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. Everyone communicated via a dedicated Discord server and the editing was done over a Github repository.

1.) Case Study

To find 2 interesting records that were edited by multiple annotators was the challenge here. We used the file "annotations.csv" and the corresponding "annotations_text_embeddings.npz". All files with more than one annotator were searched for. We compared all annotations by averaging the embeddings per annotator and calculated the corresponding cosine distances. The file with the largest difference (largest distance) is '568273.mp3' and the file with the highest similarity (smallest distance) is '203149.mp3'.

a.) Identify similarities or differences between temporal and textual annotations from different annotators.

The temporal windows fit together very well. The keywords in the different annotations describe two different things (violin, drone). This shows that our assumption of the correctness and accuracy of the gathered data cannot be assumed. The word drone appears in the title and the keywords (Table 1).

Table 1: File with the largest difference

index	annotator	filename	onset	offset	text
7323	1	568273.mp3	0.0	20.073243	A sharp, loud violin plays rapidly at a concert.
27703	2	568273.mp3	0.0	20.028386	A sustained ambient drone with granular and spectral textures.

The collected temporal windows match very well. The annotations in the similar files describe two identical things (cows, birds). Annotator 1 has only worked in much more detail (Table 2).

b.) To what extent do the annotations rely on or deviate from keywords and textual descriptions in the audio's metadata?

In both cases, the metadata match the text field very well (Table 1-'text', 2-'text' and 3-'title'+'keywords').

c.) Was the temporal and text annotations done according to the task description?

Yes, the annotations were made according to the task description. However, in the dissimilar case the quality is better (more detailed). The similar case was greatly simplified.

Table 2: File with the highest similarity

index	annotator	filename	onset	offset	text
4228	1	203149.mp3	0.764318	...	Cows and bulls calling and mooing
30194	1	203149.mp3	...	23.045349	cows and bulls calling and mooing
21454	2	203149.mp3	0.064283	24.191995	cows and bulls mooing
13107	1	203149.mp3	0.046322	...	birds singing in the country side
20461	1	203149.mp3	...	24.133923	Birds singing
29140	2	203149.mp3	0.064283	24.191995	birds singing

Table 3: Metadata of the files

index	filename	title	keywords
5182	568273.mp3	spectral violin drone processed through granulation and reverb	spectral, tonal, granulation, horror, drone, dark, avant-garde, ambient, violin, soundscape, experimental
6104	203149.mp3	End of the afternoon in a field, in Nebraska	CD130519T018, felix, cows, singing, birds, end, usa, field, call, bird, evening, countryside, bulls, moo, bull, sing, mooing, blume, calls, cow, fields, calling, nebraska, afternoon

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 (Figure 1).
We can see that most annotations fall below a difference in onset-/offset times of 0.1 seconds. Depending on the temporal resolution of the audio features, this might actually lead to some wrongly predicted frames in the future. Overall, most annotations seem to be pretty accurate.

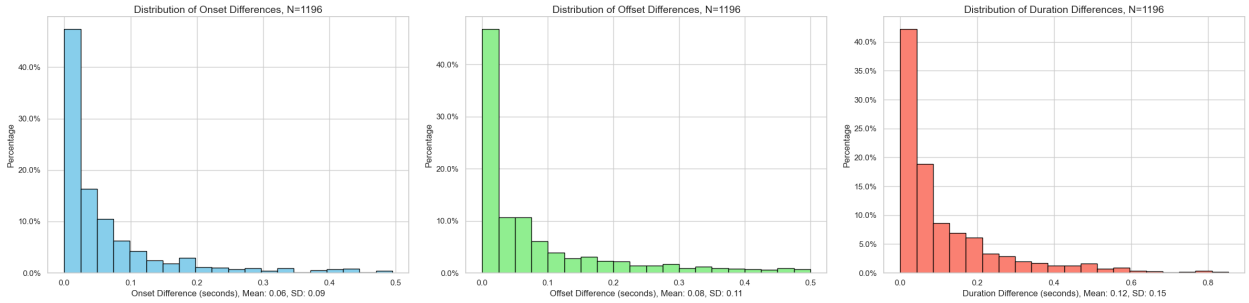

Figure 1: Absolute differences in onset-/offset times and durations for annotations corresponding to the same regions.

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 (Figure 2).
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 be the results of annotations containing false information.

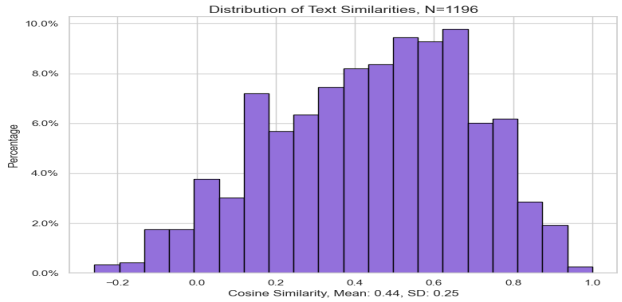

Figure 2: Cosine similarities between annotations corresponding to the same regions.

