



Proposal for the Research Grants Programme of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG)

General Information

Name: **Dr.-Ing. Jakob Abeßer**
Position: Wissenschaftlicher Mitarbeiter
Date of Birth: 03.05.1983
Nationality: German
Institution: Fraunhofer Institute for Digital Media Technology (IDMT)
Address (work): Ehrenbergstraße 31, 98693 Ilmenau
Phone: +49 3677 467288
Fax: +49 3677 467467
E-mail: jakob.abesser@idmt.fraunhofer.de
Address (private): Walter-Gropius-Straße 61, 99085 Erfurt
Phone: +49 176 61104420
Reference number of a previous DFG grant proposal: AB 675/2-1

Name: **Prof. Dr. Sebastian Stober**
Position:
Date of Birth:
Nationality:
Institution:
Address (work):
Phone:
Fax:
E-mail:
Address (private):
Phone:
Reference number of a previous DFG grant proposal:

Topic: Acoustic Scene Understanding via Listening, Understanding, and Adaptation (LUA)

Subject Area: Computer Science, Artificial Intelligence, Image and Language Processing

Keywords: Acoustic Scene Understanding, Audio Event Detection, Machine Learning

Duration (in months): 36 (New Proposal)

Summary: The development of computational methods for auditory scene analysis and machine listening has been actively researched over the last decades. One of the most challenging machine listening task is the sound event detection (SED), which involves a precise detection and classification of audio events that can be heard within a recorded acoustic scene. In environmental sound scenes, such events include chirping birds, rustling leaves, as well as human footsteps, whereas industrial scenes are often characterized by repetitive sounds originating from rotating machines, electric engines, or sawing devices. The rapid progress in the area of deep learning has led to the development of a multitude of novel data-driven algorithms in the

field of sound event detection in the last years. The applied deep learning techniques are often inspired from related research areas such as computer vision, natural language processing, and speech processing. While the human auditory system can easily decompose complex acoustic scenes to identify and focus on the most salient sound sources, algorithms driven by artificial intelligence face different challenges such as the large acoustic variability of real-life sound events, adaptivity towards previously unheard sound classes, as well as the required robustness towards changing acoustic conditions caused by recording hardware and the surrounding room acoustics.

In this project, we aim to develop methods to understand and model complex environmental or industrial sound scenes in three partial steps. In the **first** step, we want to derive a **coarse-level description of the temporal structure** of a given audio recording by **identifying segments with homogeneous timbre characteristics** and by **estimating the local degree of polyphony**, which describes how many sound events are simultaneously audible at a certain point in time. In the **second** step, we aim to derive a more detailed understanding of an acoustic scene by **detecting and classifying all individual sound events**. Furthermore, we will model the **typical order and repetition patterns of different sound events**. In order to improve the robustness against changes in the acoustic recording conditions, we will investigate methods for **domain adaptation**, which can be integrated into the AED algorithms. Due to the complexity of the task, we will use **explainable AI techniques to evaluate the meaningfulness of classification decisions**, particularly for polyphonic sound event classification scenarios. Finally, in the **third** step, we will develop methods to **recognize previously unheard sounds (acoustic anomalies)** based on different timbral and temporal novelty measures. **Continuous learning** techniques will then be used to integrate such sounds into the underlying sound classification model in order to mimic the human's capability to **continuously expand its known vocabulary of sounds**.

Project Description

1 State of the Art and Preliminary Work

State of the Art (2.5 pages)

Machine Listening Machine listening algorithms combine methods from audio signal processing and machine learning with the goal of mimicking the human's ability to perceive and analyze complex acoustic scenes. Each acoustic scene commonly includes multiple sound sources, which emit different acoustic events (sounds) over time. Machine listening algorithms find application in various scenarios such as context-aware hearing aids, acoustic condition monitoring in industrial settings, bioacoustic wildlife monitoring, IoT, and smart cities.

The two main directions of research in machine listening are acoustic scene classification (ASC) [1,2] and acoustic event detection (AED). While the goal of the former task is to summarize an entire recording using semantic descriptions such as "traffic" and "office", the latter task deals with the detection and classification of distinct sound events. In the following, the focus is on methods for acoustic event detection as they allow for a more detailed description of complex acoustic scenes.

Audio Representation Learning - using DNNs for feature learning, data compression
- Embedding approaches (openL3) [3]

- vggish, ...
- summarize outcome of grollmisch paper
- COALA [4] by aligning the learned latent representations of audio and associated tags

Audio Event Detection In this section, we will first discuss the main challenges of AED. Due to the large number of relevant publications, we will then summarize only the most relevant research trends in deep-learning based AED as published in the last three years. We refer the reader to two extensive overview articles compiled by Dang et al. [5] and Xia et al. [6] for a detailed summary of earlier AED approaches.

