# MLPC Report

**Team SPARKLING**

**Giovanni Filomeno** **Katharina Hehenwarter** **Elīna Emīlija Ungure** **Kathrin Hofmarcher**

## Contributions

Katharina Hehenwarter did the Case Study and wrote the conclusion (Task 1 and 5). Elīna Emīlija Ungure worked on Annotation Quality (Task 2). Kathrin Hofmarcher investigated Audio Features (Task 3). Giovanni Filomeno looked into Text Features (Task 4) and made the slides. Everyone contributed to the report by writing about their tasks.

## 1 Case Study

→438153.mp3

The audio with the filename "438153.mp3" and the title "G17-07-Cattle Roundup.wav" (retrieved from the metadata.csv) was annotated by the annotators "4990913873726557849864133405302676392750611324650759682869018099518271234941" (for shortness named here 'Annotator 1') and "6899111265188446986752251952205782765329438939709985607240156637228874596385 2" (for shortness named here "Annotator 2"). The annotators were also given the following keywords: "Cows, Vintage, Optical, Animals" and the description: "Cattle Roundup. Cowboys call. Horses gallop. Some mooing.". Annotator 2 annotated in general 12 events, while Annotator 1 annotated 7 events. Both annotators identified that in this audio are sounds made by cows and a man, as it was described in the metadata. Furthermore, Annotator 2 heard a sheep bleating in the last 2 seconds of the audio, while Annotator 1 identified this sound as a mooing cow. This could mean that Annotator 1 followed the given keywords during the annotation task more than Annotator 2. Both annotated each sound as a new region and annotated all animal sounds, while Annotator 2 annotated the sounds made by people in more detail with "shouting, talking, laughing, whistling". Annotator 1 also missed the continuous sound which can be heard throughout the whole audio which Annotator 2 described as a "steady rumbling sound produced by a vehicle". For the textual representation, Annotator 2 tends to use more detailed phrases, while Annotator 1 uses more brief and general phrases. Both of them explained the sound source very well and stick to the requirement "One Description → One Sound". But both of them lack mostly the contextual description on what environment the sound suggests and the descriptive information ("a person shouting", "a sheep bleating"). Also two of the annotations made by Annotator 1 were not independent ("more cows mooing").

→441791.mp3

The audio with the filename "441791.mp3" and the title "180081 Bee Against Window.WAV"(retrieved from the metadata.csv) was annotated by the annotators "7526205831694568980367141074683662123375335303003550042342837024713291313900 3" (for shortness named here "Annotator 1") and "8018067142874428479292091530341476192021309155303501668119887028799200001861 4" (for shortness named here "Annotator 2"). The annotators were furthermore given the following keywords: "Buzzing, Annoying, OWI, Window, Bee" and the description: "An African Bumble Bee buzzing against a window". Annotator 1 annotated 15 events in general, while Annotator 2 annotated 8 events, which suggests that Annotator 2 merged the buzzing of the bee into larger regions while Annotator 1 selected more and smaller regions. Annotator 1 is in all his annotation descriptions talking about a fly making noises while Annotator 2 talks about a bug. This suggests that both of them were not influenced by the given keywords and the audio description which are both talking about a bee. Also Annotator 2 is referring in one of the annotations to an "object" which 'disturbs it in the flight' which could be the mentioned window (see description). Both of the annotators missed the quiet continuous bird chirping in the background due to the low volume of the audio. Also both annotators use the same text for almost all of their

annotations (Annotator 1: "fly buzzing" (sometimes included: repeatedly/quietly); Annotator 2: "A bug is flying and buzzing around quietly in a steady pattern"). Comparing these two shows that the annotations of Annotator 1 mostly lack context (object/window not mentioned) and mostly have no temporal information even for differing duration times of the single annotation sequences. Annotator 2 on the other hand has two annotations which are not independent of each other. The first annotation is described as "A bug is flying and buzzing around close to the listener in a steady pattern", while the following has the description "A bug is flying and buzzing around closer to the listener". which suggest combining these 2 statements that the buzzing sound increased.

## 2 Annotation Quality

Different methods and statistics can be used to asses the overall quality of the audio annotations. Here, the similarity between different annotators on the same files is evaluated. Furthermore, the overall dataset is analyzed to get a better understanding of the precision, quantity, and quality of the dataset.

