Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

MLPC Report - Task 2: Data Exploration

Team OBSERVE

Johannes Grafinger **Jonas Gantar** **Leonhard Markus Spanring** **Reinhard Josef Pötscher**

Contributions

Reinhard and Johannes (Group 1) were responsible for tasks 1.) Case Study and 2.) Annotation Quality
Leonhard and Jonas (Group 2) were responsible for tasks 3.) Audio Features and 4.) Text Features of the report
All of us together were responsible for task 5.) Conclusions
In the same constellation we created this report. We all worked on the presentation together, with Group 2 providing its content. We held regular meetings, where each group presented their results up to that point and the other critically reviewing their work.

1.) Case Study

To find 2 interesting records that were edited by multiple annotators, we first looked in "metadata.csv" to see which files had more than one annotator. This resulted in a list of 149 files.
We then looked at the "metadata_title_embeddings.npz" and the "metadata_keywords_embeddings.npz" in order to be able to draw some conclusions. At the same time, we also checked the standardization of the embedding.

The approach was that files with very clear embedding (target: 1.0 in one place) lead to very clear annotations. However, this did not produce any useful results.

An important assumption here is the correctness and accuracy of the titles and keywords.

The next approach used the file "annotations.csv" and the corresponding "annotations_text_embeddings.npz". All annotations with more than one annotator were searched for. This resulted in a list of 731 annotations and 1468 annotations in total. There were 6 files with 3 annotations each.

From the associated text embeddings, we compared all embeddings of an annotator and an annotation with all other annotators by distance and subsequently removed 2 files. The file with the largest deviation is '568273.mp3' and the file with the smallest deviation is '203149.mp3'.

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

a.) How precise are the temporal annotations?

Since we did not have any ground truth for when events occur within the files, we simply compared temporal differences of annotations from different annotators corresponding to the same region. A pair of annotations were said to correspond to the same region if both their respective onset and offset times separately do not deviate by more than 0.5 seconds. This of course introduces a trade-off between False Positives and False Negatives. Then, the absolute differences in onset-/offset times and durations were gathered and plotted in histograms, shown in Figure 1.

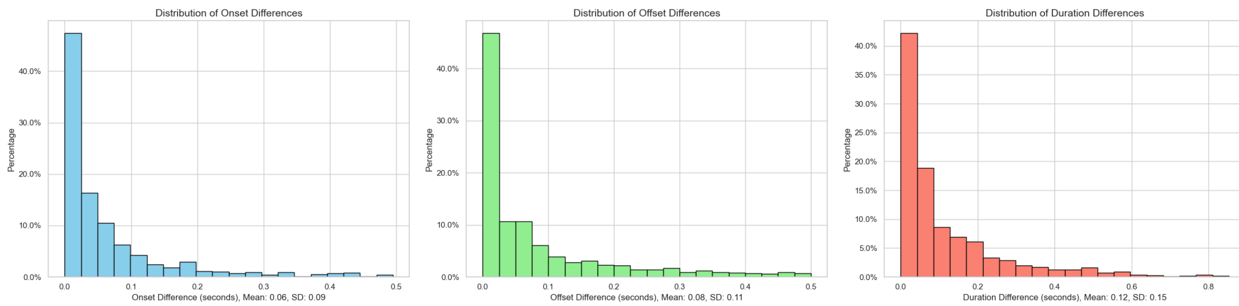

Figure 1: Absolute differences in onset-/offset times and durations for annotations corresponding to the same regions.

We can see that most annotations fall below a difference in onset-/offset times below 0.1 seconds. Depending on the temporal resolution of the audio features, this might actually lead to some falsely predicted frames in the future. Overall, most annotations seem to be pretty accurate.

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

Using the same criterion as described above, we collected pairs of annotations corresponding to the same regions. For each of these pairs we then simply fetched their annotation text embeddings and calculated their cosine-similarity. The results can be seen in Figure 1.

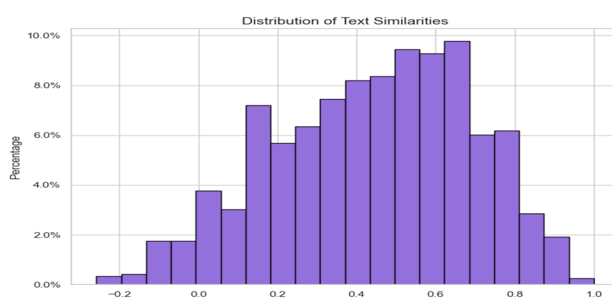

Figure 2: Cosine similarities between annotations corresponding to the same regions.

