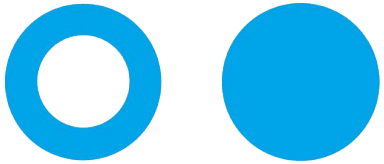


Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Environmental sound classification using deep learning

Supervisors:
Prof. Dr. Paul Plöger
M.Sc. Anastassia Coast Makers

Structure of presentation

- Problem formulation.
- Deliverables
- Proposed model
- Dataset
- Time, frequency and spectrogram view of data

Problem formulation

- Intelligent sound detection(ISR) is a technology for identifying sound events that exist in the real environment and embedding such ability in machines or robots.
- Environmental sound classification(ESC) serve as a fundamental steps of ISR.
- The main goal of ESC is to precisely classify the class of a detected sound.
- Automatic speech recognition (ASR) and music information recognition (MIR) will be inefficient when applying to non stationary ESC task because environmental sound is non stationary in nature.
- Therefore, it is essential to develop an efficient ISR system for environment sound(ES) recognition.
- The main obstacles of current algorithms for ESC task are
- 1) log-mel and MFCC are the most widely used acoustic features applied to ESC, but were originally designed for ASR and MIR. ES is non stationary signal and using single features lead to the failure of capturing the import information about ES.
- 2) In recent years neural network has gained popularity for environmental sound classification. However performance is still unsatisfactory.
- Hence, there is a need to develop appropriate auditory features and novel neural network models to achieve high categorization accuracy for ESC tasks.

Deliverables

Minimum Viable

- Literature survey related to environmental sound classification techniques.
- Analysis of state of the art.
- Collection of appropriate environmental sound dataset for training and testing.

Expected

- Develop a model for evaluation.
- Train the model with gathered datasets.

Desired

- Evaluation of the proposed model with different datasets.
- Comparison of proposed model accuracy with different dataset and previously developed model accuracy.

Proposed architecture

- According to [2], temporal and spectral characteristics of the sound needs to be considered during processing of the signal to make a classification.
- Representing the sound signal with temporal and spectral characteristics maximize the information content pertaining to sound signal.
- A considerable number of research [1] works indicate that combined features performed better than only use one feature set in ESC tasks.
- The proposed model in [1] has achieved good accuracy rate compared to the previously proposed models.
- However the model in [1] doesn't incorporate the temporal dependency of the extracted features.
- The proposed model (Figure 1) takes into consideration both temporal and spectral characteristics.
- 1) Combined features are extracted from the sound
- 2) The features is fed into four four-layer(Two branches, each having four layers of CNN) to extract the spectral features.
- 3) RNN-LSTM is used to find the temporal dependency.
- 4) Two branches output is fused using the Dempster–Shafer evidence theory.

Courtesy:Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion. Yu Su 1,2,* , Ke Zhang 1, Jingyu Wang 1 and Kurosh Madani 2