### 2.1 Temporal Annotation Precision

This subsection addresses how precisely the annotators marked the onsets and offsets. To measure this, files that are annotated by multiple people are filtered to create a separate dataset. The annotations are then grouped into clusters using the intersection over union, with the threshold set to 0.5. This ensures that the temporal annotations that describe the same sound are viewed as such in the calculations. For each cluster, the mean absolute deviation (for shortness named here 'MAD') of both the onsets and the offsets is calculated by $\frac{1}{n}\sum_{i=1}^{n} x_i - mean(x)$ to see how closely the annotators agree on the start and end time of the sound. Figure 1 summarizes the averaged MAD of all clusters in a file separately for onsets and offsets, in other words, each mark represents the MAD of an onset or an offset of a single file. From Figure 1, it can be concluded that the annotators mainly agreed on the temporal onsets and offsets with little deviations and a few exceptions. This conclusion is also supported by the weighted MAD, which for the onset is 0.202 seconds and for the offset is 0.226 seconds. The MAD is weighted to ensure that files with more clusters contribute more to the final result. This contributes to the calculation of the overall temporal precision score, which for onsets is 0.8 and for offsets is 0.77, again indicating good overall precision of onsets and offsets.

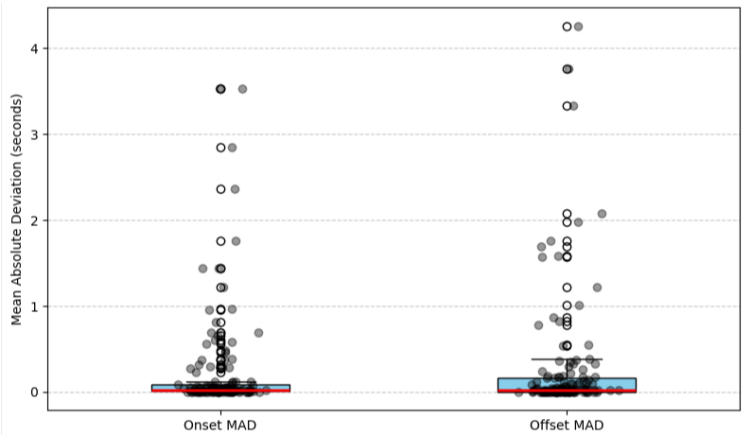


Figure 1: Distribution of Onset and Offset MADs

### 2.2 Similarity in Text Annotations

To evaluate how similar text annotations are between annotators, the same subset of files with multiple annotators is used as in subsection 3.1.

Firstly, the similarity between the clusters is measured. This is achieved by transforming the annotations into TF-IDF vectors, calculating a cosine similarity matrix, and extracting the mean similarity score. The result from this method is a similarity score of 0.137, which indicates that the annotators do not use the same words to describe the same event, and only very few words overlap.

Secondly, to confirm that the results yielded are trustworthy, a second method was employed to check the similarity between text annotations. The text annotations were turned into sets of unique words, and the Jaccard similarity is

used to derive the similarity between the annotations. For this method, the clusters were ignored, and all temporal annotations represented the file. In addition, if there were more than 2 annotators per file, the average of all the annotations represented the final similarity score of the file. The overall score is 0.221 and is derived by averaging all the file scores. This score is only slightly higher than the score from the first method, indicating that both scores represent the true similarity quite well.

Both methods indicate that the vocabulary used in annotations varies significantly. Since none of the methods can recognize synonyms, it could be that this is the cause of the low text similarity score. After investigating 1 to 2 examples, the same conclusion can be drawn. In the future, this could be solved by agreeing on a set of vocabulary to be used to annotate.

### 2.3 Summary of Overall Annotation Quality

The dataset consists of 2000 annotated files. The minimum number of annotations per file is 1, and the maximum is 73, with the average of 4 (rounded up from 3.9875) and the median of 2 annotations per file. To find statistics about distinct sound events, the keywords from the metadata are used. Unique keywords are counted per file in all temporal regions to ensure that if the same sound is repeated, it is counted as one sound source, and averaged, which comes to 10.43 distinct sound events per file. This number seems quite high, and after investigating the keywords manually, it seems that some keywords are adjectives or synonyms. To improve the accuracy of the estimation, a method that filters out adjectives and synonyms could be implemented.

