MLPC 2025 Task 2: Data Exploration

Tara Jadidi, Florian Schmid, Paul Primus

April 2025

Context

Kepler Intelligent Audio Labs (KIAL), a soon-to-be-founded innovative AI startup, aims to collect a general-purpose dataset with strong temporal annotations to train sound event detection systems for their customer (see Figure 1 for a schematic illustration of sound event detection).

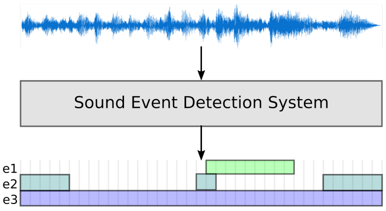

Figure 1: **Sound Event Detection (SED)** systems take an audio recording as input and predict acoustic events of interest ($e1, e2, e3$) *including* their temporal onsets and offsets.

To make this data set reusable for many potential future customers, KIAL decides to annotate sounds with free text annotations instead of relying on a fixed set of categories (see Figure 2). For a detailed description of the project, look at the project description on our Moodle page.

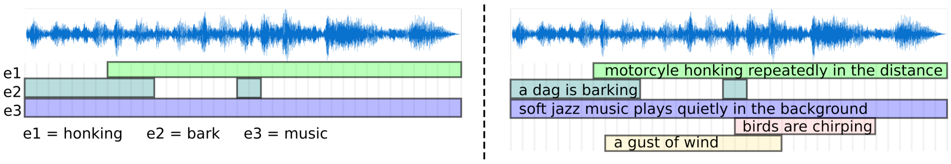
e1
e2
e3
e1 = honking
e2 = bark
e3 = music
motorcyle honking repeatedly in the distance
a dag is barking
soft jazz music plays quietly in the background
birds are chirping
a gust of wind

Figure 2: *Left:* Typically, SED data sets are only annotated for a fixed set of labels ($e1, e2, e3$). *Right:* Instead of relying on a fixed set of labels, KIAL decided to use arbitrary free text descriptions. Text annotations can be mapped to labels of interest ($e1, e2, e3$) by matching synonyms or using word embedding models.

KIAL has divided the project into these five stages:

1. **Data Collection:** Collect a large dataset of audio recordings, containing a large variety of potential target sounds and events. (✓)
2. **Data Annotation:** Once the data is gathered, human annotators carefully annotate sounds in the recordings with textual descriptions and temporal onsets and offsets. (✓)
3. **Data Analysis:** The dataset undergoes a thorough examination through exploratory data analysis, allowing for a deeper understanding of its characteristics and potential challenges before further processing. (← we are here)
4. **Model Training & Selection:** KIAL's R&D team needs to demonstrate the usefulness of the new approach to the managers by *training machine learning models* for one or multiple fictitious customers (each customer is represented by a fixed set of labels).

5. **Challenge:** A customer requests a sound event detection system for a fixed set of labels. Provide predictions on a hidden test set to prove to the customer that the developed system has a high detection performance and win an (also fictitious) contract.

KIAL has already collected the dataset by scraping publicly available audio recordings on the web (step 1) and annotated the recordings with textual annotations (step 2).

Task Outline: Data Exploration (25 points)

Deadline: Thursday, April 24th, 23:59
Submission: Submit your team's report and slides via Moodle.
Group Work: Talk to your team early and decide on a distribution of workload, a schedule, and deadlines. *Check in often.* Plan enough time to help your colleagues if they struggle to complete a task. Let us know early if one of your teammates is unresponsive.

This task is mandatory. Content and formal requirements are outlined below.

1 Report (max. 22 points)

For the first part of this assignment, you will have to write a report based on the template provided on Moodle. Your report needs to discuss *all of the following points*:

1. **Case Study** (2 points): Find two interesting recordings with at least two annotators and multiple annotations. Compare the temporal and textual annotations, and try to answer the following questions:
   (a) Identify similarities or differences between temporal and textual annotations from different annotators.
   (b) To what extent do the annotations rely on or deviate from keywords and textual descriptions in the audio's metadata?
   (c) Was the temporal and text annotations done according to the task description?
2. **Annotation Quality** (6 points): Use *all audio recordings annotated by multiple annotators* to address the following points quantitatively.
   (a) How precise are the temporal annotations?
   (b) How similar are the text annotations that correspond to the same region?

   Use *the complete data set (or a subset)* to address the following points quantitatively.