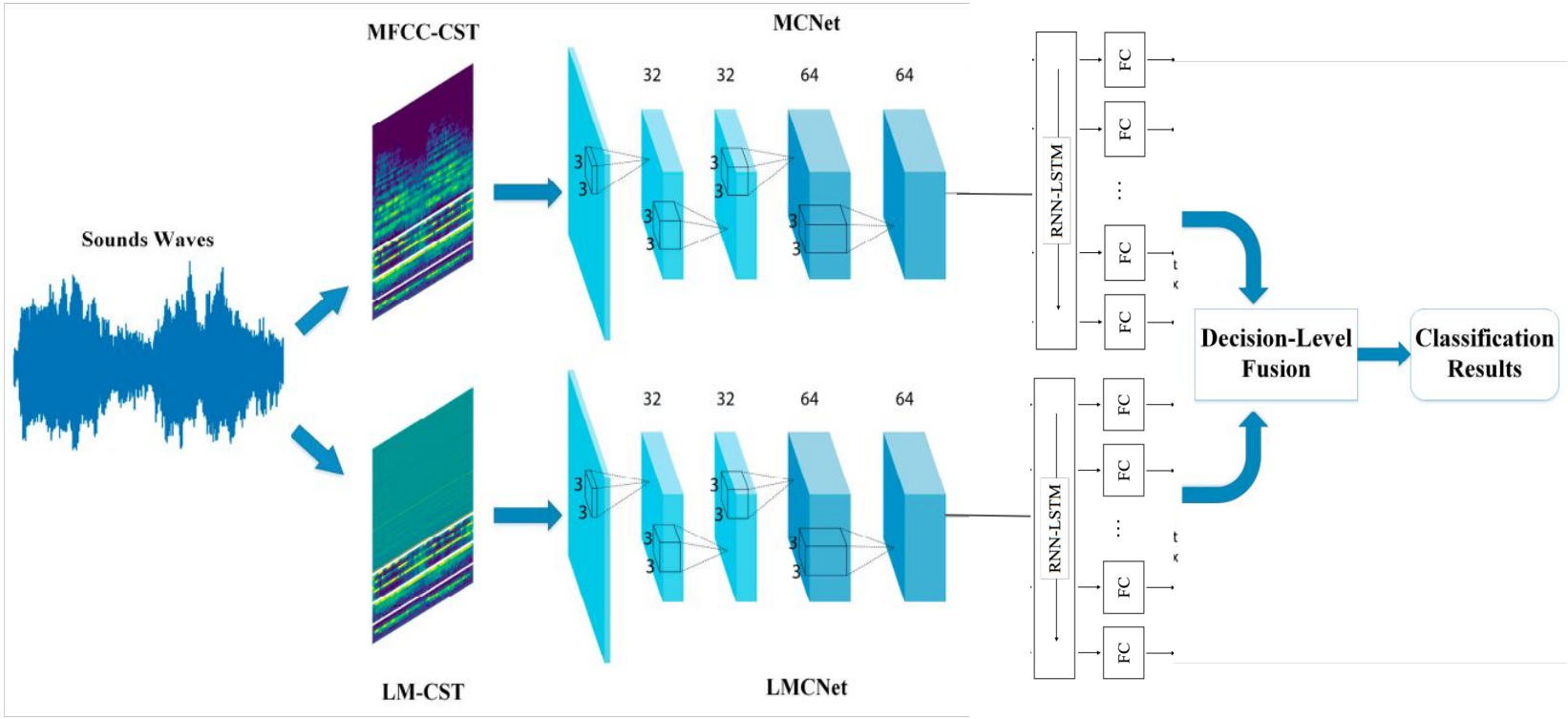


Figure 1: The proposed overall framework

Existing model and their performance

Courtesy: RARE SOUND EVENT DETECTION USING 1D CONVOLUTIONAL RECURRENT NEURAL NETWORKS Hyungui Lim1, Jeongsoo Park1;2, Kyogu Lee2, Yoonchang Han1

- "Rare sound event detection using 1D convolutional recurrent neural networks." [2]:
- Proposed framework in [2] is shown in the figure 2 below.
- The framework consist of four major steps: 1) log-amplitude mel-spectrogram extracted from audio 2) Spectral features are extracted with 1D ConvNet 3) RNN-LSTM is used to find the temporal dependency. 4) sound event detection and the onset time of audio.
- Dataset used for evaluation: TUT Rare Sound Events 2017, The dataset consist of three classes with background noise, ‘baby crying’, ‘glass breaking’, and ‘gunshot’.
- Result: Baseline: DCASE 2017 baseline system, ER-error rate. Ensemble is used to find overall effect.

Performance of baseline and proposed system in the evaluation set.

	ER		F-score	
	baseline	proposed	baseline	proposed
Baby crying	0.80	**0.15**	66.8	**92.2**
Glass breaking	0.38	**0.05**	79.1	**97.6**
Gunshot	0.73	**0.19**	46.5	**89.6**
Overall	0.64	**0.13**	64.1	**93.1**

Courtesy: RARE SOUND EVENT DETECTION USING 1D CONVOLUTIONAL RECURRENT NEURAL NETWORKS Hyungui Lim1, Jeongsoo Park1;2, Kyogu Lee2, Yoonchang Han1