The average word count for all annotators is 7.58. There are 28 annotators from 326 who annotated on average with 4 or fewer words, of which 6 used 3 or fewer words. These 28 annotators most likely lack details in annotations, making some of them less qualitative. Figure 2 displays and summarizes the variance between annotators. Each bar represents the average word count of all of the annotations produced by an annotator. It also shows that multiple people used more than 20 words per annotation on average, which might indicate that the text includes unimportant details and assumptions about the sound source.

Out of 2000, 59 annotations consist of 1 word and 509 consist of 2 or fewer words; these annotations are almost certainly of poor quality or outliers. Furthermore, there are also audio files for which the annotators cannot agree on the sound source, and annotations that have spelling mistakes (for instance, annotator "6187541200552586547127486465356195121808148462834891043910784263399802279564O" for file 574542.mp3 misspelled rhythm as "rythm"). A simple method to fix spelling mistakes would be to do an automated spell check using a spell-check tool like Grammarly. The few annotations consisting of less than 1 word could be annotated again, as there are not so many. Moreover, a test could be run to identify files on which 2 annotators completely disagree, and a third party could manually choose the more appropriate one.

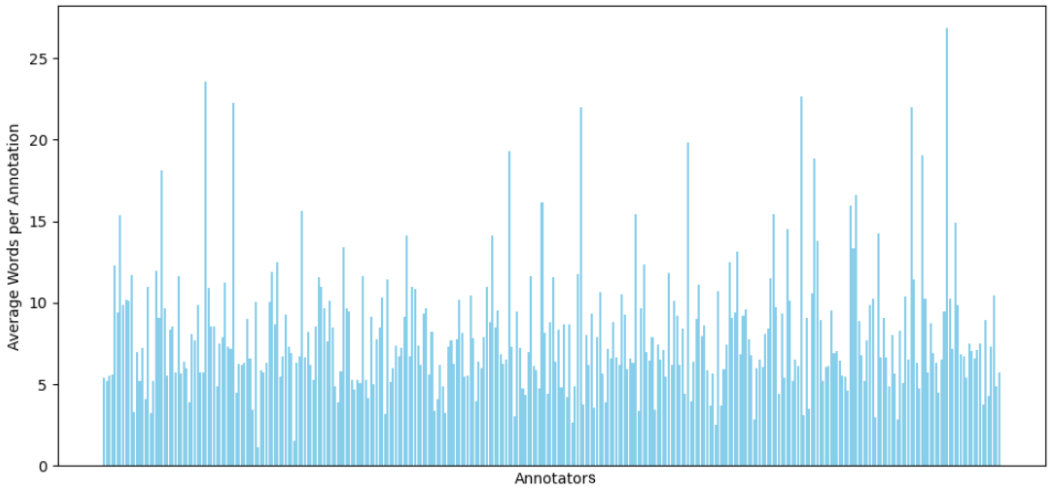


Figure 2: Average Word Count per Annotator

# 3 Audio Features

The audio features are: embeddings, melspectrogram, mfcc, mfcc delta, mfcc delta2, flatness, centroid, flux, energy, power, bandwidth, contrast and zerocrossingrate.

## 3.1 Most useful Audio Features

Each audio feature alone does not cluster the audios good since audios with keywords from the same group (e.g. insects) are not together but strayed over all datapoints. Also e.g. the audio features melspectrogram and mfcc together did not work well for clustering, which does not seem strange since since the single features do not show anything. The downprojection of all features with PCA led to a better result and was therefore mainly used in the next steps.

## 3.2 Meaningful clusters of audio features

To find clusters, Kmeans clustering to 30 clusters was done once only with the features melspectrogram and mfcc (Figure 3), which led to a poor result. The with PCA downprojected features (Figure 4) and not downprojected features, were better as with only the 2 features but also not too good when looking at the bad weather sounds. With 15 clusters the result for the PCA-downprojected data was much worse than with 30 clusters.

