Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Information Technology

Intermediate Presentation Bachelor Thesis
Deep embedded Music

Fabian Gröger

17.04.2020

Fabian Gröger
Bachelor Student Computer Science
School of Information Technology
Lucerne University of Applied Sciences and Arts

FH Zentralschweiz

Agenda

- **Project overview**
- **Project management**
- **Related Work / State of the art (Triplet loss and Tile2Vec)**
- **Datasets used**
- **Ideas and concepts (Approaches)**
- **First results**
- **Summary and outlook**
- **Q&A**

Project Overview:
Unsupervised embedding space for audio

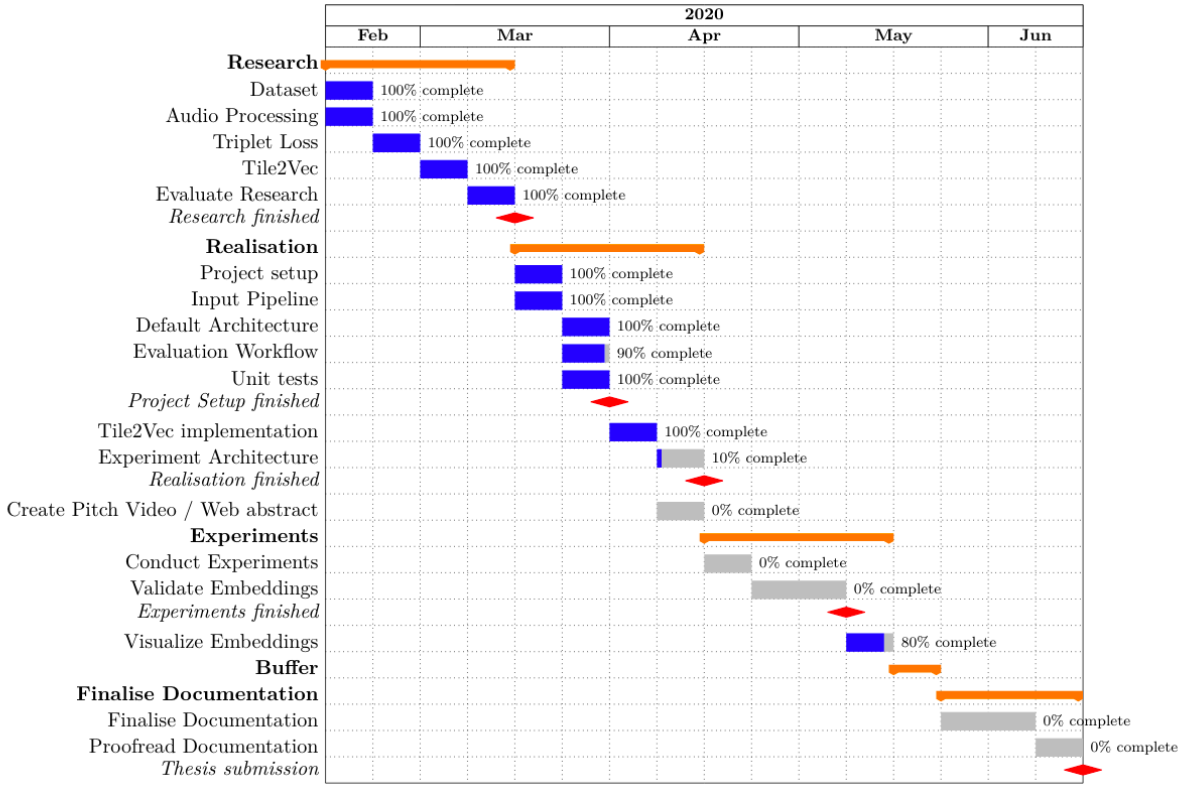
Idea

- Create an **embedding space for audio**, where distances have meaning
- Adapt **Tile2Vec**, an image embedding algorithm, to audio
- Evaluate the **performance on the DCASE 2018 task 5 dataset**
- Use **unsupervised machine learning** to embed samples
- **Train a simple classifier** on the embeddings
- The resulting embedding algorithm should be **applied exploratively to music**

