Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Information Technology

Intermediate Presentation Bachelor Thesis
Deep embedded Music

Fabian Gröger

21.04.2020

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**
- **Datasets used**
- **Ideas and concepts**
- **First results**
- **Summary and outlook**
- **Q&A**

Project Overview:
Unsupervised embedding space for audio

Idea

- Create an **embedding space for audio**, using **unsupervised machine learning**
- Adapt **Tile2Vec**, an image embedding algorithm, to audio
- Evaluate the **performance on the DCASE 2018 task 5 dataset**
- **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** based on neural networks
- **Evaluate** the architecture on a **music dataset**

Project Management

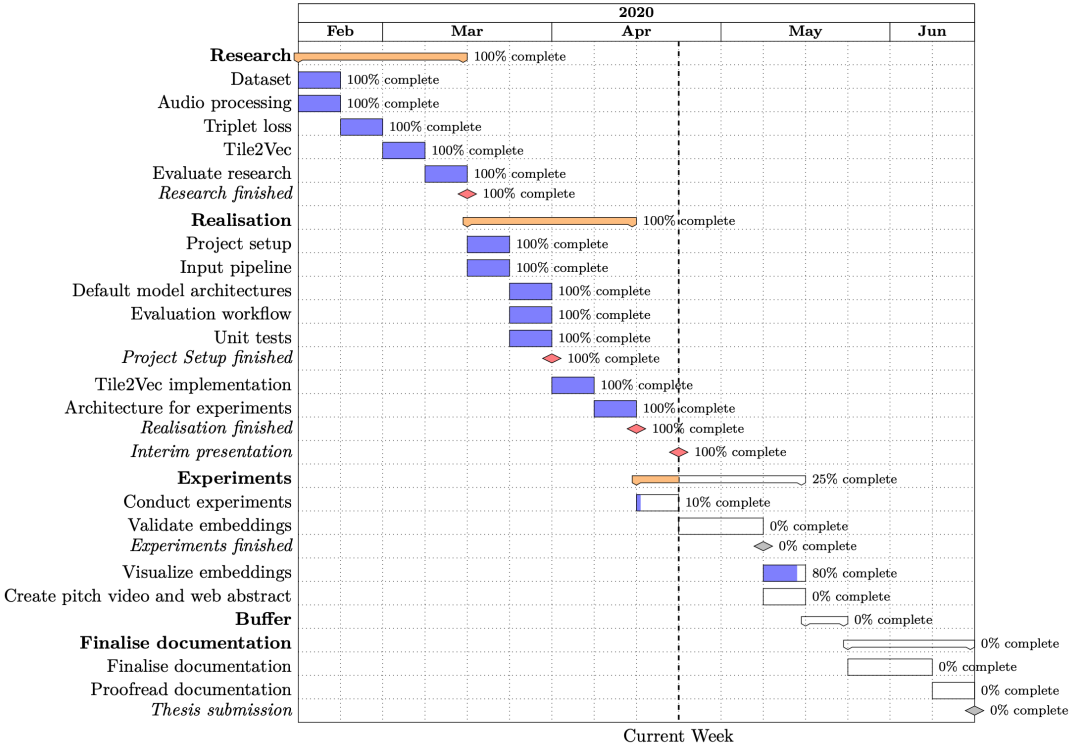

Fig. 1 Project plan

Related Work:
Intro to Digital Signal Processing

- **Digital Signal Processing (DSP)** takes a real-world signal which has been digitised and mathematically manipulates it
- **Sound signals** can be defined as pressure variations travelling through the air, which are often referred to as **sound waves**
- **Raw waveform:** representations of a signal in the **time-domain**
  x-axis: represents time
  y-axis: represents amplitude
- **Spectrogram:** visual representation of the spectrum of **frequencies of a signal as it varies with time**
  x-axis: represents time
  y-axis: represents frequency
  colours: magnitude of the frequency at a time