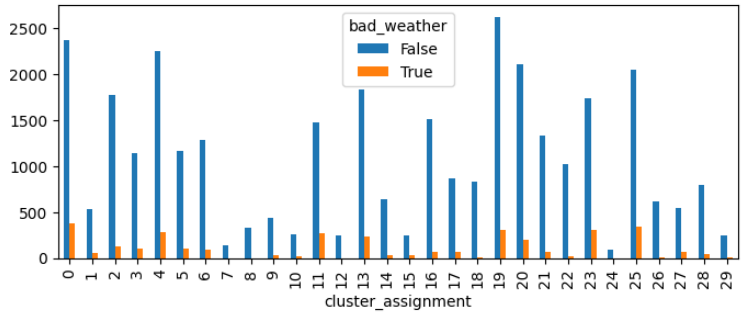


Figure 3: Clustering result for bad weather with features melspectrogram and mfcc

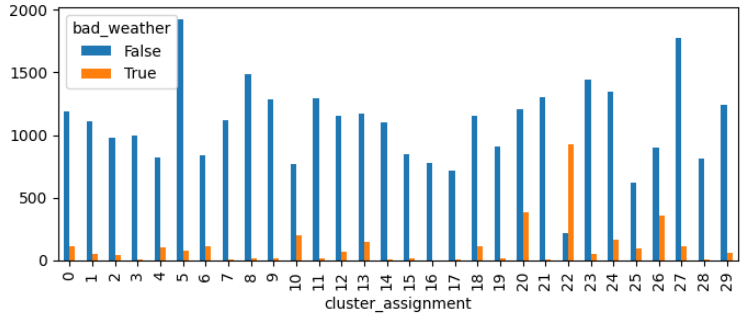


Figure 4: Clustering result for bad weather with PCA

# 4 Text–Feature Exploration (Task 4)

## 4.1 Choosing the number of clusters

To determine a sensible $k$ for $k$-means on the 35 826 annotation-embeddings (1024-D), we swept $k \in [2, 30]$ and computed Silhouette[1], Davies–Bouldin (DB; lower = better), and the within–cluster sum of squares (SSE). Figure 5 shows that while the *maximum* Silhouette occurs at $k \approx 29$, the DB minimum and the SSE *elbow* both lie around $k \approx 12$. We therefore adopt $k = 15$: interpretable size with only a marginal drop in Silhouette.

[1] Higher = better separation.

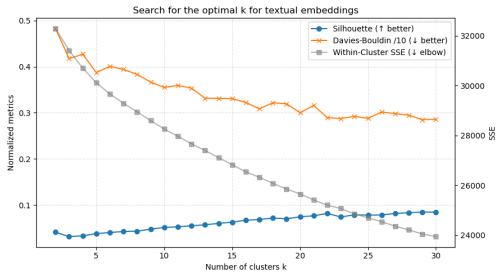


Figure 5: Internal-validity curves for $k$-means.

### 4.2 Dog & Cat labelling function

A simple regex `\b(?:dog|dogs)\b` / `\b(?:cat|cats)\b` serves as a *labelling function* for the two classes requested.[2] Projected into the audio-embedding space, Dog and Cat form visibly compact clouds with little overlap (Figs. 6 – 7). A quantitative check gives a global Silhouette of 0.163 (Dog 0.151, Cat 0.189), confirming that dog/cat frames are *tighter* than the background although not perfectly separated.

**PCA vs. t-SNE vs. UMAP.** Figure 6 (linear PCA) already reveals a loose "stripe" structure: both pets occupy the upper half of the ellipse-shaped embedding cloud, yet they overlap by $\approx 30\,\%$. The non-linear projections sharpen that picture: UMAP (Fig. 7 left) compresses the main dog cluster into a tight blob and exposes three small cat sub-clusters, while t-SNE (Fig. 7 right) goes one step further and almost tears the two species apart (global geometry is distorted, but local cohesion is maximal). The gradual Silhouette rise from PCA $\rightarrow$ UMAP $\rightarrow$ t-SNE (0.08 / 0.16 / 0.21 on the same 15 k-sample subset) quantifies this visual impression: non-linear methods capture the fine-grained acoustic cues (bark vs. meow) that PCA cannot separate in a strictly linear space.

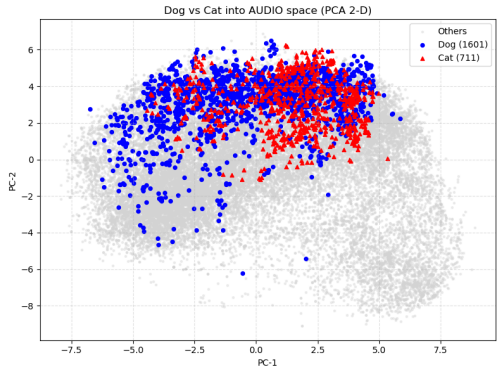


Figure 6: Dog vs Cat in AUDIO space – PCA 2-D.