Computational methods face several challenges in the analysis of complex acoustic scenes. The first challenge is the **large variety of different sound characteristics**: Sound events of interest range from nature sounds like bird calls, rustling leaves, and rain drops, over machine-made sounds like motor engines, braking noises, or chainsaws, to human-made sounds like voice, laughter, or screams. Different sounds can be characterized as structured or unstructured, stationary or non-stationary, repetitive or without any predictable and repetitive nature. Also, sound events are composed of diverse acoustic building blocks such as short transients, noise, and harmonic signal components. As a second challenge, the **temporal boundaries of environmental sounds are often ill-defined**. This complicates both the sound-level annotation as well as the sound event detection. For example, the loudness of passing vehicles like cars or trains gradually increase in the beginning and continuously fade into the background noise level in the end.

As a third challenge, sound events appear either in the foreground or background depending on the position of a sound source in a given acoustic scene. If multiple sounds appear simultaneously, they **overlap and often blend into novel and unheard mixture sounds**. One promising approach to unravel such mixtures from multi-channel audio recordings is to apply sound localization techniques such as beam-forming, which aim to estimate the spatial position of sound sources in order to separate them (**REF**). In this project however, we will not investigate a spatial sound localization since most machine listening datasets that we aim to exploit only provide monaural sound recordings. The main reason is that single-channel or stereo recording setups are easier to setup and less expensive in most machine listening application scenarios. As a final challenge, the **durations** of acoustic events cover a **large range** from very short (gun shot, door knock, or shouts) to very long and almost stationary (running machines or climate sounds such as wind or rain). Since most AED methods analyze fixed sized audio segments, the classification of acoustic events is complicated, if the sound duration exceeds the analysis window size. As one possible solution, multi-scale approaches such as the AdaMD algorithm [7] use **multiple time resolutions** for classification simultaneously.

State-of-the-art AED algorithms consistently use deep neural networks to process audio recordings based on sample-level data (end-to-end) or based on derived two-dimensional representations such as the Short-time Fourier Transform (STFT) spectrogram or the mel spectrogram. The most common neural network architectures are convolutional neural networks (CNN) or convolutional recurrent neural networks (CRNN). Both architecture types include a front-end, where multiple convolutional layers are used to learn sound-specific features from the above-mentioned signal representations. As a back-end, CNN models use fully-connected layers for sound classification whereas CRNN models apply recurrent layers such as Gated Recurrent Units (GRU) or Long Short-Term Memory (LSTM) layers to model the temporal

progression of such features for acoustic event detection.

In the following, we will discuss several the most relevant recent advances on these basic neural network architectures for AED. In order to be able to classify longer sound events, the receptive field of CNNs can be increased by using dilated convolutions [8] or by replacing the convolutional and recurrent layers in a CRNN model by depthwise separable and dilated convolutional layers as shown in [9]. As a side effect, the number of model parameters and the training time can be reduced significantly.



Recurrent neural network focus on the most recent past in modelling sequential information due to their limited memory. Hence, their ability to model long-term context, which is crucial for classifying longer sound events, is limited. **As a solution, attention mechanisms were proposed in Natural Language Processing (NLP) to focus the classification to a defined time period within the signal [10].** Xu et al. introduce attention mechanisms at multiple positions of a CRNN network for AED [11]. In each layer of the model front-end, the output of an additional convolutional layer is used as **multiplicative gating** to allow to focus on particular regions on the intermediate feature maps and to control the information flow to the next model layer. A similar gating mechanism is added to the output of the recurrent back-end layer to allow for a temporal localization of sound events. Frame-level features of sound events such as successive spectrogram frames are highly correlated. In order to avoid that models with self-attention overfit to such irrelevant correlations and instead focus on sound-level relationships, the attention width needs to be constrained. Pankajakshan et al. propose a memory-controlled sequential self attention [12], where different attention spans are used and combined using a multi-head attention structure, which allows to perform AED on multiple time resolutions simultaneously.

Different alternatives to the common cascade of convolutional layers in the model front-end were developed in the AED literature. As one example, Ding and He propose the adaptive multi-scale AED (AdaMD) approach [7], which uses an hourglass network to compute and process features on multiple scales. The hourglass network uses a **decoder/encoder** structure to first reduce the input feature resolution to a low-dimensional representation and then gradually restoring it. Intermediate feature maps are used as input to four individual recurrent layers (bi-directional GRU), whose results are combined for the final prediction. This allows to train individual predictions which are specialized to detect sounds at different time-frequency scales.



AED Datasets are often imbalanced since some sound classes naturally occur less frequent than others. Transfer learning has shown to be a promising approach to pre-train networks on large dataset before finetuning them to smaller task-specific datasets. Often, intermediate feature representations (embeddings) are extracted using pre-trained models and used as input for AED classifiers. Here, given a CRNN model, one or multiple convolutional layers can be fixed after being pre-trained on a source task and the remaining layers can be fine-tuned on a new target task [13]. Other commonly used embeddings are the OpenL3 embeddings [3] as well as the embeddings proposed by Kumar et al. [14]. As shown by Grollmisch et al. in [15], such embeddings can also be successfully applied to classification tasks in related domains such as music information retrieval and industrial sound analysis.