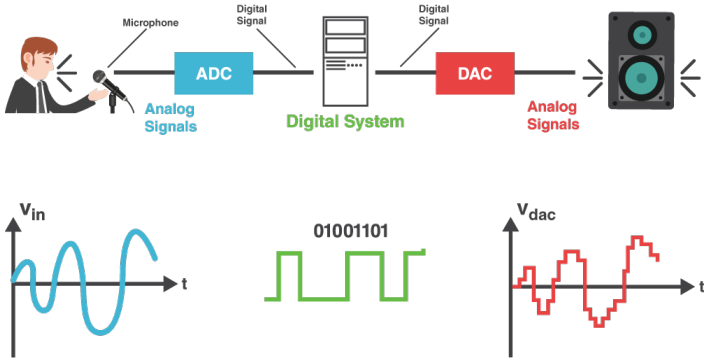

Fig. 2 DSP illustration

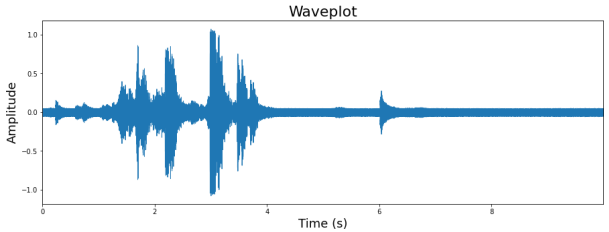

Fig. 3 Waveplot

References: [1] [2]

Related Work:
Intro to Digital Signal Processing

- **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**
- **Log Mel spectrogram:** visual representation of the frequencies in the log Mel scale

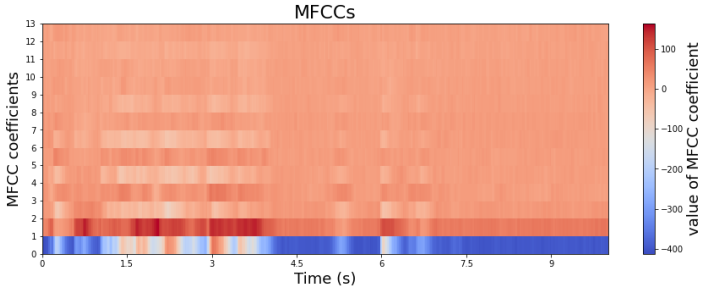

Fig. 4 Visualisation of MFCC

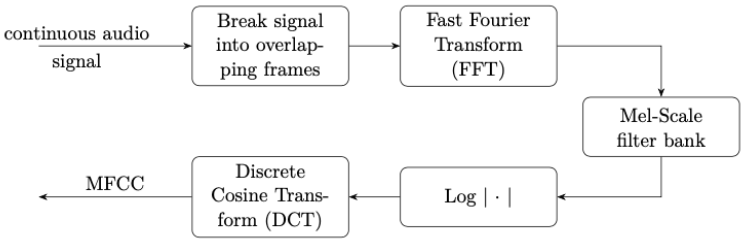

Fig. 6 Visualisation calculating MFCC

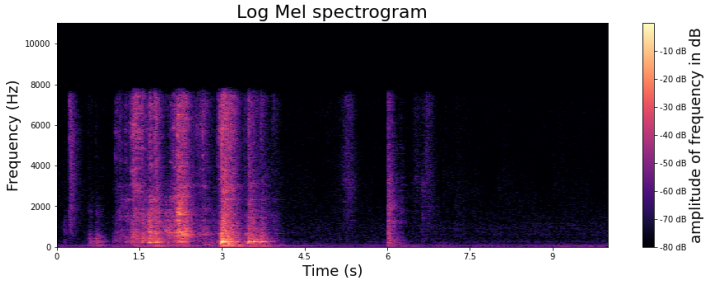

Fig. 5 Log Mel spectrogram

References: [2]