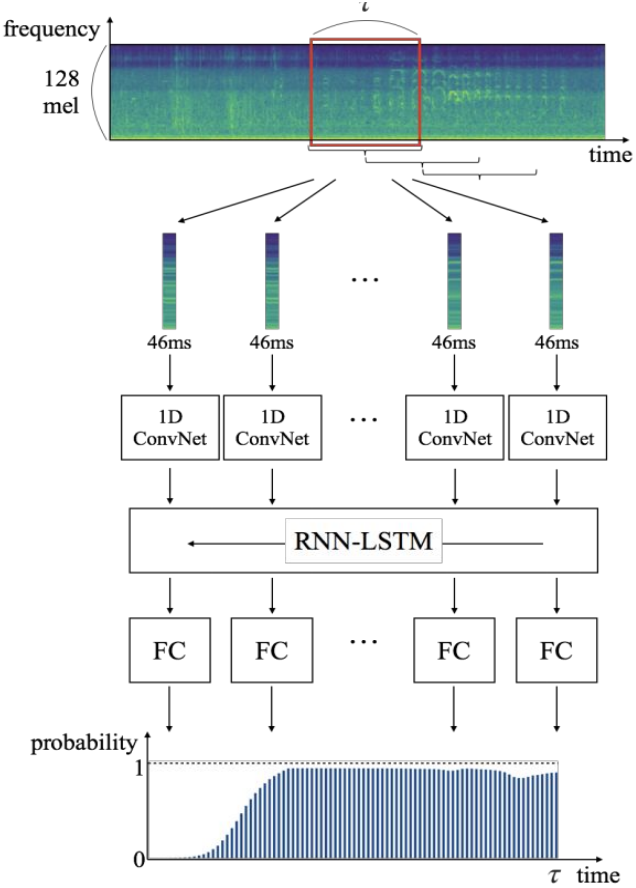


Figure 2: Overall framework of the proposed method

Existing model and their performance

Courtesy:Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion
Yu Su 1,2,* , Ke Zhang 1, Jingyu Wang 1 and Kurosh Madani 2

- Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion [2]:
- Combined features (Figure 3) are extracted from the sound and this features is fed into four four-layer(Two branches, each having four layers of CNN) CNN (Figure 4).
- Two branches output is fused using the Dempster–Shafer evidence theory (Figure 4) to compose TSCNN-DS model and prediction is made.
- Steps involved: 1) Feature Extraction and Combination (Figure 3) 2) Structure of the MCNet and LMCNet 3)Dempster—Shafer Evidence Theory-Based Information Fusion (Figure 5).
- Dataset used for evaluation: UrbanSound8K dataset (8732 labeled urban sounds, the length is less than or equal to 4s), air conditioner (ac), car horn (ch), children playing (cp), dog bark (db), drilling (dr), engine idling (ei), gunshot (gs), jackhammer (jh), siren (si) and street music (sm).
- The classification accuracy of CNN-based ISR system is nearly equal to 97% in ESC tasks.

Results

Courtesy:Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion. Yu Su 1,2,* , Ke Zhang 1, Jingyu Wang 1 and Kurosh Madani 2

Table 1. Class-wise accuracy of four models with four-layer CNN evaluated on UrbanSound8K

Class	LMC (LMCNet)	MC (MCNet)	MLMC	TSCNN-DS
ac	98.6%	99.9%	99.2%	99.9%
ch	93.9%	91.4%	93.2%	94.2%
cp	97.3%	93.9%	96.1%	97.5%
db	92.6%	90.4%	94.2%	95.3%
dr	94.8%	95.0%	95.7%	97.2%
ei	98.9%	99.6%	98.5%	99.6%
gs	88.6%	91.1%	85.9%	95.4%
jh	93.2%	95.9%	91.1%	97.1%
si	98.6%	98.3%	98.5%	98.9%
sm	95.0%	97.4%	94.1%	96.9%
Avg.	95.2%	95.3%	94.6%	97.2%

Table2. Comparison of classification accuracy with other models on UrbanSound8K datasets.

Model	Feature	Accuracy
Piczak [28]	LM	72.7%
Tokozume [35]	Raw Data	78.3%
Zhang X. [32]	Mel	81.9%
Zhang Z. [6]	LM-GS	83.7%
Li [20].	Raw Data-LM	92.2%
Boddapati [25]	Spec -MFCC-CRP	93%
LMCNet	LM-C	95.2%
MCNet	M-C	95.3%
TSCNN-DS	**MC and LMC**	**97.2%**

Courtesy:Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion
Yu Su 1,2,* , Ke Zhang 1, Jingyu Wang 1 and Kurosh Madani 2