Acoustic Anomaly Detection The task of anomaly detection (AD) aims at detecting deviations from a “normal” system behaviour by analyzing different data modalities such as time series data, audio, or video recordings. AD is highly relevant in many application scenarios ranging from medical heart monitoring [16] to predictive maintenance of industrial facilities

[17]. From the machine learning perspective, the biggest challenge of AD is that in most cases, only examples from the normal system state are available while recordings from the anomalous state are often hard to acquire. One of the main reasons is that it is hard to predict or too costly to simulate all possible fault types of a machine, which might occur in the future.

Chalapathy and Chawla provide a detailed overview over application scenarios and algorithmic approaches for anomaly detection in [18]. In section 2, we briefly introduce the general topic and later focus on methods which are most promising for the task of acoustic anomaly detection.

In most approaches, anomaly detection (AD) is implemented by combining distribution modeling and outlier detection or based on a prediction error. As examples for the former approach, the normal state data distribution can be modeled using techniques such as Gaussian Mixture Models (GMM) or Support Vector Machines (SVM) before the class likelihood of a novel data instance can indicate whether its an anomaly or not. In the second approach, an anomaly score is computed from the deviation between a measured quantity at a certain point in time and its prediction based on previous observations.

Time-series analysis techniques are most often used to detect local anomalies in long-term data recordings. Traditionally, methods such as are auto-regressive models, Kalman filters, and Hidden Markov Models (HMM) were used for time-series analysis. Nowadays, non-linear methods based on recurrent and convolutional neural networks are mostly used to derive predictions on future behaviour.


As an alternative to time-series approaches, autoencoder architectures use an encoder network to represent an input signal using a low-dimensional representation before reconstructing the original signal using a decoder network. If such an autoencoder is trained on examples from the normal state distribution, it is likely to fail on reconstructing anomalous signals and to return higher reconstruction error values. Popular variants are fully-connected, convolutional, as well as variational autoencoders. Suefusa et al. propose to interpolate the center frame from its previous and subsequent spectral frames [?], which has been shown beneficial particularly for AD in non-stationary machine sounds. ADD Dcase tech report and challenge [19]

Instead of performing AD on raw measurement data, low-dimensional signal representations are processed during the time-series analysis. For instance, Lin et al. [?] use a Variational Autoencoder (VAE) to first extract low-dimensional embeddings from short time series segments. Then, they use an LSTM network to predict the embedding of the next time frame and derive the anomaly score from the deviation between the measured and predicted embedding. Another way to reduce the data dimensionality is to initially perform a subsampling [?, ?].

Audio Segmentation - analogy to music information retrieval

- brief summary - traditional approaches
- representation learning (unsupervised learning using triplet loss [20])
- pre-processing step in semi-supervised annotation approaches for candidate segment selection [?]
-

Machine Listening Datasets As a general challenge in fields such as computer vision, natural language processing, and machine listening, large amounts of training data are required to capture and learn the natural variability in the data. For the tasks of acoustic event


detection, various datasets were collected over the last decade and published to stimulate further research. While most of the dataset are rather small and include only up to 20 different sound classes, the AudioSet [21] is a remarkable exception with around 2.1 million audio segments taken from a taxonomy of 527 sound classes. In the following, we want to highlight some of these datasets with importance to the LUA project. As another prominent example, the “Dataset for Environmental Sound Classification (ESC-50)”¹ include 2000 recordings from the categories animal sounds, nature sounds, human-made sounds, interior (domestic), as well as exterior sounds and was used in over 50 AED publications as evaluation dataset. 

In all runs of the annual “DCASE” evaluation campaign (Challenge on Detection and Classification of Acoustic Scenes and Events) since 2016, several tasks related to acoustic event detection in domestic and outdoor recordings were proposed with their unique datasets². As one example, the TUT Rare Sound Events dataset³ includes recordings of the rare and anomalous sound event classes “baby crying”, “glass breaking”, and “gunshot” mixed with different background scenes and can potentially be used as evaluate set for anomaly detection (WI4).

With a special focus on sound detection in urban acoustic scenes, both the MAVD dataset⁴ and the Urban Sound Dataset⁵ were recently proposed. The sound events include different vehicle types such as cars, busses, and motorcycles as well as particular machine sounds from jackhammers, air conditioner, or running engines.

In the field of industrial sound analysis, several audio datasets were published from Fraunhofer IDMT [17] including among others recordings of different machine types. The ToyADMOS dataset [22] and MIMII dataset [23] were recently published as public benchmarks for acoustic anomaly detection in industrial settings. In total, they cover over 30000 recordings of different operating states of vehicles such as toy cars, conveyors, and trains as well as machines such as valves, pumps, and fans.

The CHIME-HOME dataset⁶ includes 6.8 hours of audio recordings in domestic environment. As an example of a dataset for environmental monitoring, we want to mention the BirdVox-70k⁷, which provides annotated field recordings for bird migration monitoring.