Related Work:
Intro to Neural Networks

- **Artificial Neural Network (ANN or NN)** are a computational model which were inspired by the way the brain works
- NN **consist of many neurons** (Perceptron)
- Perceptron's consist of **input(s)**, **bias** and **an activation function** which are used to calculate an output
- **Feedforward Neural Networks** contain **multiple neurons arranged in layers**, where **each neuron is connected to the adjacent** one and information **only moves forward**
- **Multi Layer Perceptron (MLP)** contain an **input**, **output** and **one or more hidden layers**

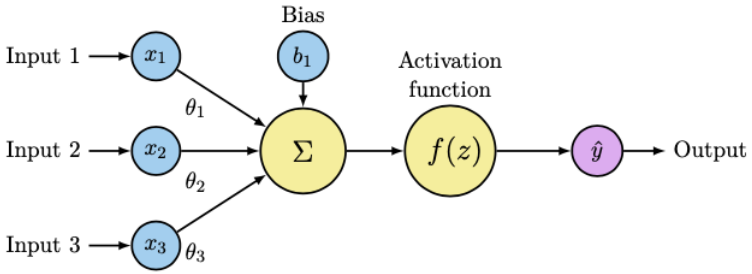

Fig. 7 Visualisation of a Perceptron

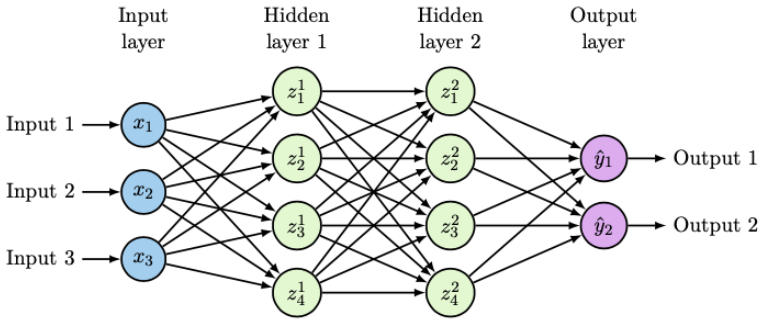

Fig. 8 Visualisation of an MLP

References: [3]

Related Work:
Intro to Neural Networks

- **Convolutional Neural Network (CNN)** specialises 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

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

Eq. 1 Convolutional operation

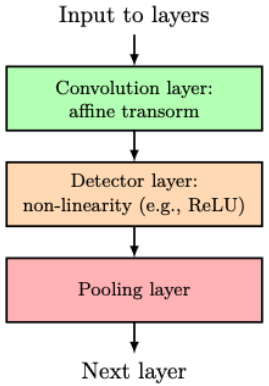

Fig. 9 CNN layer structure

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

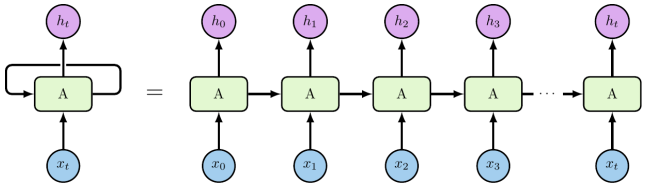

Fig. 10 Visualisation RNN

References: [3]

Related Work:
Triplet Loss

- Learn distributed **embeddings by the notion of similarity and dissimilarity**
- 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

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

Eq. 2 Euclidian distance

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

Eq. 3 Triplet loss

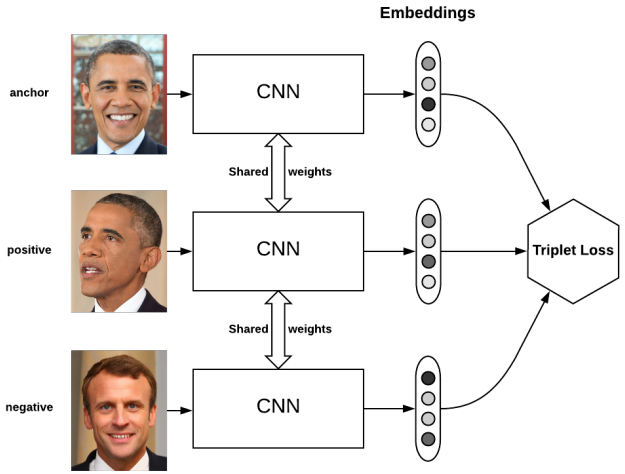

Fig. 11 Triplet loss visualisation

References: [4] [5]

