## 2. Annotation Quality (6 points): Use the audio recordings annotated by multiple annotators to answer the following questions:

### (a) How precise are the temporal annotations?

A group of graphs showing different types of annotation

AI-generated content may be incorrect.

Figure 2.1.1 – Temporal Annotation Differences

Temporal annotation precision (Figure 2.1.1) was evaluated by comparing the onset and offset timings of overlapping regions from different annotators on the same audio file, considering only segments where a temporal overlap was present. The key metrics were Euclidean distance, onset and offset differences, and their means.

The results show a distinct spike around 0 sec, meaning the majority of annotators strongly agree on start and end times. All four distributions show a long positive skew. While the onset and offset distributions show consistent gradual decay starting before the 5 second mark, the Euclidean distance analysis highlights a secondary hump in the 12-20 second range, indicating a stronger disagreement pattern here.

The results point to high general temporal precision, however, a significant amount of annotations do differ in timing, pointing to possible ambiguity in some of the audio that makes it harder to segment.

### (b) How similar are the text annotations that correspond to the same region?

A screenshot of a calculator

AI-generated content may be incorrect.

Table 2.1.2 - Similarity of Overlapping Text Annotations Stats

A graph of a bar graph

AI-generated content may be incorrect.

Figure 2.1.2 – Similarity of Overlapping Text Annotations

Most overlapping annotations show low textual similarity across different annotators. The cosine similarities ranges from -0.29 – 0.75, with a mean of 0.088 and a median of 0.072, indicating that many annotators differ in their wording or focus (see Table 2.1.2). The histogram (Figure 2.1.2 – Similarity of Overlapping Text Annotations) confirms this by showing a concentration of values between 0 and 0.2, suggesting generally weak agreement in textual descriptions even when annotators mark the same time span.

## 2. 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?

Please see Table 2.2.1 – Per File Measures for per file statistics, and figures Figure 2.2.1.1 – Annotations Per File and Figure 2.2.1.2 – Distinct Sound Events Per File.

|  |  |  |
| --- | --- | --- |
|  | Annotations/File | Distinct Sound Events/File |
| Total Unique Files | 9026 | 9026 |
| Total Annotations | 35826 | 35826 |
| Mean | 3.9692 | 2.4253 |
| Median | 2 | 2 |
| STD | 4.4254 | 1.9063 |
| Min | 1 | 1 |
| Q1 | 1 | 1 |
| Q2 | 2 | 2 |
| Q3 | 5 | 3 |
| Max | 96 | 27 |

Table 2.2.1 – Per File Measures

A graph with blue and white bars

AI-generated content may be incorrect.

Figure 2.2.1.1 – Annotations Per File

A graph of a number of events

AI-generated content may be incorrect.

Figure 2.2.1.2 – Distinct Sound Events Per File

### (b) How detailed are the text annotations? How much does the quality of annotations vary between different annotators?

|  |  |
| --- | --- |
|  | Text Detail |
| Total annotations | 35826 |
| Mean word count | 7.4874 |
| Median | 7 |
| STD | 4.6315 |
| Min | 1 |
| Q1 | 4 |
| Q2 | 7 |
| Q3 | 9 |
| Max | 88 |

Table 2.2.2.1 – Text Annotation DetailA graph of a number of words

AI-generated content may be incorrect.

Figure 2.2.2.1 – Word Count Per Annotation

**How much does the quality vary between annotators?**  
A screenshot of a computer

AI-generated content may be incorrect.

Table 2.2.2.2 – Word Count Per Annotator

A graph of a number of words

AI-generated content may be incorrect.

Figure 2.2.2.2 – Average Word Count Per Annotator

The annotations have an average word count of 7.85 word (see Table 2.2.2.2 – Word Count Per Annotator). The histogram in Figure 2.2.2.2 – Average Word Count Per Annotator shows a concentration between 5-10 words (a reasonable amount considering the annotation guidelines), however some annotators averaged over 20 words. The positive skew is indicative of this high variability.

A screenshot of a computer

AI-generated content may be incorrect.

Table 2.2.2.3 – Annotation Duration Stats Per Annotator A graph of an annotation

AI-generated content may be incorrect.