Data Augmentation During data augmentation, novel data instances for training machine listening models can be generated to facilitate model training. A first class of algorithms synthesize novel data instances. The most common group of synthesis algorithms use Generative Adversarial Networks [24] which use an adversarial training strategy to imitate new examples from observed data. Data synthesis methods either synthesize new examples as audio signals [25, 26] or as intermediate embedding vectors [27]. As an alternative to GAN-based algorithms, the SampleRNN model architecture allows for end-to-end generation of environmental sounds as shown in [28]. 

A second class of data augmentation approaches transform existing data instances to create new ones. Such transformations include audio signal processing such as pitch shifting, time

¹<https://github.com/karolpiczak/ESC-50>

²For instance: TUT Sound events 2016, Development dataset <https://zenodo.org/record/45759>

³<http://dcase.community/challenge2017/task-rare-sound-event-detection>

⁴<https://zenodo.org/record/3338727>

⁵<https://urbansounddataset.weebly.com/>

⁶<https://archive.org/details/chime-home>

⁷<https://zenodo.org/record/1226427.Wt46UWaZO8o>

stretching, or adding different types of noise [29,30]. SpecAugment and random erasing are two techniques, which apply temporal warping and different kinds of noise masking to spectrograms cite [31,32]. The most often used approach to generate novel data instances in machine listening scenarios is mix-up data augmentation [33]. Here, two or multiple data instances as well as their targets are mixed in a certain range. This allows for instance to simulate acoustic scenes with foreground and background sounds. Johnson and Grollmisch combined both data augmentation and adaptive normalization techniques in an industrial sound analysis setting [34].

Domain Adaptation bla bla bla

Open Set Classification bla bla bla

Continuous / Lifelong Learning - Active Learning for AED [?]

Evaluation Measures - newly proposed polyphonic sound detection score

- f-measure, error rate

- check Mesaros 2016 applied science article and sed eval python package

-

Preliminary Work

Jakob Abeßer studied computer engineering (Diplom) and media technology (Ph.D.) at Ilmenau University of Technology. Since 2008, he has been working in the field of semantic music processing, audio signal processing, and machine learning at the Fraunhofer Institute of Digital Media Technologies (IDMT) in Ilmenau. From 2012 to 2017, he has been a researcher as part of the interdisciplinary Jazzomat Project⁸ at the Liszt School of Music, Weimar. Since 2018, he is a principal investigator of the DFG-funded project ISAD⁹, where one main research topic is to develop methods to detect and classify sound events in complex music recordings. Since 2018, Jakob Abeßer focused his research towards machine listening tasks such as acoustic scene classification, acoustic event detection, and anomaly detection and actively participated in related publicly-funded and industry projects.

As an example, we want to mention three studies that are closely related to the LUA-project. In [35], a distributed noise monitoring approach for urban environments is presented. The developed sensor units simultaneously noise level measurement as well as sound event detection on a hardware platform with limited computational performance. In [2], a detailed overview over almost 100 scientific publications on deep-learning based acoustic scene classification is provided. Among other aspects, the article discusses several topics such as domain adaptation, open-set classification, and attention-based neural networks, which are highly relevant for the project proposed here.

Since he started working at Fraunhofer IDMT, Jakob Abeßer supervised more than ... student theses. Also he was involved in the management and execution of various projects

⁸“Melodisch-rhythmische Gestaltung von Jazzimprovisationen. Rechnerbasierte Musikanalyse einstimmiger Jazzsoli” (PF 669/7-1)

⁹“Informed Sound Activity Detection in Music Recordings” (AB 675/2-1, MU 2686/11-1)

related to music and audio signal processing ranging from industry projects to publically funded projects.

Benefits for Collaboration

1.1 Project-Related Publications

2 Objectives and Work Programme

2.1 Anticipated Total Duration of the Project

36 months

2.2 Objectives

The overall objective of the LUA project is to develop techniques for analysis and understanding of acoustic scenes based on long-term audio recordings. We aim to develop an adaptive system for auditory scene analysis, which will be based on artificial intelligence methods. It will combine several algorithms for listening, understanding, and adaptation in order to detect and recognize the most salient sound events in a given acoustic scene. As two relevant scenarios, we will evaluate the developed techniques using both outdoor sound scenes captured in urban and environmental locations as well as indoor scenes captured in industrial settings such as factories and manufacturing lines. This approach will allow for a rigid evaluation as each scenario presents its own challenges in terms of common sound classes with unique sonic characteristics and typical background noises. We will address each partial step of the Listening/Understanding/Adaptation paradigm in separated objectives (O1), (O2), and (O3).

(O1) Listening Real-life acoustic scenes commonly include both stationary and moving sound sources emitting permanent or occasional sound events. In this project, the **focus is on clearly defined sound events of short and medium durations**, which can exhibit a wide range of acoustic characteristics such as short signal transients towards longer noise-like and harmonic signal components. In general, sound events often overlap which complicates their precise detection and classification. Therefore, we first aim to analyze the structure of a given acoustic scene recording on a coarse-level by identifying segments with homogeneous timbre characteristics. Furthermore, we aim to measure the time-localized sound polyphony, i.e., the number of simultaneously audible sound sources, as an additional cue to characterize sound scenes.

