y. (2016). SampleRNN: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837..
  
- [3] · Suh, S., Park, S., Jeong, Y., & Lee, T. (2020). Designing acoustic scene classification models with CNN variants. DCASE2020 Challenge, Tech. Rep.
  
- [4] Faustino, P. S. (2019). Crackle and wheeze detection in lung sound signals using convolutional neural networks.
  
- [5] Minami, K., Lu, H., Kim, H., Mabu, S., Hirano, Y., & Kido, S. (2019, October). Automatic classification of large-scale respiratory sound dataset based on convolutional neural network. In 2019 19th International Conference on Control, Automation and Systems (ICCAS) (pp. 804-807). IEEE
  
- [6] · <https://deepmind.com/blog/article/wavenet-generative-model-raw-audio>.
  
- [7] · <http://www.secmem.org/blog/2019/08/18/wavenet/>.

---

# 감사합니다