To-Do

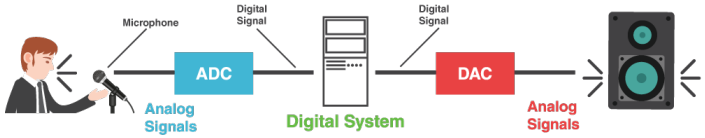
- Adapt **Tile2Vec** algorithm to music with **Tensorflow 2.0**
- Find an **appropriate audio feature** to represent the samples
- Find an **embedding architecture** on the basis of neural networks
- **Evaluate** the architecture on a **music dataset**

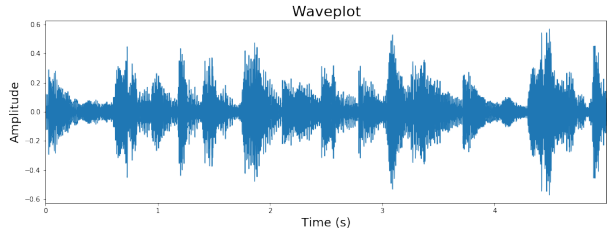
Project Management



Related Work:
Intro to Digital Signal Processing

- **Digital Signal Processing (DSP)** takes a real-world signal which has been digitised and mathematically manipulates it
- Signals need to be processed so that the information they contain can be analysed
- Converters such as an **Analog-to-Digital converter (ADC)** takes a real-world signal and turns it into a digital format
- **Digital-to-Analogue (DAC)** does the opposite of the ADC and converts the digital signal back to an analogue
- **Sound signals** can be defined as pressure variations travelling through the air, which can be described as waves and therefore often called **sound waves**
- **Raw waveform:** representations of a signal in the **time-domain**
  y-axis: represents amplitude (loudness)
  x-axis: represents the time

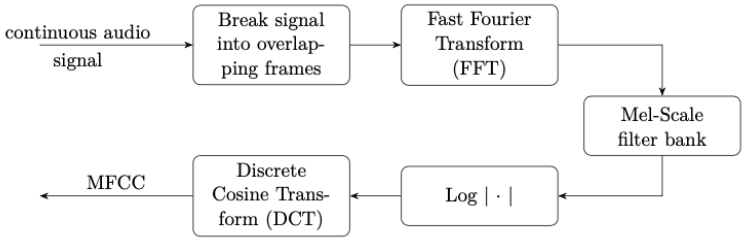


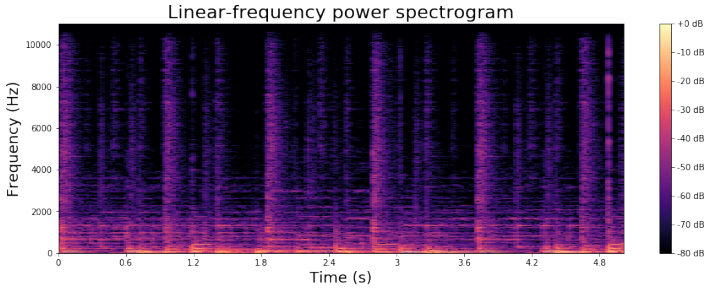


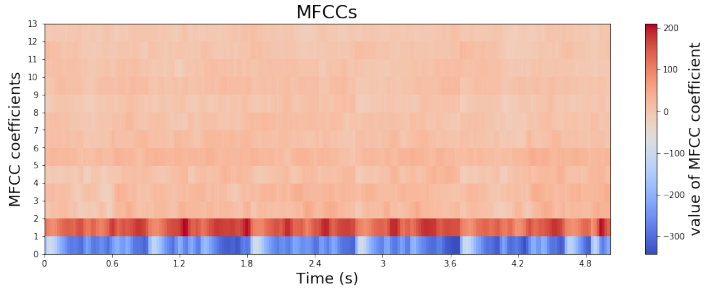
References: [1]

Related Work:
Intro to Digital Signal Processing

- **Spectrogram:** visual representation of the spectrum of **frequencies of a signal as it varies with time**
  x-axis: time
  y-axis: frequency
  colours: magnitude of the frequency at a time
- **Log Mel spectrogram:** visual representation of the signal transformed into the log scale
- **Mel Frequency Cepstral Coefficients (MFCC) focus on the audio signal** and discard information such as background noise
- MFCC were designed to **mimic the human hearing**