(O2) Understanding In the second objective, we aim to derive a deeper understanding of acoustic scenes by modeling two main aspects. As a first aspect, we aim to **detect and classify individual sound events** and **organize the corresponding sound classes in taxonomy-like structures based on timbre similarity measures**. Secondly, we want to model the regularity of appearance of certain sound events within recorded acoustic scenes by **investigating their specific temporal order and possible repetition cycles**. Most common sound event classification models are able to distinguish up to 20 sound event classes. In this project, we target large-scale sound taxonomies of between 100 and 200 sound classes. For this purpose, we will exploit and combine publically available machine listening datasets as will be further detailed in objective (O4). As most existing machine listening datasets contain either monaural or

stereo audio recordings, we will not investigate the task of spatial sound localization in this project.

Based on suitable training data, we will compare recently proposed deep neural network architectures based on convolutional recurrent neural networks (CRNN) and attention mechanisms w.r.t. their ability to classify the large variety of targeted sound classes. Parallel to the architectural choice of the model, suitable input representations of the analyzed audio signals need to be identified. Whereas most recent machine listening models use two-dimensional spectrum-based input features, a direct processing of audio samples in an end-to-end learning paradigm will be investigated as promising alternative. Furthermore, we will compare flat and hierarchical classification strategies. Here, we will exploit taxonomies, which we derive from relationships between different sound classes based on meaningful timbral similarity measures. Two approaches are planned to implement an ongoing quality control and sanity check of the sound recognition model to be trained. First, sound similarity measures and corresponding taxonomies will be validated by human listening tests. Second, we will investigate the meaningfulness of automatic sound classifications of the neural network based algorithm by applying techniques from the research field Explainable Artificial Intelligence (xAI). These methods for instance allow to identify parts of the input feature representation such as particular frequency bands, which have contributed most to a particular classification decision.

(O3) Adaptation In order to mimic the human’s capability of lifelong learning, we will investigate in the third objective how to extend the sound classification model in order to adapt itself to previously unheard sound scenes. Initially, we will test domain adaptation approaches to reduce the influence of differing acoustic conditions between existing machine listening datasets integrated in the training set as will described in objective (O4). Then, potentially novel sound events, which have not been part of the model’s training data before, need to be identified. For this purpose, we will compare two approaches. As first approach, we will first investigate the paradigm of open-set classification, where a classification model can express that it cannot classify a detected sound with high confidence. As a second approach, we interpret this task as anomaly detection and implement different novelty measures to identify novel and previously unheard sounds. Here, we will not just focus on timbral novelty but also on changes in the repetition patterns and appearance order of known sounds. Finally, the sound recognition model will be extended by continuous learning strategies in order to allow for a regular model re-training and extension of the sound taxonomy.

(O4) Data Handling The amount of sound event classes and their respective examples included in common machine listening datasets up to this day lag behind computer vision datasets such as the ImageNet dataset with over 21 thousand object classes [36]. In objective (O4), we will build up a large database of annotated audio recordings, which allow to train and evaluate the sound recognition model investigated in (O2). Special focus will be put on the questions, how to best combined datasets with different annotation schemes. Some datasets only provide weakly-labeled annotations, where sound events are annotated as tags without specifying their exact appearance time and duration. In contrast, strongly-labeled annotations provide the additional segment information for each sound appearance. As three main evaluation scenarios, we will construct a novel dataset of long-term real-life recordings from various urban, environmental, and industrial sound scenes, which will become the main


		Ilmenau	Joint	 Magdeburg
Y1	Q1	(WI1) Homogeneity-based Segmentation	(WJ1) Datasets & Data Augmentation	(WM1) Domain Adaptation
	Q2			
	Q3			
	Q4			
Y2	Q1	(WI2) Polyphony Estimation	(WJ2) Sound Event Detection & Classification	(WM2) Taxonomy Learning & Hierarchical Classification
	Q2			
	Q3			(WM3) Model Validation
	Q4			
Y3	Q1	(WI3) Temporal Sound Model		(WM4) Continuous Learning
	Q2			
	Q3			
	Q4			

Figure 1: Three-year project structure illustrates eight individual work packages of the two working groups in Ilmenau (IDMT) and Magdeburg (OVGU) as well as two joint work packages.

test data for the methods to be developed in objectives (O1) - (O3).

2.3 Work Programme Including Proposed Research Methods

Work Packages Ilmenau

(WI1) Homogeneity-based Segmentation In this work package, we aim to extract a coarse-level segmentation of long-term audio recordings into segments with homogeneous timbre characteristics. Such segments could for instance include stationary sounds such as rotating machines, passing trains, or noise-like sounds such as rain or running water. In Music Information Retrieval (MIR), similar attempts were made to identify segments such as verse or chorus in music recordings [37]. A common approach is to start with identifying points in time in which certain tonal or harmonic characteristics change. This task is referred to as boundary detection. In a first step, novelty curves are extracted, which measure how likely a segment change is at a certain point in time. In a second step, segments are annotated with the same or different labels based on pre-defined similarity measures.