We definitely see a similarity between the texts, as the bulk of the similarities lies above 0. We were a bit surprised that the average similarity is only 0.44, as we expected it to be way higher. We even found some similarities < 0, which might represent annotations containing some misinterpretations.

c.) How many annotations did we collect per file? How many distinct sound events per file?

The number of annotations per file was easily calculated by grouping the corresponding dataframe by filenames.

Estimating the number of distinct sound events per file was a bit more complicated. For each set of annotations of one annotator for one file, we fetched the corresponding annotation embeddings and computed their pairwise cosine similarities. Say there N annotations, they were then put into a $N \times N$ similarity matrix. This matrix was then filtered such that all values below a similarity threshold of 0.8 were set to 0 and all others to 1. This should give connected components within the graph corresponding to that matrix where the annotations are extremely similar, most likely because they describe the same sound event. Therefore the number of connected components is our estimate for the number of distinct sound events. If there were multiple annotators per file, the number was averaged and rounded. The results are shown in 3.

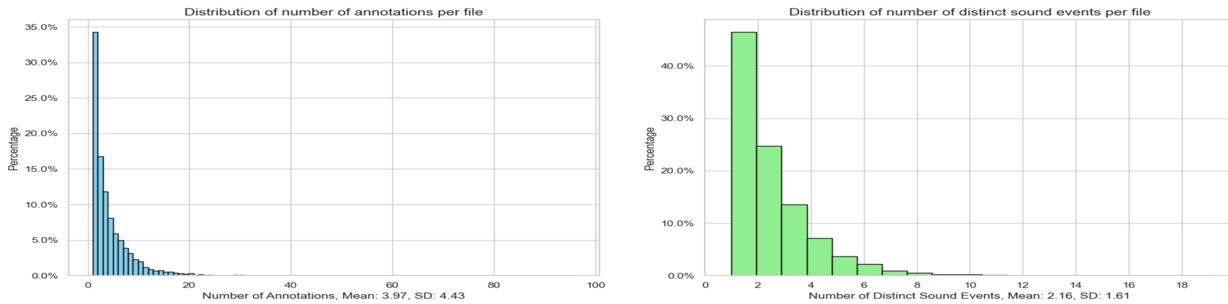

Figure 3: Distributions of annotations and distinct sound events per file.

The results seem reasonable. We have an average of 3.97 annotations and 2.16 distinct sound events per file. Interestingly, the maximum number of annotations for a single file.

d.) How detailed are the text annotations? How much does the quality of annotations vary between different annotators?

As quality metrics for a single annotation,, we decided to compare the Text-Token-Ratio (TTR, number of unique word divided by the number of words), the number of spelling errors and the number of words. Analyzing the textual annotations for different annotators led us to the following 4.

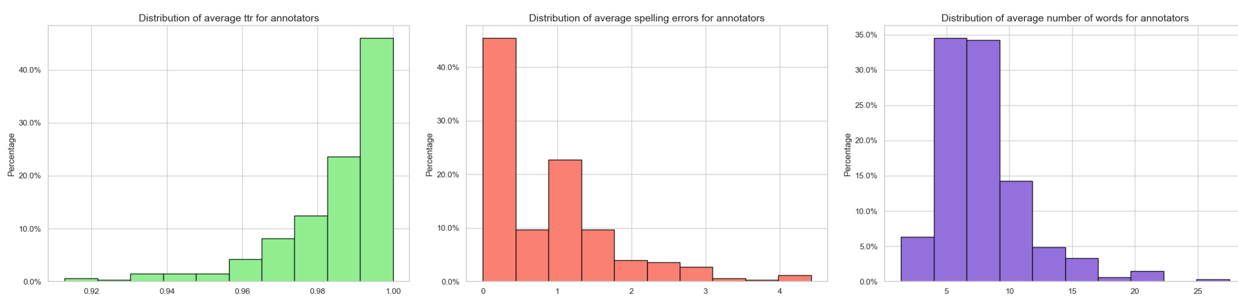

Figure 4: Distributions of quality metrics over all annotators.

The best metric out of these for checking how detailed the annotations are would most likely be the number of words per annotation. As the average annotator has written an average of 7.85 words per annotation, most of them seem to be reasonably detailed. As for the variation in annotation quality: we were surprised to see that almost half of the annotators had on average more than 1 spelling error in their annotation. There also seem to be some with extremely low word counts. The standard deviation 3.53 for the average number of words seems reasonable as well. Most annotations should be of sufficient quality. TTR turned out to be useless, as most annotations are so short that it would be very hard to find duplicate words.

e.) 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).