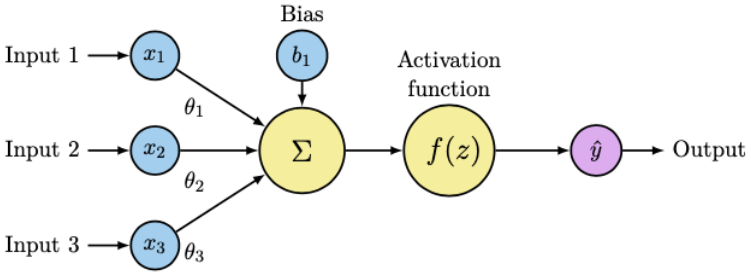


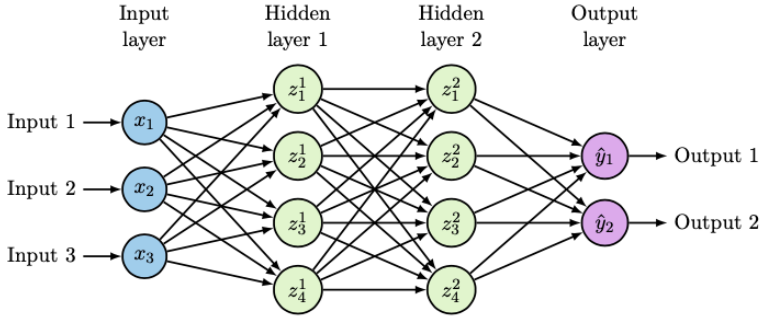




References: [3]

- **Artificial Neural Network (ANN or NN)** is a computational model that is inspired by the way the brain works
- NN **consists of many neurons** (Preceptrons)
- A Preceptron consist of **input(s)**, **a bias** and **an activation function** which are used to calculate an output
- **Feedforward Neural Networks** contain **multiple nodes arranged in layers**, where **each node is connected to the adjacent** ones
- Information only moves forward
- **Multi Layer Preceptron (MLP)** contains an **input**, **output** and **one or more hidden layers**

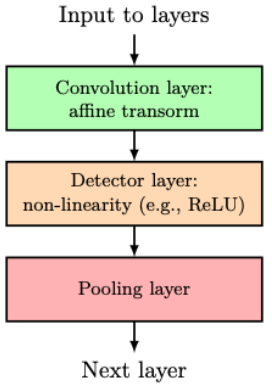




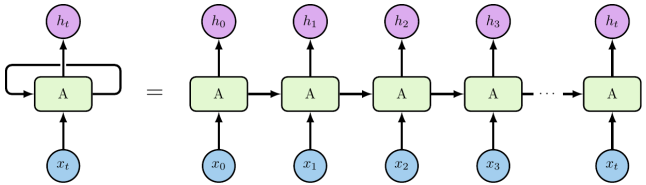
Related Work:
Intro to Neural Networks

- **Convolutional Neural Network (CNN)** specialise in working with **grid like data** (e.g. sound, images)
- CNN are networks that **use convolution** in place of general matrix multiplication in at least one of their layers

$$S(i,j) = (K * I)(i,j) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} I(i-m, j-n) \cdot K(m,n)$$



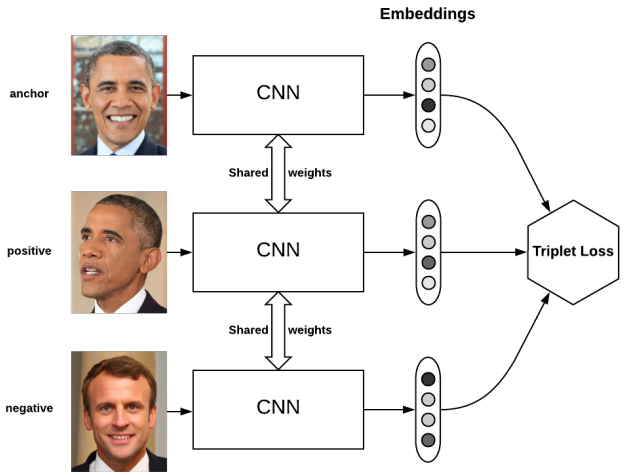
- **Gated Recurrent Unit (GRU)** belong to the family of **recurrent neural network (RNN)**, which focus on **processing sequential data**
- RNNs use a **cyclic loop** to pass the information from the step before to the next along with the input of the current step
- GRU were introduced to **solve the vanishing gradient** problem in RNNs (input from far behind looses importance)
- GRU have **two gates**, an **update and reset gate** which decide what information will be passed to the output