As shown by McCallum in [20], siamese networks can be trained based on Constant-Q spectrograms of music recordings such that the learnt embeddings allow to compute novelty curves for boundary detection from. We want to transfer this approach of representation learning for the segmentation of environmental sound recordings. Due to the large variety of timbre characteristics, we will focus instead on processing either spectrogram-like representations such as Mel spectrograms or audio signals using end-to-end modeling approaches. McCallum uses an unsupervised training approach with the assumption that audio segments which are close in time are likely to sound similar. In contrast, we aim to study further supervised training objectives based on existing sound class labels and their relationship in sound taxonomies. Furthermore, we aim to study, how the temporal overlap of multiple sounds will affect the novelty curves and the segmentation results.

(WI2) Polyphony Estimation While the segmentation estimated in (WI1) describes the temporal structure of a recorded acoustic scene, the local number of simultaneously audible sounds (polyphony degree), provides a complementary view by measuring how the sound density change over time. Related tasks such as ensemble size estimation [38] and local polyphony estimation in piano recordings [?] were investigated in the field Music Information Retrieval (MIR). While pitched music instruments such as guitars or piano share at least a common harmonic overtone structure, environmental sounds cover a large variety of spectral characteristics. This makes the polyphony estimation in machine listening scenarios, which to the best of our knowledge is a previously unexplored task, particularly challenging.

In this work package, we plan to approach the polyphony estimation (PE) task by first performing a broad categorization of each segment identified in (WI1) into sound categories such as transient, harmonic, or noise-like sounds. Also, we plan to use unsupervised clustering approaches to identify particular local patterns in time-frequency representations of the acoustic scene, which can indicate distinct sound events or components thereof. Based on the global sound database, which will be collected in (WJ1), we will synthetically generate well-defined mixtures of different sound types with different degrees of polyphony, which will be for training and evaluating different algorithmic approaches for PE.

(WI3) Temporal Sound Model Most sound types appear multiple times within an acoustic scene. Examples for repeating sounds are bird singing, passing cars, or rotating machines. In this work package, we build upon the detected and classified sound events in (WJ2). We will investigate

- Learn typical temporal appearance / repetition patterns / order for different sound events

(WI4) Anomaly Detection - *Implement evaluate open-set classification / reject mechanisms to detect novel and previously unheard sound types*

In this work package, we aim to detect previously unheard sounds (timbre novelty) as well as known sound events, which occur in unusual order or temporal frequency of repetition (temporal novelty). For the first novelty type, we will quantify the timbre dissimilarity of a novel sound w.r.t. previously detected sound events (WP 1.2, WP 2.1). For the second novelty type, we use the STS model of a given acoustic scene as developed in WP 2.3. to predict the order and repetition frequency of future sound events. As baseline systems, will re-implement several state-of-the-art AAD algorithms¹⁰, which work on a frame-level instead of event-level. We plan to evaluate both strategies in different anomaly detection scenarios from both urban and industrial sound scenarios.

Work Packages Magdeburg

(WM1) Domain Adaptation

(WM2) Taxonomy Learning & Hierarchical Classification - Unsupervised vs. semi-supervised taxonomy creation (data-driven vs. Knowledge-driven) based on sound similarity

¹⁰Based on the results of the DCASE 2020 Challenge task 2 - Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring.

measures - Flat vs. hierarchical classification - check Bones et al 2018 Frontiers in Psychology [?]

(WM3) Model Validation - Defining & implementing task-specific evaluation measures

- novel evaluation measures for polyphonic AED [39]

- use XAI methods (LRP and alike) as sanity check, that relevant temporal-spectral components are recognized by the models (e.g. using synthetically created mixtures)

- for references: LRP on ISA [40] -> artificial unwanted bias added -> missbehaviour of model could be visualized

(WM4) Continuous Learning - Evaluate continuous / active learning approaches to

- o update the SED model by life-long learning of novel sound classes

- o adapt the sound taxonomies (for hierarchical sound classification)

Joint Work Packages

(WJ1) Datasets & Data Augmentation - Compilation and merging of various publically available machine listening datasets to global sound database

- annotation quality check

- Unifying different annotations schemes (weak vs. strong labels, single vs. multilabel) and audio formats (different sample rates, number of channels) - (what can we learn about sound taxonomies from analyzing multiple datasets?)

- DA techniques to create timbre variety, mix synthetic sound scenes etc...

(WJ2) Sound Event Detection & Classification - build upon large-scale sound database (see WP 4)

- Comparative study of audio representations for deep neural network based sound modeling (spectrogram-based vs. sample-based end-to-end)

- Comparative study of neural network architectures (CNN, CRNN, attention-based)

- Integration of dynamic temporal attention span (short vs. long sound durations)

- Assessment of data augmentation techniques to improve the robustness generalizability

- adaptive multi-scale detection (AdaMD) method [7] for multi-scale feature extraction

2.4 Data Handling

- use publically available data resources (datasets, annotations)

- what results do we plan to publish?