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 of N annotations from one annotator for one file, we fetched the corresponding annotation embeddings and computed their pairwise cosine similarities. 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 to turn the matrix into. This then gives 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 (Figure 3).
The results seem reasonable. We have an average of 3.97 annotations and 2.23 distinct sound events per file.

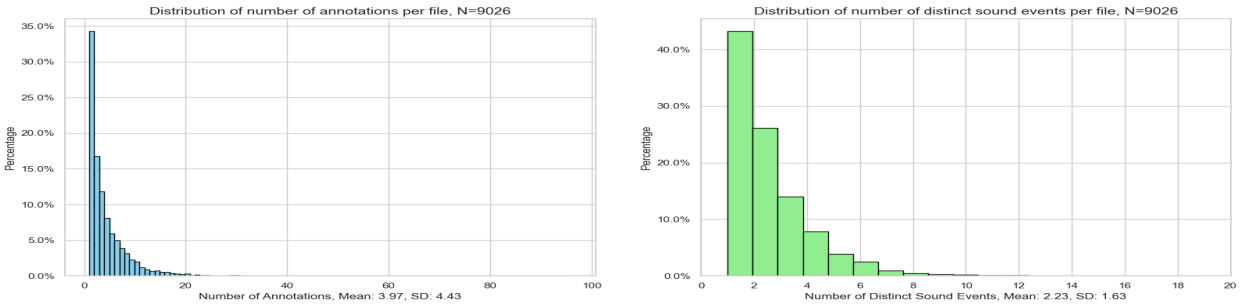

Figure 3: Distributions of annotations and distinct sound events per file.

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 (using a simple off-the-shelf spell checker) and the number of words. Analyzing the textual annotations for different annotators led us to the following graphs (Figure 4). 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 mostly 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

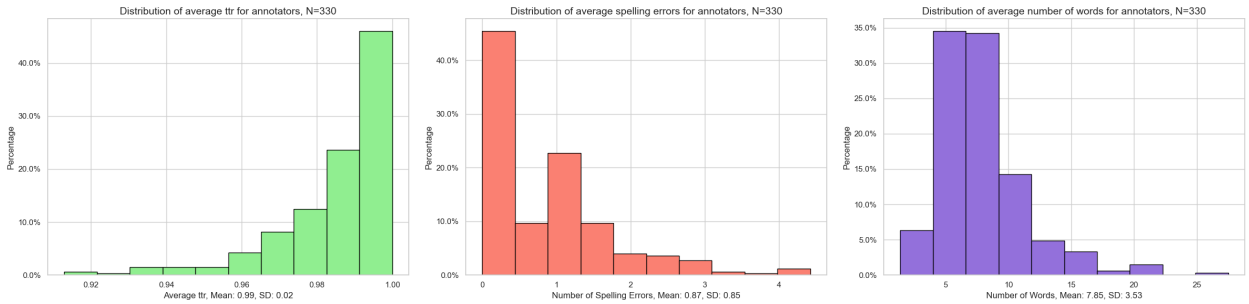

Figure 4: Distributions of quality metrics over all annotators.

dissimilar from metadata embeddings. When it comes to fixing typos and spelling errors, one could simply use an off-the-shelf library 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%.
The threshold is reached with 52 Principle Components (Figure 5). The 10 most important feature groups (embeddings, ZCR, mel-spectrogram, mfcc, etc.) of each component are then counted.

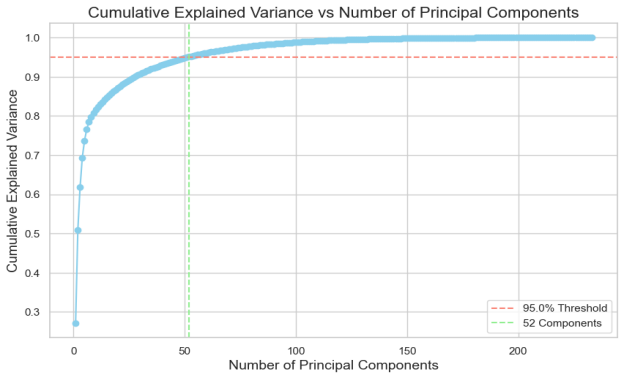

Figure 5: Cumulative Explained Variance using PCA on the audio features

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

This means that 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 from task a) are excluded. From the important features, the annotated and unannotated parts are extracted. The snippets of audio features are then concatenated into a single array where the first [:len(X)] elements are annotated and the rest are unannotated features. T-SNE is then used to create a 2-dimensional vector for each section. The vectors are displayed on figure 6

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. Clustering was applied to the raw data instead of the t-SNE down projection, as it provided a better distinction between annotated and unannotated features.
While cluster 9 consists almost entirely of unannotated features, this is not so clear for all the other clusters (Figure 7). This is probably due to the annotation errors, as most of the unannotated parts are either not completely silent or are small gaps between annotations that should still be part of the annotation.