As our previous analysis has shown: yes, there are some poor quality annotations, i.e. outliers. This may be some consisting only of a single word or not relating to the audio because of some misconception. These could simply be removed by checking the word count of the annotations and removing the ones for which the text embedding is really dissimilar from metadata embeddings. When it comes to fixing typos and spelling errors, one could simply use an off-the-shelf library (like in the above section) to go over the texts and correct them.

3.) Audio Features

a.) Which audio features appear useful? Select only the most relevant ones or perform a down projection for the next steps.

As a method to extract the most important audio features, PCA was chosen.

The first step was to flatten them into a single array to make them usable for PCA. Using flattened data, the goal was to reach a cumulative explained variance of 95%.

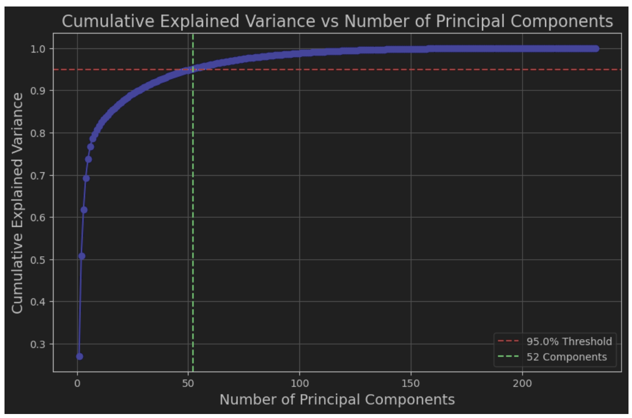

Figure 5: Cumulative Explained Variance using PCA on the audio features

The threshold is reached with 52 Principle Components. Now the 10 most important feature groups (embeddings, ZCR, mel-spectrogram, mfcc, etc.) of each component are counted.

Top contributing feature groups across the first 52 components:

Feature Group: mfcc, Count: 338; Feature Group: melspectrogram, Count: 82; Feature Group: embeddings, Count: 76; Feature Group: contrast, Count: 14; Feature Group: centroid, Count: 3; Feature Group: energy, Count: 2; Feature Group: power, Count: 1; Feature Group: flatness, Count: 1; Feature Group: bandwidth, Count: 1; Feature Group: zerocrossingrate, Count: 1; Feature Group: flux, Count: 1

As a result, Δ-mfcc and $\Delta\Delta$-mfcc can be ignored.

b.) Extract a fixed-length feature vector for each annotated region as well as for all the silent parts in between.

.

First the unimportant features are excluded and the annotated, as well as unannotated parts get extracted. The snippets of the audio features are concatenated into a single array, where the first [:len(X)] elements are annotated and the rest unannotated features. T-SNE is then used to create a 2-dimensional vector for each section.

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?

K-means was used to cluster the audio features. The clustering was applied on the raw data instead of the t-SNE down projection since it yielded a better differentiation of annotated and unannotated features.

While cluster 9 consists almost only of unannotated features, it is not so clear for all the other clusters. This is probably due to the fact, that most unannotated parts are either not completely silent or are small gaps between annotations that should still be part of the annotation.

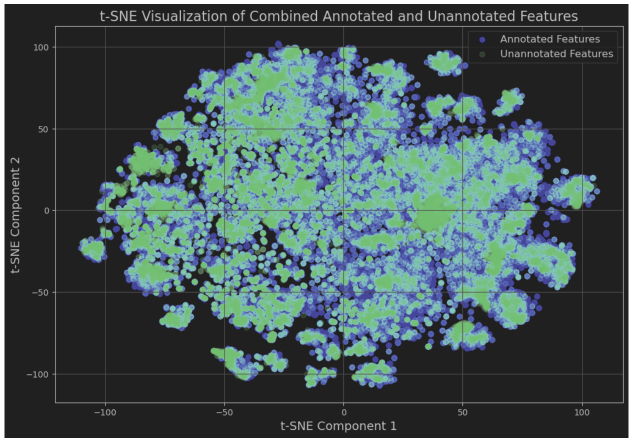

Figure 6: Audio Features T-SNE

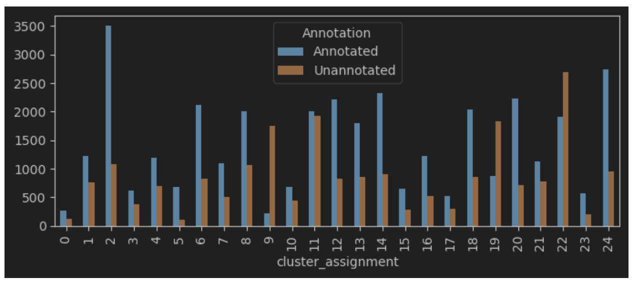

Figure 7: Clustered Audio Features

4.) Text Features

a.) Cluster the text features. Can you find meaningful clusters?