Related Work:
Tile2Vec

- **Adapt triplet loss** to an **unsupervised setting** using the **distributional hypothesis** from natural language (**words appearing in similar contexts** tend to have **similar meanings**)
- $\mathbf{x_p}$ lies within a **neighbour radius** from the anchor
- $\mathbf{x_n}$ lies outside of the **neighbour radius**
- $\boldsymbol{l^2}$ **– normalisation**, to force embeddings to be within a hypersphere

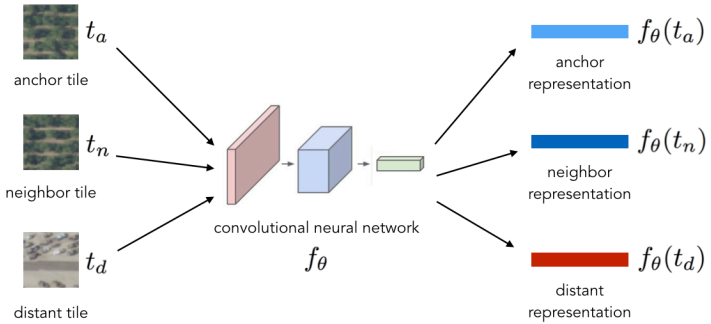

Fig. 12 Visualisation Tile2Vec

References: [6] [7]

Datasets used:
DCASE Challenge 2018 – Task 5 dataset

- Derivative of the **SINS dataset**, a continuous recording of one person living in a vacation home over one week
- **DCASE: 7 microphone** arrays (4 linearly arranged) in the combined living room and kitchen
- Continuous recordings were labelled and split into **audio segments of 10s**
- As development set, **approximately 200 hours of data from 4 microphone arrays**
- As evaluation set, data is provided from **all the microphone arrays**
- **Winner:** IBM Team, **macro-averaged F1** score of **88.4%**

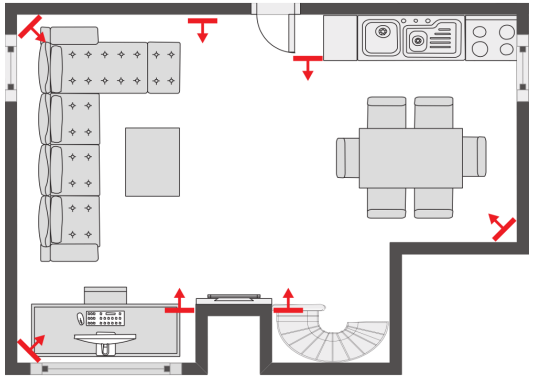

Fig. 13 DCASE microphones distribution

References: [8]

Datasets used:
Music Dataset

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

Tbl. 1 Music dataset

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
- **Triplet selection:**
  Songs split into segments, anchor and neighbour belong to the same audio, opposite belongs to a different one (temporal proximity)
- **Models:**
  state-of-the-art CNN, GRU, CRNN architectures
- **Application to music:**
  same as for noise detection set
- **Metrics:**
  Embedding: triplet loss, distance to neighbour and opposite
  Classifier: sparse categorical cross-entropy loss, sparse categorical accuracy, macro-averaged F1 score