- website / demonstrators, source code, sound examples?

2.5 Other Information

2.6 Explanations on the Proposed Investigations Involving Experiments on Humans, Human Material or Animals

2.7 Information on Scientific and Financial Involvement of International Cooperation Partners

3 Bibliography Concerning the State of the Art, the Research Objectives, and the Work Programme

References

- [1] D. Barchiesi, D. D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic Scene Classification: Classifying environments from the sounds they produce,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.
- [2] J. Abeßer, “A Review of Deep Learning Based Methods for Acoustic Scene Classification,” *applied sciences*, vol. 10, no. 6, 2020.
- [3] J. Cramer, H.-h. Wu, J. Salamon, and J. P. Bello, “Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings,” in *IEEE ICASSP*, Brighton, United Kingdom, 2019, pp. 3852–3856.
- [4] X. Favory, K. Drossos, T. Virtanen, and X. Serra, “COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations,” 2020. [Online]. Available: <http://arxiv.org/abs/2006.08386>
- [5] A. Dang, T. H. Vu, and J. C. Wang, “A survey of Deep Learning for Polyphonic Sound Event Detection,” in *Proceedings of the International Conference on Orange Technologies (ICOT)*, Singapore, Singapore, 2017, pp. 75–78.
- [6] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang, “A Survey: Neural Network-Based Deep Learning for Acoustic Event Detection,” *Circuits, Systems, and Signal Processing*, 2019.
- [7] W. Ding and L. He, “Adaptive multi-scale detection of acoustic events,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 28, no. X, pp. 294–306, 2020.
- [8] Y. Li, M. Liu, K. Drossos, and T. Virtanen, “Sound Event Detection Via Dilated Convolutional Recurrent Neural Networks,” pp. 286–290, 2020.
- [9] K. Drossos, S. I. Mimilakis, S. Gharib, Y. Li, and T. Virtanen, “Sound Event Detection with Depthwise Separable and Dilated Convolutions,” 2020. [Online]. Available: <http://arxiv.org/abs/2002.00476>
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in *Proceedings of the Conference on Neural*

Information Processing Systems (NIPS), vol. 8, no. 1, Long Beach, CA, USA, 2017, pp. 8–15.

- [11] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, “Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 121–125.
- [12] A. Pankajakshan, H. L. Bear, V. Subramanian, and E. Benetos, “Memory Controlled Sequential Self Attention for Sound Recognition,” 2020. [Online]. Available: <http://arxiv.org/abs/2005.06650>
- [13] P. Arora and R. Haeb-Umbach, “A Study on Transfer Learning for Acoustic Event Detection in a Real Life Scenario,” in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Luton, UK, 2017, pp. 1–6. [Online]. Available: <http://ieeexplore.ieee.org/document/8122258/>
- [14] A. Kumar, M. Khadkevich, and C. Fugen, “Knowledge Transfer from Weakly Labeled Audio Using Convolutional Neural Network for Sound Events and Scenes,” in *IEEE ICASSP*, 2018, pp. 326–330.
- [15] S. Grollmisch, E. Cano, C. Kehling, and M. Taenzer, “Analyzing the Potential of Pre-Trained Embeddings for Audio Classification Tasks,” in *European Signal Processing Conference (EUSIPCO)*, Amsterdam, The Netherlands, 2020.
- [16] L. Ahrens, J. Ahrens, and H. D. Schotten, “A machine-learning phase classification scheme for anomaly detection in signals with periodic characteristics,” *Eurasip Journal on Advances in Signal Processing*, vol. 2019, no. 1, 2019.
- [17] S. Grollmisch, J. Abeßer, J. Liebetrau, and H. Lukashevich, “Sounding Industry: Challenges and Datasets for Industrial Sound Analysis (ISA),” in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, 2019.
- [18] R. Chalapathy and S. Chawla, “Deep Learning for Anomaly Detection: A Survey,” jan 2019. [Online]. Available: <http://arxiv.org/abs/1901.03407>
- [19] S. Grollmisch, D. Johnson, J. Abeßer, and H. Lukashevich, “IAEO3 - COMBINING OPENL3 EMBEDDINGS AND INTERPOLATION AUTOENCODER FOR ANOMALOUS SOUND DETECTION,” in *Detection and Classification of Acoustic Scenes Events*, 2020.
- [20] M. C. McCallum, “Unsupervised Learning of Deep Features for Music Segmentation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, UK: IEEE, may 2019, pp. 346–350. [Online]. Available: <https://ieeexplore.ieee.org/document/8683407/>
- [21] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio Set: An Ontology and Human-Labeled Dataset for Audio Events,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 776–780.