To identify meaningful clusters in the annotation text feature space, several dimensionality reduction and clustering strategies were tested. The best results were achieved by first applying t-SNE with a perplexity of 100 to the dataset. This produced visually distinct and well-separated groups.
Following dimensionality reduction, K-Means clustering was applied to the 2D t-SNE output, yielding 24 clusters. To evaluate the quality of these clusters, the most common words in each cluster were analyzed. Clusters dominated by a few semantically consistent keywords (e.g., "dog", "bark", "puppy") were considered well defined., while clusters with a wide range of unrelated terms were considered less consistent.
Overall, most clusters represented one or two distinct topics, indicating that the annotation embeddings were highly clusterable. For example, insect sounds, bad weather and cat noises were each grouped into their own unique clusters, showing semantic similarity through spatial proximity in the feature space.

b.) Design a labeling function1 for classes dog and cat. Do the annotations labeled as dog or cat sounds form tight clusters in the text and audio feature space?

To evaluate whether semantically similar annotations form tight clusters, labeling functions were created using keyword matching, similar to the ones introduced in the lecture. Simple rule based filters were used to identify dog and cat sounds. These functions relied on a small set of keywords:
Dog: "dog", "bark", "puppy", "growling"
Cat: "cat", "kitten", "meow", "purr"
These functions proved very accurate, identifying large numbers of relevant samples with minimal false positives across the board. The cat related samples clustered almost entirely within a single cluster (cluster 4), showing high cluster purity. Dog related samples appeared primarily in two clusters (clusters 18 and cluster 21), which were spatially adjacent in the 2D projection, indicating that the clustering identified a connection between those clusters even if they were not clustered together. These results demonstrate that the text embedding and clustering approach captured meaningful semantic structure in the data, clustering annotations labeled cat or dog sounds closely.

c.) How well do the audio feature clusters align with text clusters?

To explore the relationship between text based and audio based clusters, the cluster assignments from the text features were overlaid with those from the audio features. The goal was to determine whether semantically grouped annotations also clustered similarly based on their audio characteristics.
The results varied depending on the specific. For example, cat related samples appeared almost entirely within a single audio cluster. In contrast, other text based clusters, such as musical instruments, were distributed across multiple audio clusters.
This discrepancy can be explained by the fundamental difference between the two feature spaces.The text embedding reflects semantic meaning, while the audio embeddings reflect sound similarity. For instance, the sound of a drone and a violin are acoustically similar (sometimes even mistaken for one another by humans as seen in point one) but they are semantically distinct. As a result, they may cluster together in the audio space but remain separated in the text space.
Overall there is still a high degree of alignment between many clusters across both clustered embeddings, showing that text and audio aligns to a substantial degree.

5.) Conclusions

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?

6.) Submission of MLPC reports

Please read the instructions below carefully and follow them faithfully. Note that this template is based on the official Neurips 2023 template. In your report, you may use three levels of headings, as described in what follows.

7.) Headings: first level

This is a first level heading.

a.) Headings: second level

This is a second level heading.

a.).1 Headings: third level

And this is a third level heading. Make sure to structure your report s.t. no deeper levels are necessary.

8.) Footnotes, Figures and Tables

a.) Footnotes

Footnotes should be used sparingly. Note that footnotes are properly typeset *after* punctuation marks.[1]

b.) Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. You may use color figures. Please refer to all your figures in text, by using e.g., Figure 8.

c.) Tables

All tables must be centered, neat, clean and legible. Please refer to all your tables in text, by using e.g., Table 2.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package.[2]

[1] As in this example.
[2] `https://www.ctan.org/pkg/booktabs`

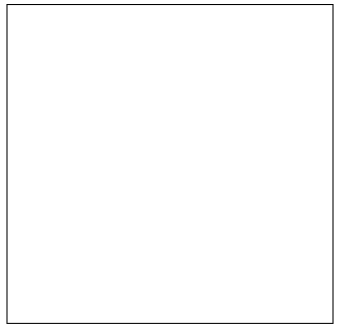

Figure 8: Sample figure caption.

Table 2: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

9.) Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (this will result in a deduction of points). Please note that pages should be numbered, and adhere to the given *page limit* to avoid further point deductions. Your final submission should be a `pdf` file.