References: [3]

Related Work:
Triplet Loss

- Introduced by Google for face recognition
- Learn distributed **embeddings by the notion of similarity and dissimilarity**
- Siamese networks are trained, which share their weights, on the different inputs
- triplets consisting of an **anchor x_a**, a **positive x_p** and a **negative x_n**
- α is denoted as a **margin**, which puts a limit on how far the network can push the negative sample away to improve the loss



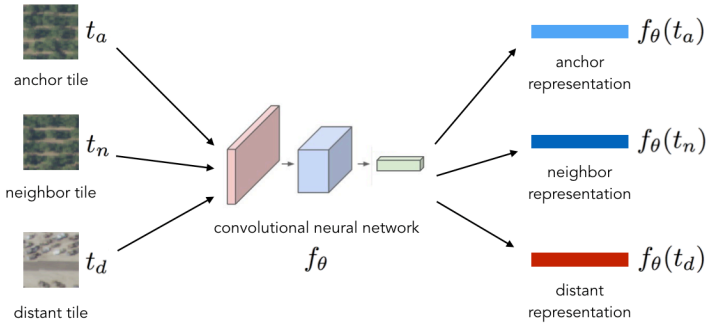
$$d(p,q) = ||q-p||_2 = \sqrt{(q_1-p_1)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$

$$\mathcal{L} = \left[\left|\left|f_\theta(x_a) - f_\theta(x_p)\right|\right|_2^2 - \left|\left|f_\theta(x_a) - f_\theta(x_n)\right|\right|_2^2 + \alpha\right]_+$$

References: [2]

Related Work:
Tile2Vec

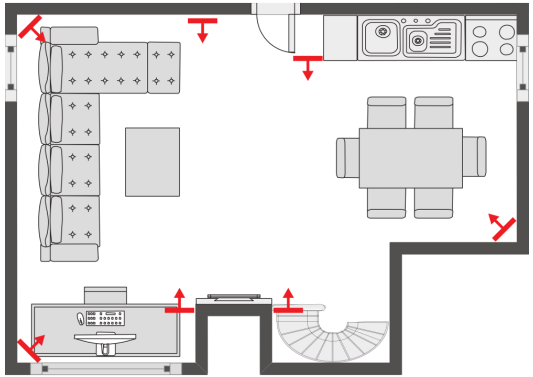
- Introduced by Stanford University
- **unsupervised representation learning algorithm**
- extends the **distributional hypothesis** from natural language to spatially distributed data, **words appearing in similar contexts** tend to have **similar meanings**
- Introduce a l^2 **– regularisation** term, to force embeddings to be within a hypersphere
- $\mathbf{x_p}$ lies within a **neighbour radius** from the anchor
- $\mathbf{x_n}$ lies within a **opposite radius** from the anchor



References: [3]

Datasets used:
DCASE Challenge 2018 – Task 5 dataset and Music Dataset

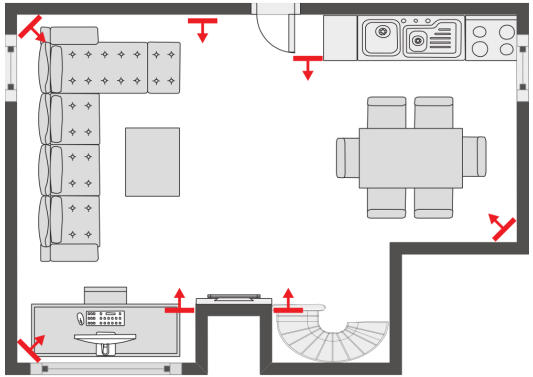
- Derivative of the **SINS dataset**
- Continuous recording of one person living in a vacation home over a period of one week
- **DCASE: 7 microphone** arrays in the combined living room and kitchen
- Microphone array consists of 4 linearly arranged microphones
- Continuous recordings were split into **audio segments of 10s**
- Each audio segment contains **4 channels**
- As development set, **approximately 200 hours of data from 4 sensor nodes** along with the ground truth is given
- As evaluation set, data is provided from **all the sensor nodes**
- **Winner:** IBM Team, **macro-averaged F1** score of **88.4%**



References: [4]