   (a) How many annotations did we collect per file? How many distinct sound events per file?
   (b) How detailed are the text annotations? How much does the quality of annotations vary between different annotators?
   (c) Are there any obvious inconsistencies, outliers, or poor-quality annotations in the data? Propose a simple method to filter or fix incorrect or poor-quality annotations (e.g., remove outliers, typos, or spelling errors).
3. **Audio Features** (6 points): Load and analyze the audio features:
   (a) Which audio features appear useful? Select only the most relevant ones or perform a down projection for the next steps.
   (b) Extract a fixed-length feature vector for each annotated region as well as for all the silent parts in between. The most straightforward way to do this is to average the audio features of the corresponding region over time, as shown in the tutorial session.
   (c) Cluster the audio features for the extracted regions. Can you identify meaningful clusters of audio features? Do the feature vectors of the silent regions predominantly fall into one large cluster?

4. **Text Features** (6 points): Load and analyze the text features of the annotations (file: `annotations_text_embeddings.npy`):
   (a) Cluster the text features. Can you find meaningful clusters?
   (b) Design a labeling function[1] for classes *dog* and *cat*. Do the annotations labeled as dog or cat sounds form tight clusters in the text and audio feature space?
   (c) How well do the audio feature clusters align with text clusters?
5. **Conclusions** (2 points): What conclusions can you draw from your analysis for the next phases of the project?
   (a) Is the dataset useful to train general-purpose sound event detectors?
   (b) Which biases did we introduce in the data collection and annotation phase?

The report **must not exceed a page limit of 6 pages**, of which **in total at most 4 pages should be text**.

2 Statement of Contributions

In addition to addressing these questions above, **add a statement of the contributions of all team members** as indicated in the template.

3 Slide Set (3 points)

The complementary slide set should present the results from your written report in a concise manner. More precisely, you will have to answer the questions corresponding to one of the sub-topics outlined in the previous section. The specific topic is determined based on the first letter of your group name, i.e. A for Team Aberrant, or B for Team Bed. To find your topic, determine the according letter, and find your topic in the following list:

First letter of group name	Topic
A, C, E, M, Q	Case Study & Conclusion
B, F, I, L, N, P	Annotation Quality
D, G, J, R, T, U, W	Audio Features
H, K, O, S, V, Y, Z	Text Features

Table 1: Assignment of topics for the slides based on the first letter our your group name.

The **upper limit for the number of slides you should prepare is 4** (excluding an additional title slide that should contain your group name and the member names).

4 Dataset

The dataset download links are available on Moodle. The format and content of the dataset are described in detail in the slide deck of the corresponding classroom meeting. Please refer to that slide deck for information on the audio features and the file formats. Have a look at the slide deck of our first tutorial session to see how the dataset can be loaded.

[1] A function that takes a class of interest and a textual annotation (or text features) as input and returns `True` if the annotation refers to the given class and `False` otherwise.

5 Grading

Each of the five topics given in the task outline above will be evaluated according to the following criteria:

- Thoroughness & Completeness: Have you thought about the problem, and answered every question?
- Clarity: Are the ideas, features, algorithms, and results described clearly? Based on your descriptions, could the reader reconstruct your experiments?
- Presentation: Did you select an appropriate way of communicating your results, e.g., did you use meaningful plots where helpful?
- Correctness: Is the proposed procedure/experiment sound, correct?
- Punctuality: The reports must be submitted in time. Any delay will result in reduced grades. Specifically, submitting on the day after the deadline will deduct 1/3 of the points, submitting on the second day after the deadline will deduct 2/3 of the points; submissions on the third day after the deadline or later will be rejected.

You will be awarded all points for the slide set if it addresses the assigned topic, is within the slide limit, and is submitted before the deadline.

6 Summary

- Completing Task 2 is a requirement to pass this course.
- Look at the given questions and answer all of them appropriately in a written report. Make sure to use the report template provided to you via Moodle. Adhere to the given page limit (max. 6 pages where at most 4 pages can be text) and include a statement about the contributions of all team members
- Create a set of slides tackling the questions of one of the topics. The topic is determined by the first letter of your group name. Make sure to adhere to the slide limit for this step as well (max. 4 + 1 title slide).
- Upload the written report as well as your slides to Moodle before the deadline.
- You will get a maximum number of 22 points for your written report, and 3 points for the slide set.