MLPC Project 2025

Tara Jadidi, Florian Schmid, Paul Primus

March 2025

The (Fictitious) Story

Kepler Intelligent Audio Labs (KIAL), a soon-to-be-founded innovative AI startup, specializes in developing *acoustic sound event detection* technology for *a diverse range of customers*.

Sound Event Detection (SED) systems analyze audio recordings to detect acoustic events and their temporal onset and offset. These systems are useful for various application scenarios, including:

- **Healthcare:** Hospitals might use it to monitor patients by detecting events such as coughing, sneezing, crying, shouting, hiccupping, or snoring.
- **Transport & Logistics:** Harbor operators could implement an automated surveillance system to recognize sounds such as cars, trucks, trains, ships, helicopters, fire alarms, or sirens.
- **Smart Homes:** Home automation systems can benefit from detecting speech, music, running kitchen appliances, and other household sounds.

Figure 1 illustrates how KIAL's model should detect domestic events in an audio recording for a smart home application.

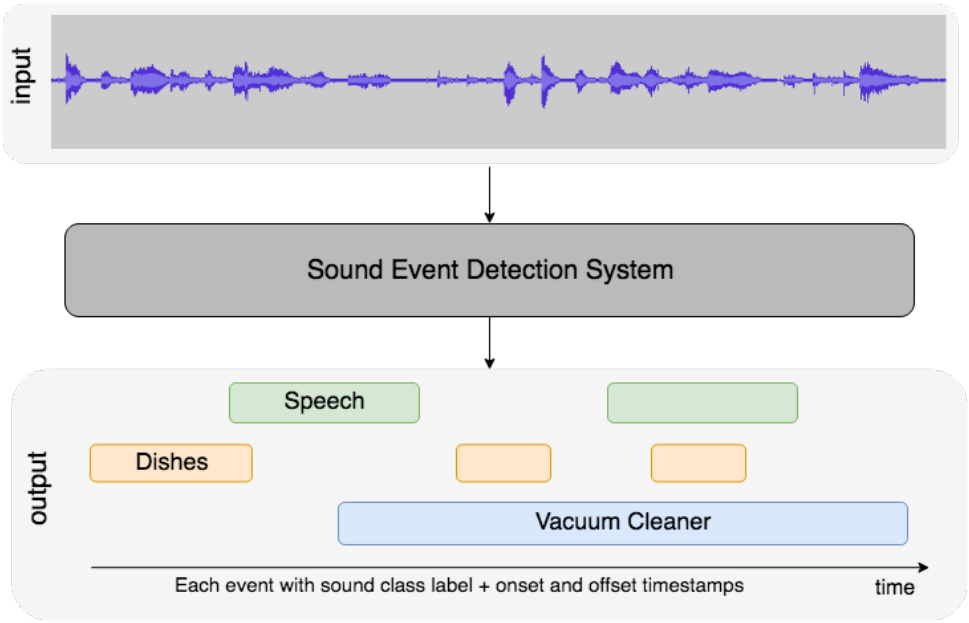

Figure 1: **Sound Event Detection (SED)** systems take an audio recording as input and predict acoustic events of interest (here: speech, dishes, and vacuum cleaner) *including* their temporal onsets and offsets. Figure taken from the DCASE website.

The Typical Approach

KIAL's approach (so far) to develop SED systems for their customers follows a well-established procedure, which can be divided into five major phases:

1. **Data Collection:** The process begins with the customer defining and providing the sound events of interest, which serve as the to-be-predicted labels of the system. A large dataset of audio recordings, which contains the target events, is then collected.
2. **Data Annotation:** Once the data is gathered, human annotators carefully annotate the recordings by listening to them, identifying the relevant sound events along with their corresponding labels, and marking their start and end times.
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.
4. **Model Training & Selection:** Once a dataset of sufficient size and quality has been collected, the experts at KIAL will use it to *train machine learning models*. Without going into too much detail, this involves extracting features from the recordings, optimizing models to map input features to the correct sound event labels, and selecting the best model.
5. **Challenge:** KIAL is typically required to *provide predictions on a secret test set*, to give the customer an unbiased and comparable estimate of the developed model's detection performance.

The Problem with the Typical Approach

The collected audio recordings often contain a variety of sounds, including sounds that are not of interest to a particular customer. However, in the labeling procedure outlined above, *sounds that are not relevant to a specific customer, will be ignored by the annotators.* This means that these datasets *cannot be reused to train models for a different set of sound events* (e.g., for a new customer from another domain).

Example: A data set labeled for traffic noise detection (e.g., car honking) that contains sounds of people talking could also be useful to train speech detectors. However, since the annotators did not annotate the speech (the original customer was only interested in traffic sounds), the complete data set needs to undergo full relabeling, which is inefficient and costly.

Towards A More General Dataset

Observing this issue, the head of KIAL's machine learning team proposed a novel idea:

> **What if the data was annotated with textual descriptions for *every sound occurring in the scene*?** To develop a sound event detection system for a new customer, we can then create a training set by selecting those annotations that are aligned with the labels of interest.

Example: Consider an audio snippet annotated for two events:

- $[1.43, 19.2]$: "Two people are having a conversation."
- $[7.91; 10.7]$: "A car is honking repeatedly."

This snippet could then be useful to train models for customers interested in detecting either honking or speech events.

To create a training data set with labels, the text annotations need to be mapped to the labels provided by the customer. KIAL is planning to explore two strategies:

- **Simple**: Use a list of synonyms for the target labels. If a synonym is found in an annotation, map it to the target label; e.g., for speech detection, we could select all annotations containing words like 'conversation', 'speech', 'talking', etc.
- **Advanced**: Compute a similarity score between the target labels and the textual annotations using word or text embedding models [1]

Figure 2: Label Sudio annotation interface to annotate sounds in an audio recording with arbitrary textual descriptions.

Project Milestones

KIAL invites you, as a machine learning specialist, to help with the development of sound event detection systems based on a large general-purpose dataset.

You and your team will work on this project in four phases:

- **Annotation (max 10 points):** This phase is conducted individually. Using Label Studio (Fig. 2), we will collaboratively annotate sounds with free-text descriptions and temporal information. The resulting dataset will serve as the foundation for the subsequent phases of the project.
- **Data Exploration (max 25 points):** Analyze our custom dataset to uncover meaningful patterns. Can the recordings be clustered? Is there a correlation between the discovered clusters and the text embeddings? To support this analysis, we provide precomputed features, temporal annotations, textual descriptions, and their corresponding text embeddings.
- **Training a Model Based on Inferred Labels (max 40 points):** A customer presents a set of acoustic events they wish to detect. Leveraging insights from the data exploration phase, you will select training and test examples and develop sound event detection models.
- **Challenge (max 30 points):** The customer ask you to provide predictions on a secret test set. In this phase, you will refine your system and evaluate its performance on a hidden test set containing precise annotations.

[1] https://en.wikipedia.org/wiki/Word_embedding