Datasets used:
DCASE Challenge 2018 – Task 5 dataset and Music Dataset

Activity	# 10s segments	# sessions
Absence (nobody present in the room)	18860	42
Cooking	5124	13
Dishwashing	1424	10
Eating	2308	13
Other (present but not doing any relevant activity)	2060	118
Social activity (visit, phone call)	4944	21
Vacuum cleaning	972	9
Watching TV	18648	9
Working (typing, mouse click)	18644	33
Total	**72984**	**268**



References: [4]

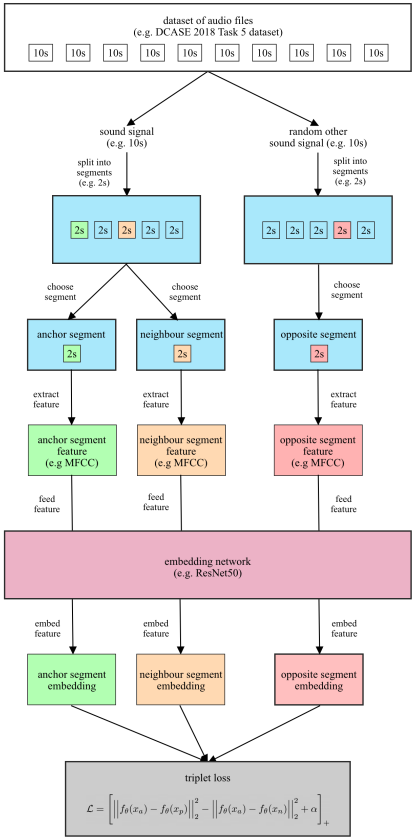
Datasets used:
Music Dataset

- **7 different music genres** with **30 songs** each
- Songs classified by genre according to Beatport.com
- Approximately **18 hours** of audio recordings
- Songs **vary in length**, from 3min to 10min
- Single channel .mp3 files

Music genre	# songs
Deep House	30
Electronica Downtempo	30
Indie Dance	30
Melodic House and Techno	30
Techno Peak Time Driving Hard	30
Techno Raw Deep Hypnotic	30
Trance	30
Total	**210**

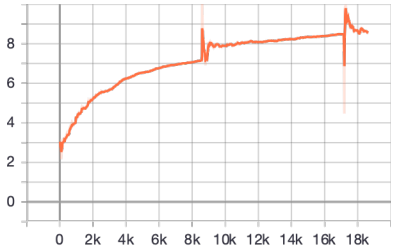
Ideas and Concepts

- **Data pre-processing:**
  None
- **Feature extraction:**
  Raw waveform, log mel spectrogram, MFCC
- **Data augmentation:**
  Shuffling and mixing
- **Triplet selection:**
  Songs split into segments, anchor neighbour belong to same audio, opposite belongs to a different
- **Models:**
  state-of-the-art CNN, GRU, CRNN
- **Application to music:**
  same as for noise detection set
- **Metrics:**
  Embedding: triplet loss, distance to neighbour and opposite
  Classifier: loss, accuracy, macro-averaged F1 score