### 4.3 Text–Audio cluster alignment

Using $k = 15$ for both modalities we re-cluster (average) audio features and compare them with the text clusters:

Table 1: Alignment metrics (valid annotations only).

| | Value | Interpretation |
|---|---|---|
| Adjusted Rand Index | 0.130 | weak agreement |
| Normalised MI | 0.258 | some shared structure |

A reordered contingency matrix (Fig. 8) shows that *some* clusters match well (e.g. Text 6 $\leftrightarrow$ Audio 7: 74 % of its samples) while others spread over several audio clusters, reflecting the fine-grained acoustic variability of the same textual concept.

[2]The stemmed variants ("dog", "dogs", ...) cover 2312 annotations ($\sim$6.4 % of the corpus).

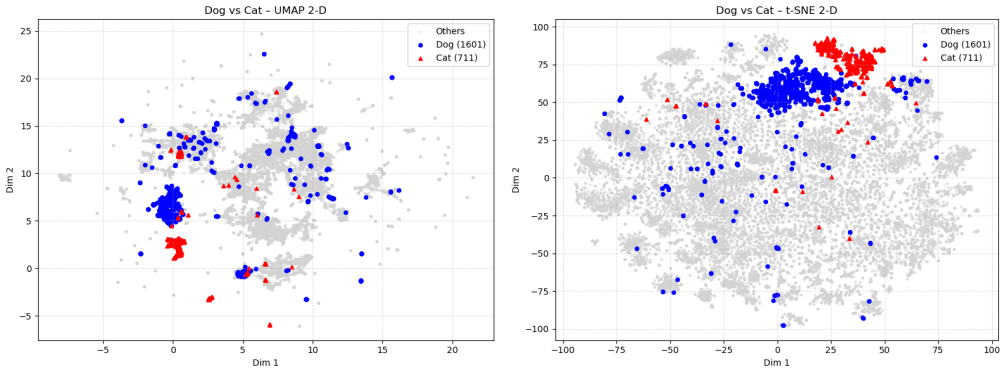


Figure 7: Same points with a non-linear projection: UMAP (left) and t-SNE (right).

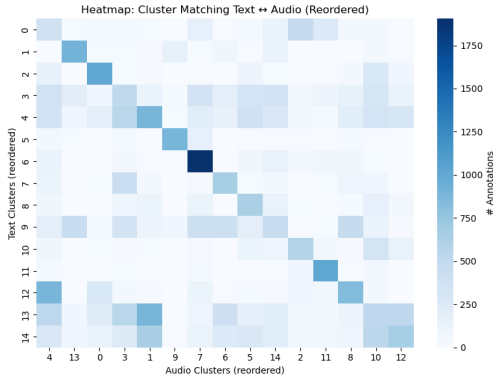


Figure 8: Text vs Audio cluster correspondence (re-ordered).

## 5 Conclusion

From the temporal analysis we found a high temporal annotation precision(MAD: 0,202s for onsets; 0,226s for offsets) which indicates that annotators in general agree on when sounds occur. Another advantage which is also good for training general-purpose sound event detectors is that we have a high diversity of sounds with an average of 10,43 distinct sound events per file. On the other hand we have an over-representation of sounds made by mammals, while rare sound categories like industrial noises are sparsely annotated and scattered across clusters. Also the variability in text annotations and also those sometimes missed sound events (mostly background sounds) may impact the model generalization. When only looking at the audio features it shows that those alone are insufficient for strong clustering but PCA-transformed features showed improved clustering outcomes. Furthermore the biases which were introduced in the data collection and annotation phase are that some annotators were strongly influenced by the provided keywords and description attached to every audio file. Also looking at the annotations individually and comparing them there exists a strong bias concerning the individual interpretation of occurring sounds which also can be found later in the varying levels of detail with which the sound events were described by the annotators(textual annotations show low similarity TF-IDF similarity: 0,137; Jaccard: 0,221). That means that some included minimal descriptions while others included speculative details depending also on the vocabulary the annotators used.