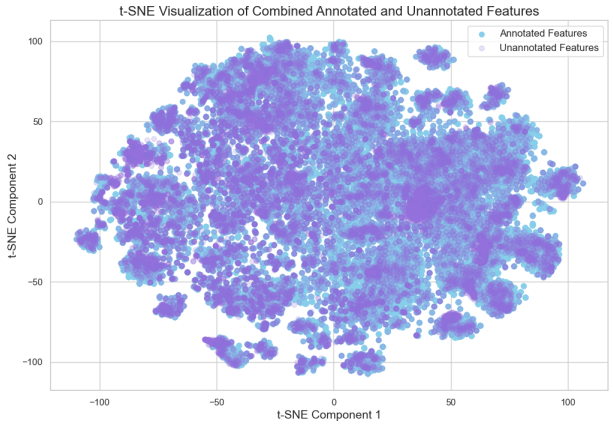

Figure 6: Feature Vectors using t-SNE

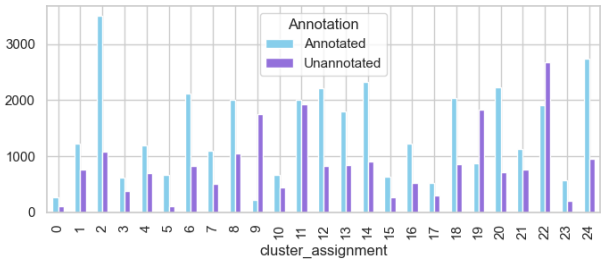

Figure 7: Clusters of Audio Features

4.) Text Features

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

To find 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 and then clustering it into 24 clusters using K-Means (Figure 8).
To evaluate the quality of these clusters, the most common words in each cluster were analyzed (Figure 9).[1] 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 high cluster quality (Figure 9).

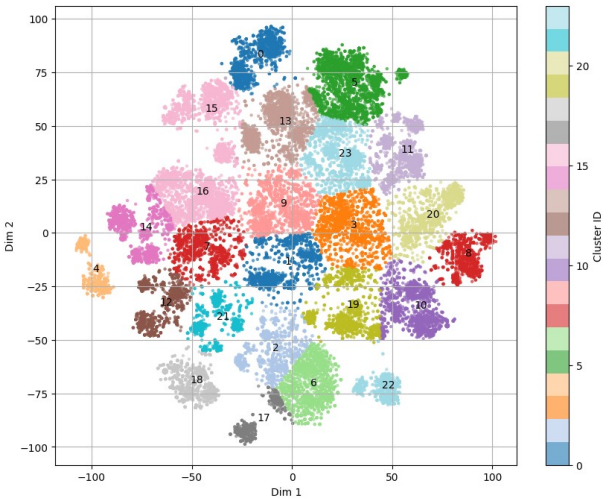

Figure 8: t-SNE of text feature space, clustered by K-Means

Cluster	Topic	Most common Words
0	Flowing Water	water, waves, splashing
1	Applause and Crowds	clapping, people, door
2	Distant Bird Calls	bird, chirping, distance
3	Metal Hammering	metallic, hammer, metal
4	Cat Sounds	cat, meowing, purring
5	Vehicles Passing By	engine, car, passing
6	Singing Birds	bird, chirping, birds
7	Humans Shouting and Singing	man, singing, loudly
8	Church Bells and Chimes	bell, ringing, metallic
9	Footsteps and Muffled Sounds	person, footsteps, walking
10	Instrumental Music	guitar, playing, piano
11	Power Tools and Construction	drill, chainsaw, loud
12	Snoring and Breathing	snoring, person, man
13	Storms: Wind and Thunder	wind, noise, thunder
14	Babies and Child Voices	baby, laughing, crying
15	Rain and Fire Crackling	water, rain, falling
16	People Talking	talking, people, man
17	Farm Animals: Sheep and Goats	sheep, bleating, birds
18	Dog Barking	dog, barking, barks
19	Horns and Drums	horn, drum, honking
20	Alarms and Sirens	alarm, siren, beeping
21	Mixed Animal Sounds	dog, cow, mooing
22	Insect Buzzing and Flying	buzzing, insect, flying
23	Train and Machine Humming	train, machine, noise