- [22] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, “ToyADMOS: A Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection,” no. Waspaa, pp. 2–7, 2019. [Online]. Available: <http://arxiv.org/abs/1908.03299>
- [23] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, “MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection,” sep 2019. [Online]. Available: <http://arxiv.org/abs/1909.09347>
- [24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Yoshua Bengio, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014, pp. 2672–2680.
- [25] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, “Deep Neural Network Based Learning and Transferring Mid-Level Audio Features for Acoustic Scene Classification,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 796–800.
- [26] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, “Integrating the Data Augmentation Scheme with Various Classifiers for Acoustic Scene Modeling,” in *Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019.
- [27] S. Mun, S. Park, D. K. Han, and H. Ko, “Generative Adversarial Networks based Acoustic Scene Training Set Augmentation and Selection using SVM Hyperplane,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*, Munich, Germany, 2017.
- [28] Q. Kong, Y. Xu, T. Iqbal, Y. Cao, W. Wang, and M. D. Plumbley, “Acoustic Scene Generation with Conditional SampleRNN,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 925–929.
- [29] J. Salamon and J. P. Bello, “Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification,” *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.
- [30] J.-X. Xu, T.-C. Lin, T.-C. Yu, T.-C. Tai, and P.-C. Chang, “Acoustic Scene Classification Using Reduced MobileNet Architecture,” *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, pp. 267–270, 2018.
- [31] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Sept, pp. 2613–2617, 2019.
- [32] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random Erasing Data Augmentation,” *Arxiv*, 2017. [Online]. Available: <http://arxiv.org/abs/1708.04896>
- [33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

- [34] D. Johnson and S. Grollmisch, “Techniques Improving the Robustness of Deep Learning Models for Industrial Sound Analysis,” in *European Signal Processing Conference (EUSIPCO)*, Amsterdam, The Netherlands, 2020.
- [35] J. Abeßer, R. Gräfe, C. Kühn, T. Clauß, H. Lukashevich, M. Götze, and S. Kühnlentz, “A Distributed Sensor Network for Monitoring Noise Level and Noise Sources in Urban Environments,” in *Proceedings of the 6th IEEE International Conference on Future Internet of Things and Cloud (FiCLOUD)*, Barcelona, Spain, 2018, pp. 318–324.
- [36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [37] M. Müller, *Fundamentals of Music Processing*. Springer, 2015.
- [38] S. Grollmisch, E. Cano, F. Mora-Ángel, and G. L. Gil, “Ensemble size classification in Colombian Andean string music recordings,” in *Proceedings of the 14th International Symposium of Computer Music Multidisciplinary Research (CMMR)*, Marseille, France, 2019.
- [39] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, “A Framework for the Robust Evaluation of Sound Event Detection,” pp. 61–65, 2020.
- [40] S. Grollmisch, D. Johnson, and J. Liebetrau, “Visualizing Neural Network Decisions for Industrial Sound Analysis,” in *Sensor and Measurement Science International (SMSI)*, Nürnberg, Germany, 2020.

4 Requested Modules/Funds

4.1 Basic Module

4.1.1 Funding for Staff

Scientific staff (full position, TVÖD 13, 36 months, Ilmenau). The staff member takes over the work in Ilmenau throughout the total duration of the project. *add tasks, required skills ...* A highly qualified candidate for this position is ...

Four student assistants (à 36 months à 30h/month, Ilmenau and Magdeburg)

4.1.2 Direct Project Costs

Equipment up to EUR 10,000, Software and Consumables

Travel Expenses

Visiting Researchers (excluding Mercator Fellows)

Expenses for Laboratory Animals

Other Costs

Project-Related Publication Expenses

5 Project Requirements

5.1 Employment Status Information

Abeßer, Jakob, Wissenschaftlicher Mitarbeiter, unbefristet (Fraunhofer IDMT, TVÖD 13, 80 %)

Stober, Sebastian, ...

5.2 First-Time Proposal Data

5.3 Composition of the Project Group

Research Group in Ilmenau

- Jakob Abeßer, Dr., TVÖD 13 (Wiss. Mitarbeiter, Fraunhofer IDMT)
- Sascha Grollmisch, PhD student (Wiss. Mitarbeiter, Fraunhofer IDMT)
- David-Scott Johnson, Dr. ((Wiss. Mitarbeiter, Fraunhofer IDMT)
- Hanna Lukashevich, Head of *Semantic Music Technologies* group (Fraunhofer IDMT)
- Stylianos Ioannis Mimilakis, PhD student, scholarship (Fraunhofer IDMT)

TODO: explain who will contribute with how many percent and with which topics / expertises

5.4 Cooperation with Other Researchers

5.4.1 Researchers with whom you have agreed to cooperate on this project

- Dr. Emmanouil Benetos, Dr. Dan Stowell (Machine Listening Lab, QMUL)
- Prof. Tuomas Virtanen, Dr. Konstantinos Drossos (Audio Research Group, Uni Tampere, Finland)
- Gael Richard (Télécom ParisTech)
- Meinard Müller (Audiolabs Erlangen)
- Prof. Mark Plumbley (Surrey)

5.4.2 Researchers with whom you have collaborated scientifically within the past three years

5.5 Scientific Equipment

5.6 Project-Relevant Cooperation with Commercial Enterprises

6 Additional Informaiton

No other application for funding of this project has been submitted. If we make such proposal, we will immediately inform the German research foundation.