Case Study

Ali Doğukan Seven, Yisong Tang, and Karel Štícha

Johannes Kepler University

1 The Case Study

1.1 A) Identify Similarities and Differences

Temporal Differences: In 44534, clear sounds were not annotated, such as barking and car noise (00:06–00:20). Annotations in 623746 did not include or oversimplified repeating synthetic sounds and the full utterance from the woman (00:03-00:12).

Textual Differences: Nearly all of the major sounds, such as flute melodies, barking, and singing, were disregarded. The emotional tone of the woman's voice (e.g., excitement) was not noted. (Unclear whether the dogs were aggressive due to distance.)

Similarity: Both annotators could have been subjected to listener-related biases toward the primary sounds, often missing background events entirely, thus resulting in different interpretations of the task.

B)Assessing Audio Metadata In 44534, only panpipe is noted, while the metadata mentions terms like "shout" and "salesman," besides other sounds like dog barks and cars are absent in the annotation therewith. For 623746, the annotation states "a woman says one word," ignoring all the singing and synthesized sounds; while metadata do suggest emergency ("alarm," "fire"), it does not get enough support from the recording itself. Some language confusion might have arisen with respect to above.

C) Assesing Text Annotation Both recordings feature weak temporal and contextual annotations. In 44534, some car sounds, dog barks, and flute melody noises are not duly timestamped. In 623746, the changes of the sound phase and the emotion in the voice are omitted. In general, the annotations are overly vague and do not fulfill the expectations of the task.

2 Conclusions

The dataset is of limited utility for general-purpose sound event detection, especially in determining whether or not an event occurred. Its high variation regarding the lengths of time events are labeled, combined with vague textual descriptions, makes it unreliable for using the dataset for more specific purposes.

Biases include a fondness for short annotations, mixed wording, and an absence of attendance with background or emotional sounds. One-word annotations are common, but they are often uninformative, limiting the usefulness of the dataset.