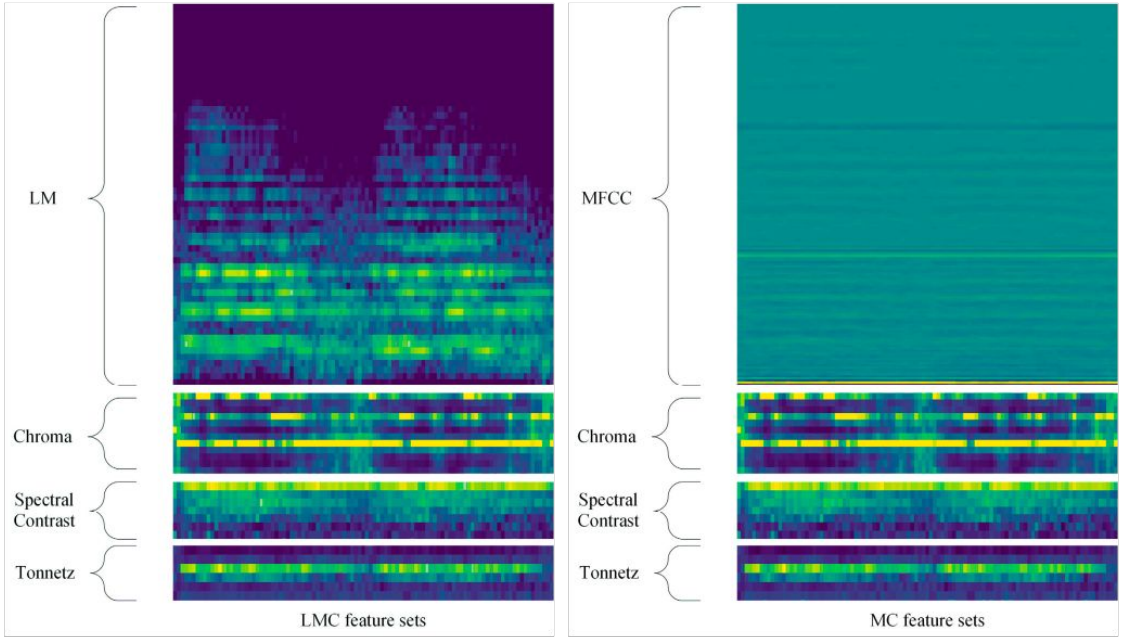


Figure 3: The spectrogram of LMC and MC feature sets

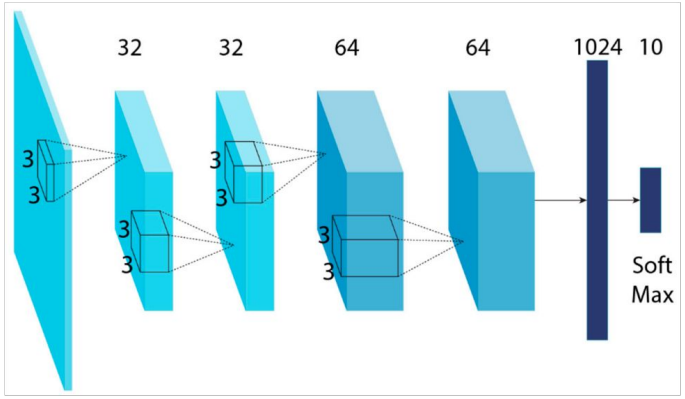


Figure 4: The architecture of the four-layer CNN

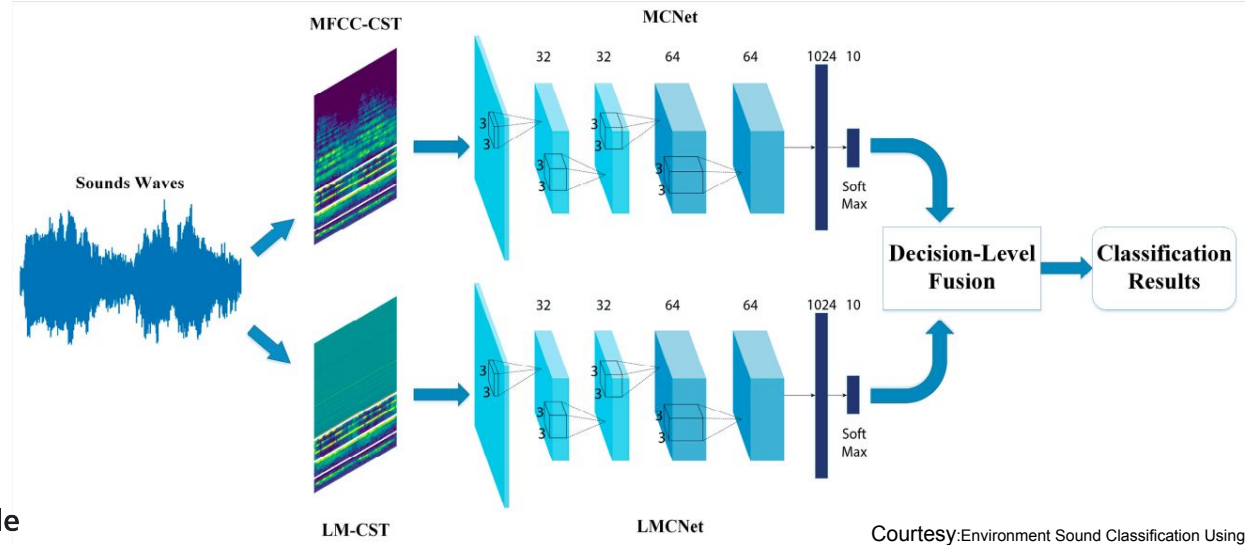


Figure 5: The overall framework

Courtesy:Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion. Yu Su 1,2,* , Ke Zhang 1, Jingyu Wang 1 and Kurosh Madani 2

Dataset

- Urbansound8k [3]:
- This dataset contains 8732 labeled sound of 4 sec length in WAV format at 22.05 kHz.
- 10 classes were defined: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jackhammer, siren, and street_music.
- Nearly 870 slices of data per class.
- DCASE 2018, Monitoring of domestic activities based on multi-channel acoustics - Evaluation dataset [4](Benchmarking available):
- Total of 72984, 10 sec audio dataset.
- The continuous recordings in home environment were split into audio segments of 10s.
- 9 classes were considered, format[activity,# 10s segment,#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..etc),18644,33]

Dataset

- Audio Events Data Set for Surveillance Applications [5];
- The MIVIA audio events data set is composed of a total of 6000 events for surveillance applications, namely glass breaking, gun shots and screams in WAV format.

	TRAINING SET		TEST SET	
	#Events	*Duration (s)*	*#Events*	*Duration (s)*
Background	–	58371,6	–	25036,8
Glass breaking	4200	6024,8	1800	2561,7
Gun shots	4200	1883,6	1800	743,5
Screams	4200	5488,8	1800	2445,4

Dataset

- DCASE 2020Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring[7]:
- **Development dataset**: Each Machine Type has three or four Machine IDs. Each machine ID's dataset consists of (i) around 1,000 samples of normal sounds for training and (ii) 100-200 samples each of normal and anomalous sounds for the test.
- **Evaluation dataset**: This dataset consists of the same Machine Types' test samples as the development dataset. The number of test samples for each Machine ID is around 400, none of which have a condition label (i.e., normal or anomaly).
- **Additional training dataset**: This dataset includes around 1,000 normal samples for each Machine Type and Machine ID used in the evaluation dataset.
- Fig 6 shows the pictorial representation of data.

Courtesy:http://dcase.community/challenge2020/task-unsupervised-detection-of-anomalous-sounds