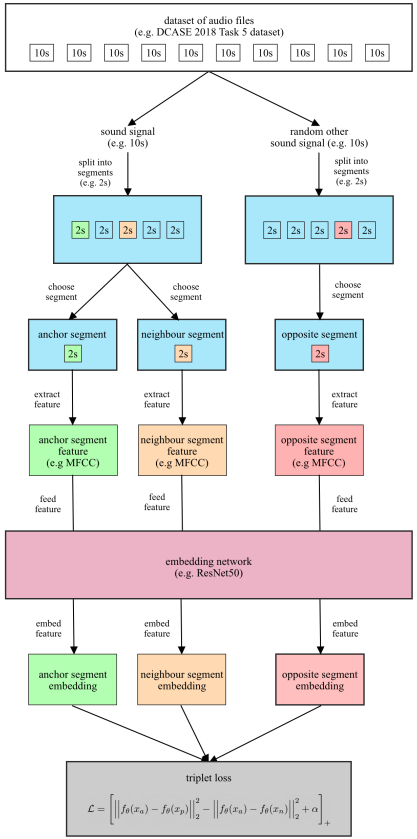

Fig. 14 Input pipeline

First results:
Metrics

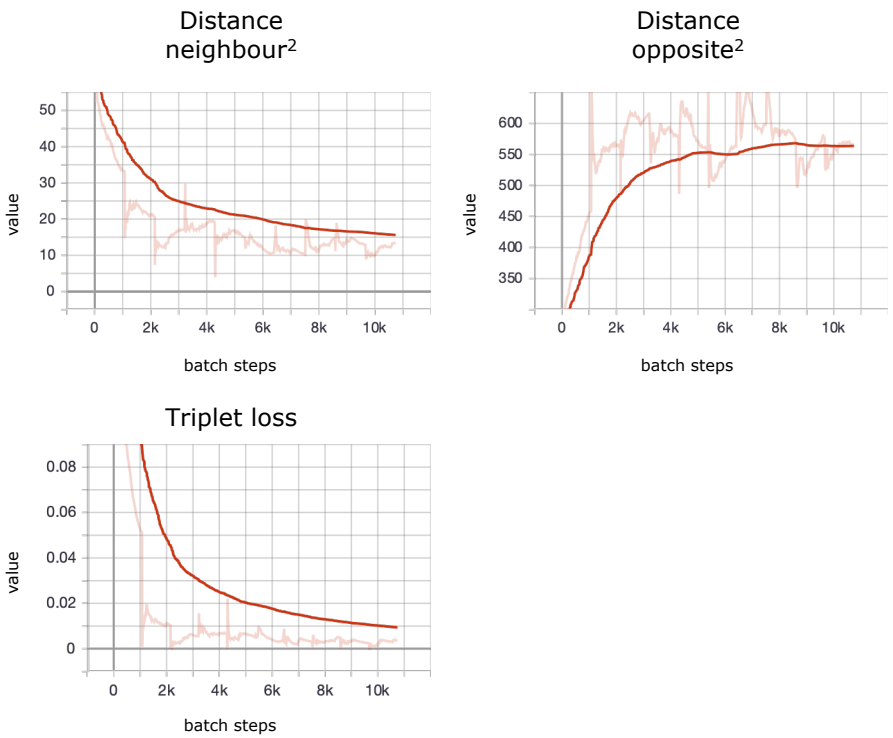

Fig. 15 Metrics

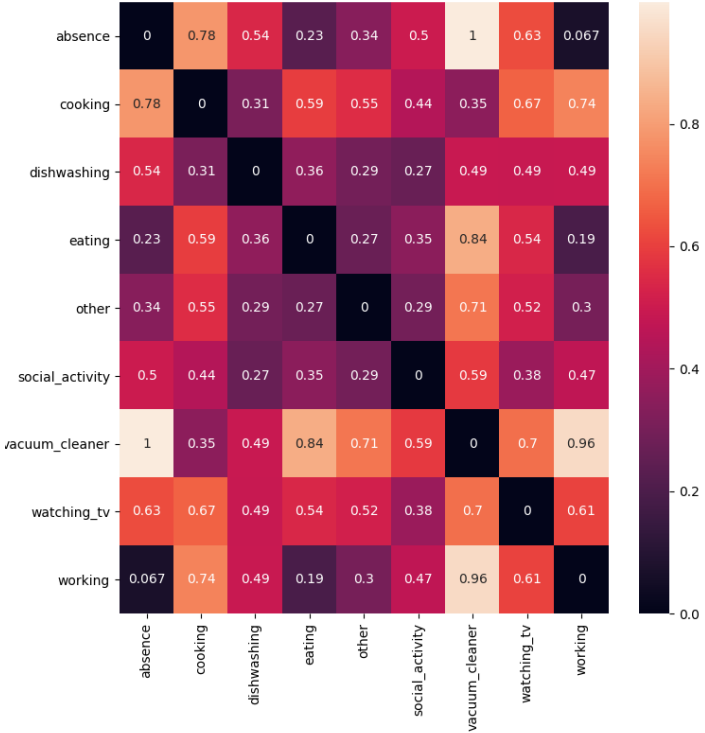

Fig. 16 Distance matrix

First results:
Projection of embeddings

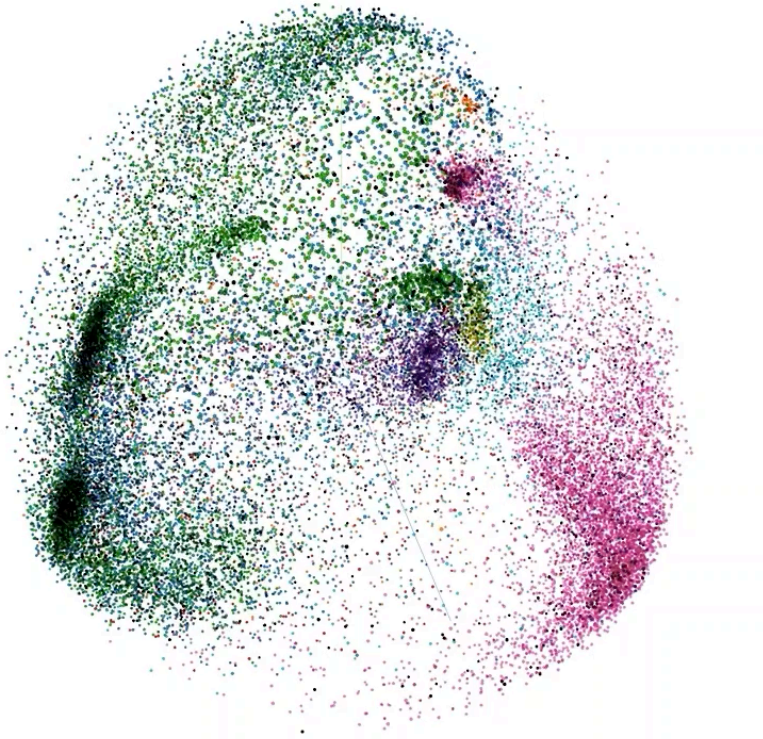

Fig. 17 Visualisation of embedding space

Summary and Outlook

Summary:

- Create an **embedding space for audio** using an **unsupervised triplet loss**
- **Evaluate the embedding** space for noise detection and music

Outlook:

- Experiment with **most suitable audio feature representation**
- Experiment with **state-of-the-art architectures**
- Apply **embedding architecture to music**
- **Hyperparameter tuning** for noise and music architecture


Q & A

References

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

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

- ***Fundamentals of Music Processing***

[2] Müller, Meinard. *Fundamentals of Music Processing*. Cham: Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-21945-5.

- ***Deep Learning***

[3] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

- ***Triplet Loss***

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

[5] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, 815–23. https://doi.org/10.1109/CVPR.2015.7298682.

References

- ***Tile2Vec***

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

[7] Jean, Neal, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. "Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data." *ArXiv:1805.02855 [Cs, Stat]*, May 30, 2018. http://arxiv.org/abs/1805.02855.

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

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