JOHANNES KEPLER
UNIVERSITY LINZ

MLPC – Team Toothpaste

Aral Cimcim, Sandro Müller, Linus Madlener, Kevin Eberl

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Feature Selection
Which audio features matter?

- **Raw input:** 768-dimensional frame wise embeddings
- **Preprocessing:**
  - Stack all frames across files StandardScaler
  - PCA → 10 principal components
- **Result:** first 10 PCs explain > 90 % of variance, so we reduce 768 → 10 without losing much.

Figure Cummulative Variance

Fixed Length Embeddings
From variable to fixed length vectors

Locate Frames:
- Onset/offset → frame indices (120ms resolution)

Projecting those frames to the X-D PC space

Average to get one vector per region (annotated and each silent gap)

$$v_r = \frac{1}{T_r} \sum_{t=1}^{T_r} f_t$$

where $f_t \in R^X$ is the PC vector at frame t

Clustering Audio Regions

Do regions group into meaningful clusters?

Input: X-D vectors for all **annotated** regions (no silences**)**
Algorithm: KMeans(n_clusteres = 8, random_state = 0)
Cluster Sizes: [X, X, X ,X , X]

Insight: Separates high-energy vs low energy events

Figure Cluster of Regions Bar chart two showcase the size of the clustered regions

Silent Regions: One Big Cluster?
Where do the silence land?

- **Build** fixed_vectors and segment_labels (“annotated” vs. “silent”)
- Embed everything via TSNE(n_components = 2, perplexity = 30)
- Colour points by label

Figure t-SNE annotated vs silent, what does silence forms a tight cluster or not?