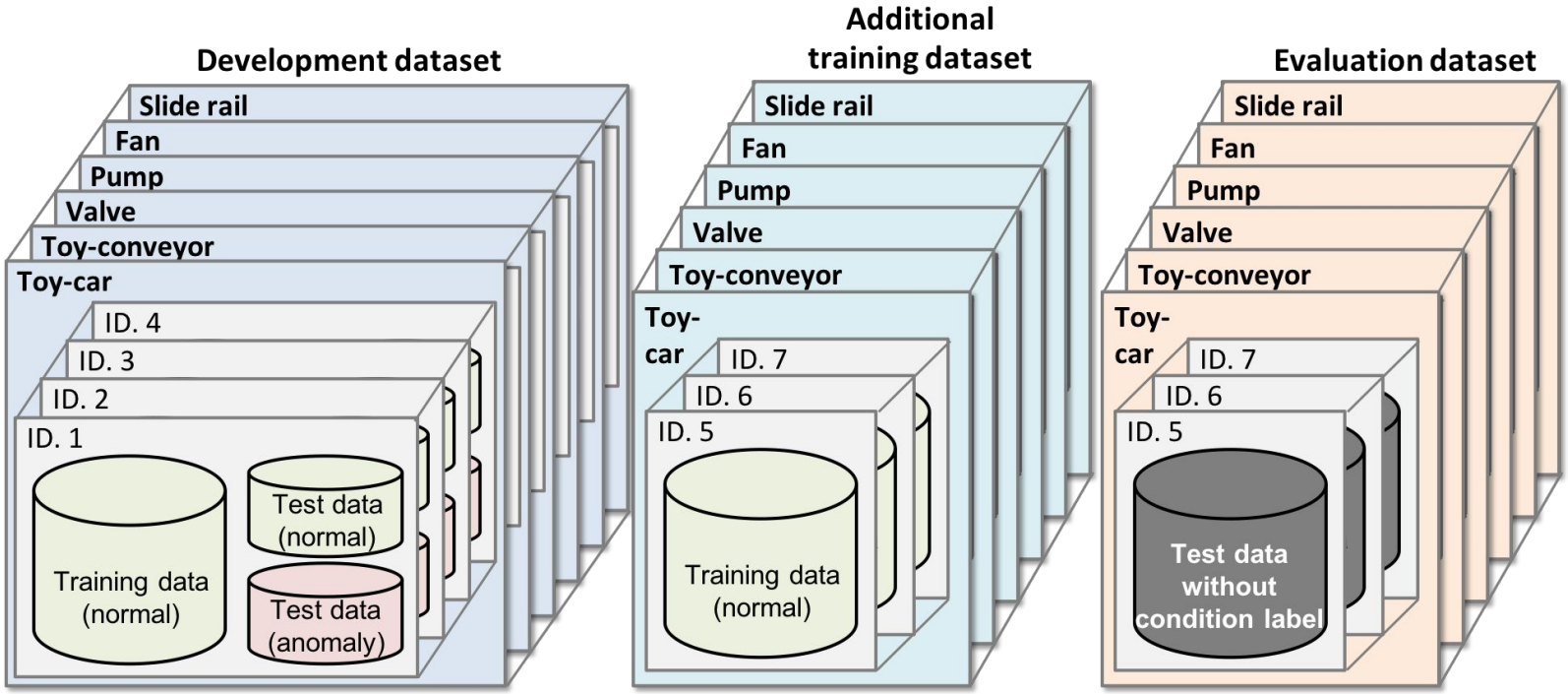


Fig 6. Development and evaluation dataset

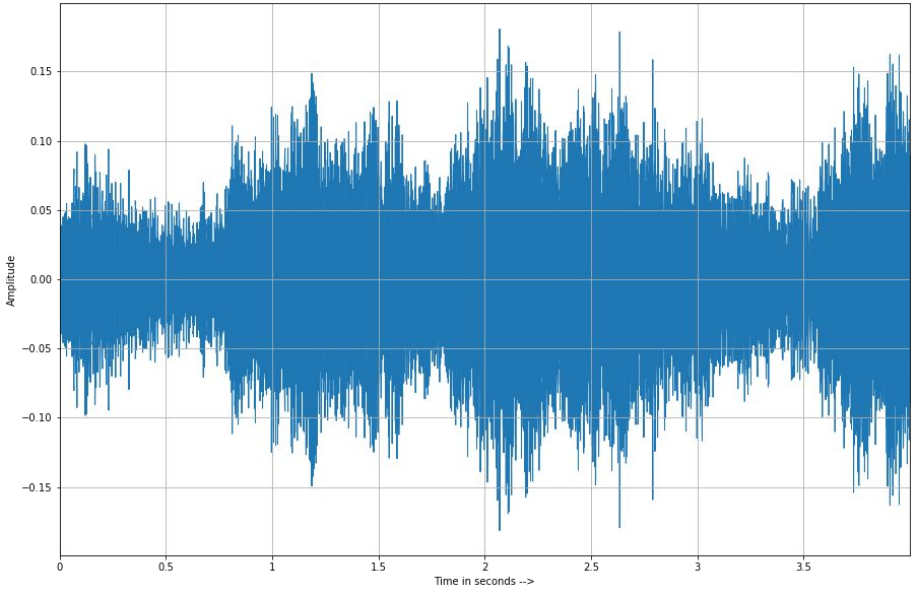


Fig 7. Time domain view of car horn sound sample

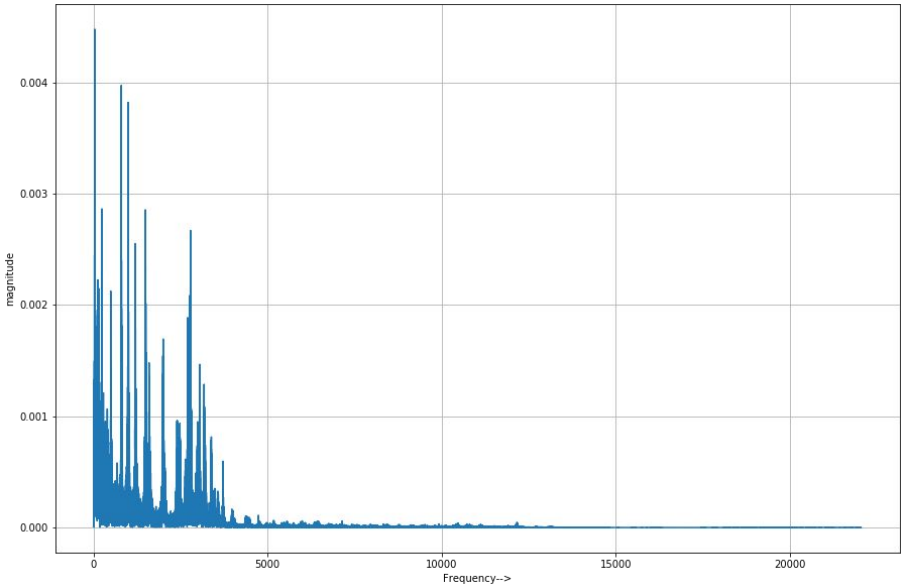
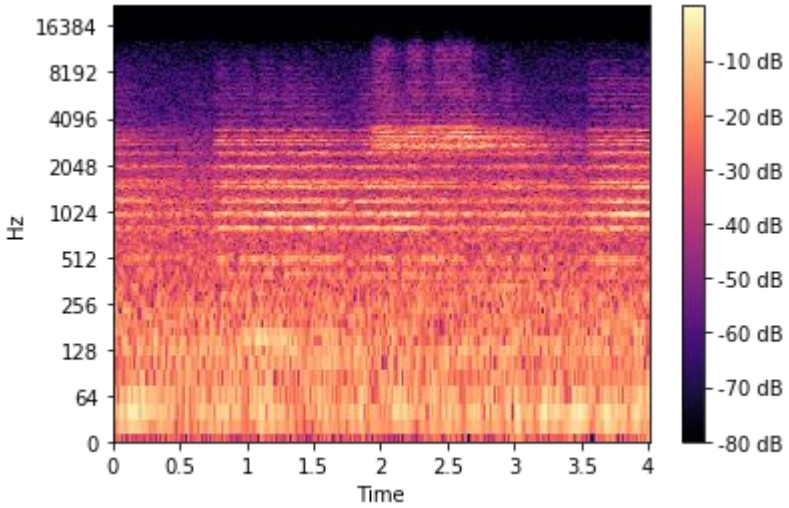


Fig 8. Frequency domain view of car horn sound sample



[8]

Fig 9. Spectrogram view of car horn sound sample.

Reference

1. Su, Y., Zhang, K., Wang, J., & Madani, K. (2019). Environment sound classification using a two-stream CNN based on decision-level fusion. Sensors, 19(7), 1733.
2. Lim, Hyungui, Jeongsoo Park, and Yoonchang Han. "Rare sound event detection using 1D convolutional recurrent neural networks." Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop. 2017.
3. https://urbansounddataset.weebly.com/urbansound8k.html
4. http://dcase.community/challenge2018/task-monitoring-domestic-activities
5. https://mivia.unisa.it/datasets/audio-analysis/mivia-audio-events/
6. https://zenodo.org/record/1160455#.W9muM1Vfi7A
7. http://dcase.community/challenge2020/task-unsupervised-detection-of-anomalous-sounds
8. https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0