Figure 9: Assigned Topics and Top Keywords per Cluster

[1] "a", "the", "and", "of", "is", "in", "with", "an", "on", "from", "by" where excluded from the count as they are non-descriptive

b.) Design a labeling function 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, relying on a small set of keywords ("dog", "bark", "puppy", "growling" for dogs and "cat", "citten", "meow", "purr" for cats)
These functions proved very accurate, identifying large numbers of relevant samples with minimal false positives. The cat related samples clustered almost entirely within a single cluster – cluster 4 (Figure 10), showing high cluster purity. Dog related samples appeared primarily in two clusters – clusters 18 and 21 (Figure 11), which were spatially adjacent in the 2D t-SNE projection. This indicates tight clusters for semantically similar topics like cats and dogs.

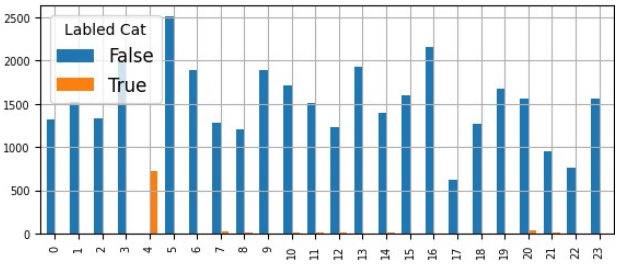

Figure 10: Samples labeled "Dog" in orange across text clusters

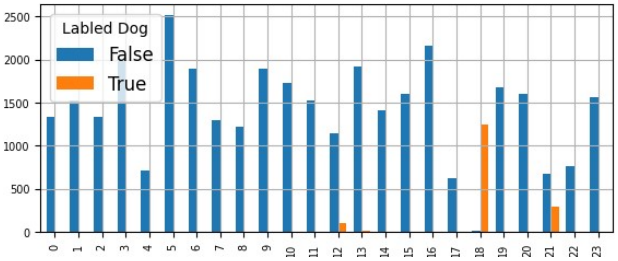

Figure 11: Samples labeled "Cat" in orange across text clusters

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

To find the relation between audio feature clusters and text clusters, the text clusters were laid over the audio clusters to see if they would fall into a specific cluster. The result varied depending on the cluster, cat and dog sounds fell almost entirely into a single audio cluster (Figure 12). Other text clusters, like musical instruments, were spread over multiple clusters (Figure 13).
This is likely because the text space is clustered based on the semantic relation of words, while the audio feature clusters are based on the similarity of sound. For instance, a drone and a violin might end up in the same audio cluster even though their semantic meanings are very different. Still, there is a clear connection between the audio feature clusters and text clusters, and many of the clusters align at least in part.

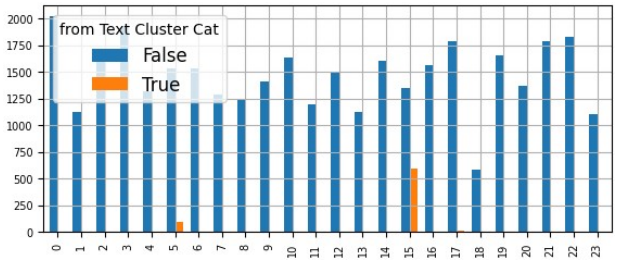

Figure 12: Spread of text cluster "Cats" across audio clusters

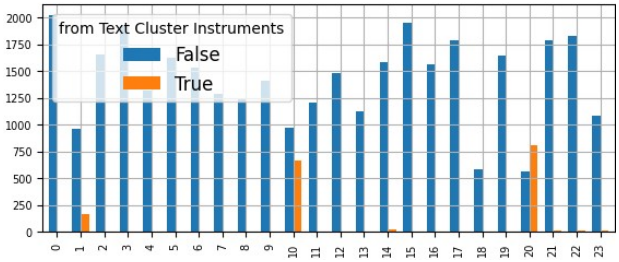

Figure 13: Spread of t. cluster "Instruments" across audio clusters

5.) Conclusions

a.) Is the dataset useful to train general-purpose sound event detectors?

While the dataset may be useful for detecting somewhat specific sounds which fall into discovered clusters, it most likely is neither large or diverse enough nor of high enough quality for truly general use.

b.) Which biases did we introduce in the data collection and annotation phase?

The data collection seems to have happened by searching for recordings from specific fields (like concerts, pets, rain & wind, ...) which probably introduced some bias. Additionally it was said that unpleasant / inappropriate recordings were specifically filtered out. As for the annotations, there definitely are some language / demographic biases which take effect in the data.