Figure 2.2.2.3 – Average Annotation Duration Per Annotator

Most annotators marked events lasting in the 5-10 second range, with an overall average of 8.38 sec (Table 2.2.2.3 – Annotation Duration Stats Per Annotator). The distribution’s long tail indicates some annotators consistently marked longer events, ranging over 20 seconds, indicating a difference in how annotators perceive events. (see Figure 2.2.2.3 – Average Annotation Duration Per Annotator) The standard deviation of 3.31 seconds indicates moderate variation in the average durations used by 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).

Annotation duration, word count, and spelling errors were the indicators used to detect inconsistencies, outliers, and poor-quality annotations. Outlier thresholds were set both by IQR and explicit selection, following best practices and empirical testing. (Table 2.2.3.1 - Duration and Word Count Metrics)

|  |  |  |
| --- | --- | --- |
|  | Duration | Word Count |
| Count | 35826 | 35826 |
| Mean | 7.3139 | 7.4874 |
| Median | 2.6609 | 7 |
| STD | 8.7611 | 4.6315 |
| Min | 0 | 1 |
| 1% | 0.1323 | 2 |
| 5% | 0.2721 | 2 |
| Q1 | 0.9712 | 4 |
| Q3 | 12.6279 | 9 |
| 95% | 26.1306 | - |
| 99% | 29.2635 | - |
| Max | 30.0447 | 88 |
| IQR | 11.6567 | 5 |
| IQR lower threshold | -16.5138 | -3.5 |
| IQR upper threshold | 30.113 | - |

Table 2.2.3.1 - Duration and Word Count Metrics

Event duration markers found no IQR lower outliers, therefore the central range is wide and tolerant. Explicit 1%/99% thresholds flagged 0.99%/1% of annotations respectively, while 5%/95% thresholds flagged 5% each, showing that most annotations are reasonable, with few consistent outliers. (See Table 2.2.3.2 – Duration Outcomes)

|  |  |  |  |
| --- | --- | --- | --- |
|  | IQR | 99% | 95% |
| Text too short (annotations) | 0 | 355 | 1792 |
| Text too long (annotations) | 0 | 359 | 1792 |
| Text too short (%) | 0% | 0.9909% | 5.0020% |
| Text too long (%) | 0% | 1.0021% | 5.0020% |

Table 2.2.3.2 – Duration Outcomes

Text description length, identified by word counts per annotation also found no lower outliers by IQR threshold. Percentile thresholds flagged 0.94% of annotations under the fifth percentile. Explicit counts found no zero-word annotations, 335 one-word, 2024 two-word, and 3146 three-word annotations. A total of 9537 annotations, marking 26.62% of the population were under the five-word threshold, arbitrarily found to be ideal for describing annotations based on the guidelines, potentially marking a large portion of the population as poor.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1% | 5% | IQR Lower | Explicit 5 |
| Word threshold | 2 | 2 | -3.5 | 5 |
| Annotation Count | 335 | 335 | 0 | 9537 |
| Annotation Proportion | 0.9351% | 0.9351% | 0% | 26.6203% |

Table 2.2.3.3 – Word Count Outcomes

Spelling was measured with Python’s pyspellchecker module to gauge effort and commitment to the task. 49.56% of annotations had no spelling errors. 32.3% had one misspelled word, nearly 15% had 2-3 misspelled words, leaving around 3% with 4 or more misspelled words.

Combining all three criteria, with duration (below 1% or above 99%), short text (≤2 word count), and a high misspell count (≥3), 15.25% of annotations were flagged as having poor quality. Using a 5% duration threshold raised this to 22.14%.

It is important to note that after filtering for files with multiple annotators, each annotator has on average annotated only 4.5 files with a STD of 2.59 files. A minimum and maximum of 1 and 27 respectively are indicative of a positive skew, and an overall small dataset that may provide inconclusive results.

Proposition:

Spelling errors can be fixed programmatically, flagging the annotation for review. Word count and duration outliers can be flagged in the same manner. A random sample of this subset will reveal the possible necessity of reannotating these flagged files. If necessary, the group of files can be reassigned to their respective annotators.