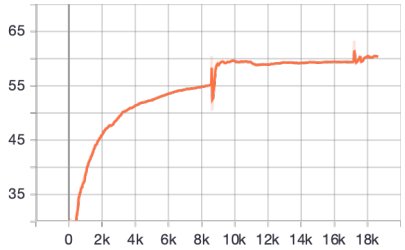


First results:
Metrics

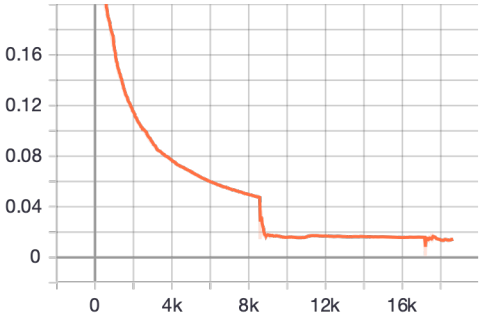
Distance neighbour²

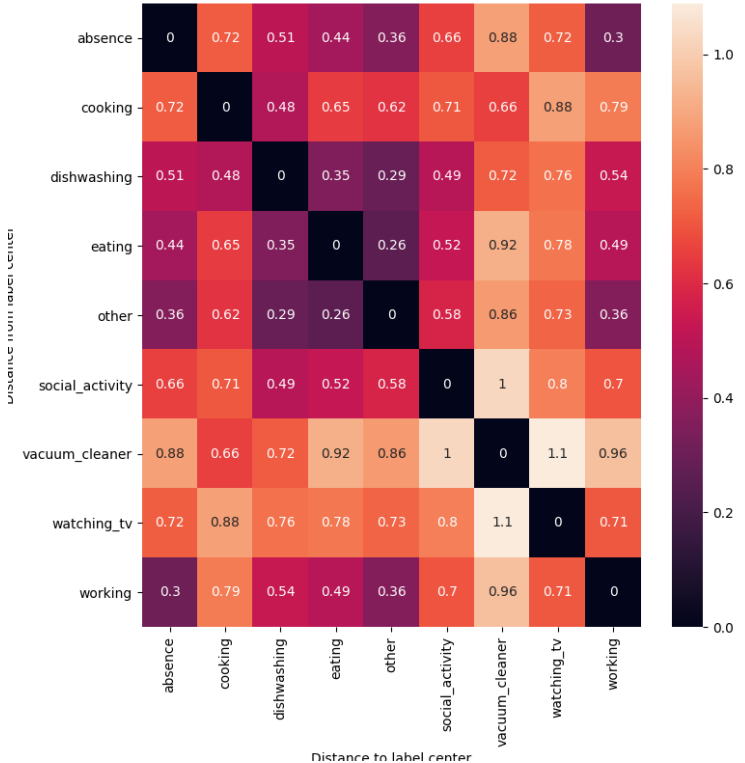


Distance opposite²

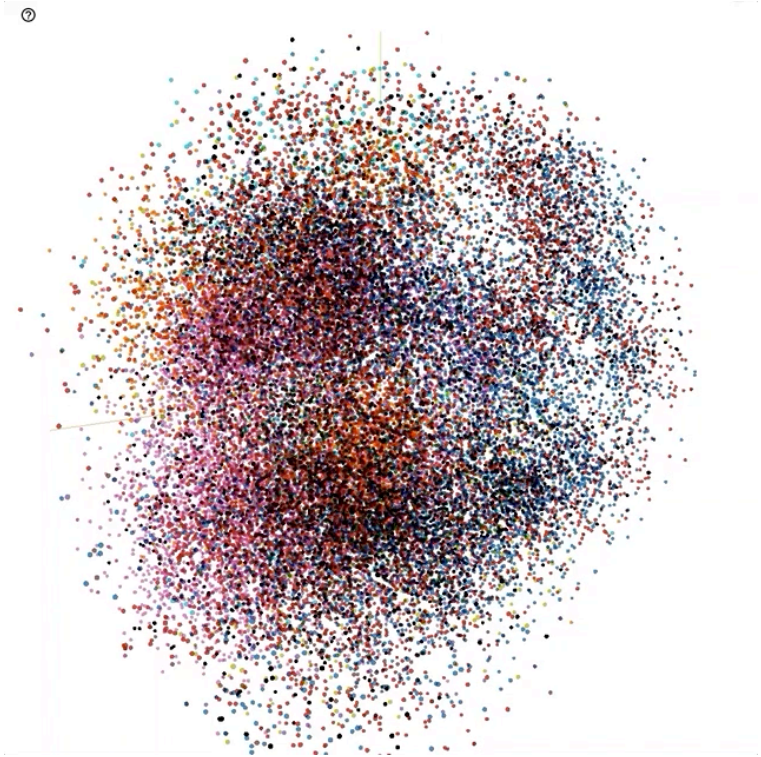


Triplet loss





First results:
Projection of embeddings



Summary and Outlook

Summary:

- Create an embedding space for audio using an unsupervised triplet loss
- Examine the embedding space

Outlook:

- Experiment with best audio feature representation for task
- Experiment with state-of-the-art architectures
- Apply same architectures to music
- Examine the clusters in the embedding space
- Hyperparameter tuning

Questions & Answers

References

- ***Introduction to Digital Signal Processing***

[1] https://www.allaboutcircuits.com/technical-articles/an-introduction-to-digital-signal-processing/

- ***Triplet Loss***

[2] https://omoindrot.github.io/triplet-loss

- ***Tile2Vec***

[3] https://ermongroup.github.io/blog/tile2vec/

- ***DCASE Challenge 2018 Task 5***

[4] http://dcase.community/challenge2018/task-monitoring-domestic-activities