

MLPC – Team Toothpaste



- Aral Cincim, Sandro Müller, Linus Madlener, Kevin Eberl

Feature Selection

Which audio features matter?

- **Raw input:** 768-dimensional frame wise embeddings
- **Preprocessing:**
 - Stack all frames across files
 - StandardScaler
 - PCA → 75 principal components
- **Result:** first 75 PCs explain > 80 % of variance, so we reduce 768 → 75

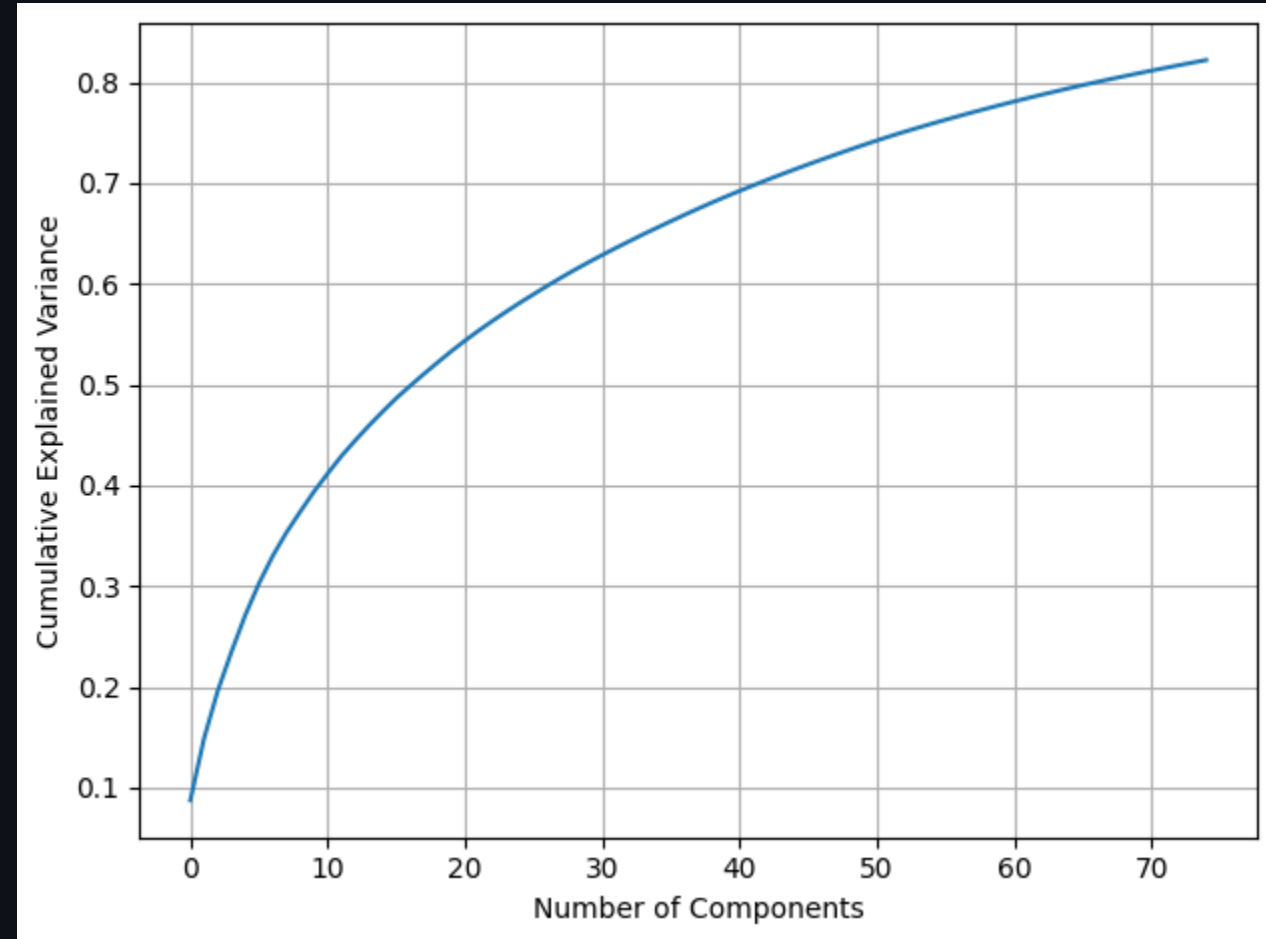


Figure 1: Cumulative Explained Variance as a Function of the Number of Principal Components

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

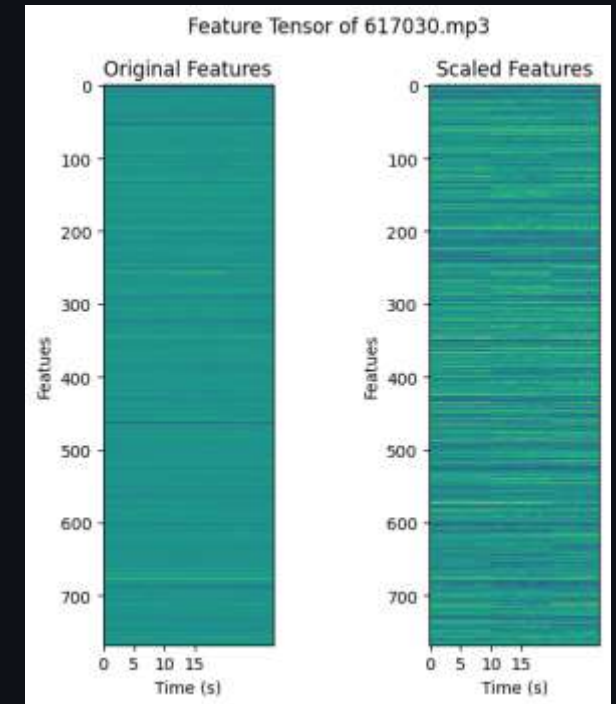


Figure 2: Scaled Features

Clustering Audio Regions

Do regions group into meaningful clusters?

Clustering

- **Algorithm:** K-Means
- **Number of Clusters (k):** 25
- **random_state = 42 , n_init = 10**

768 D → 75 D → KMeans(25)

Used less PCs to reduce noise and 25 clusters seemed to work the best by iterating over 10 possibilities

Table 1: Cluster Quality Metrics and Formulas

Metric	Score	Formula	Ideal
Silhouette Score	0.089	$s = \frac{1}{N} \sum_{i=1}^N \frac{b_i - a_i}{\max(a_i, b_i)}$	↑ higher
Calinski-Harabasz Index	936.9	$CH = \frac{\text{tr}(B_k)/(k-1)}{\text{tr}(W_k)/(N-k)}$	↑ higher
Davies Bouldin Index	2.697	$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)}$	↓ lower

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?

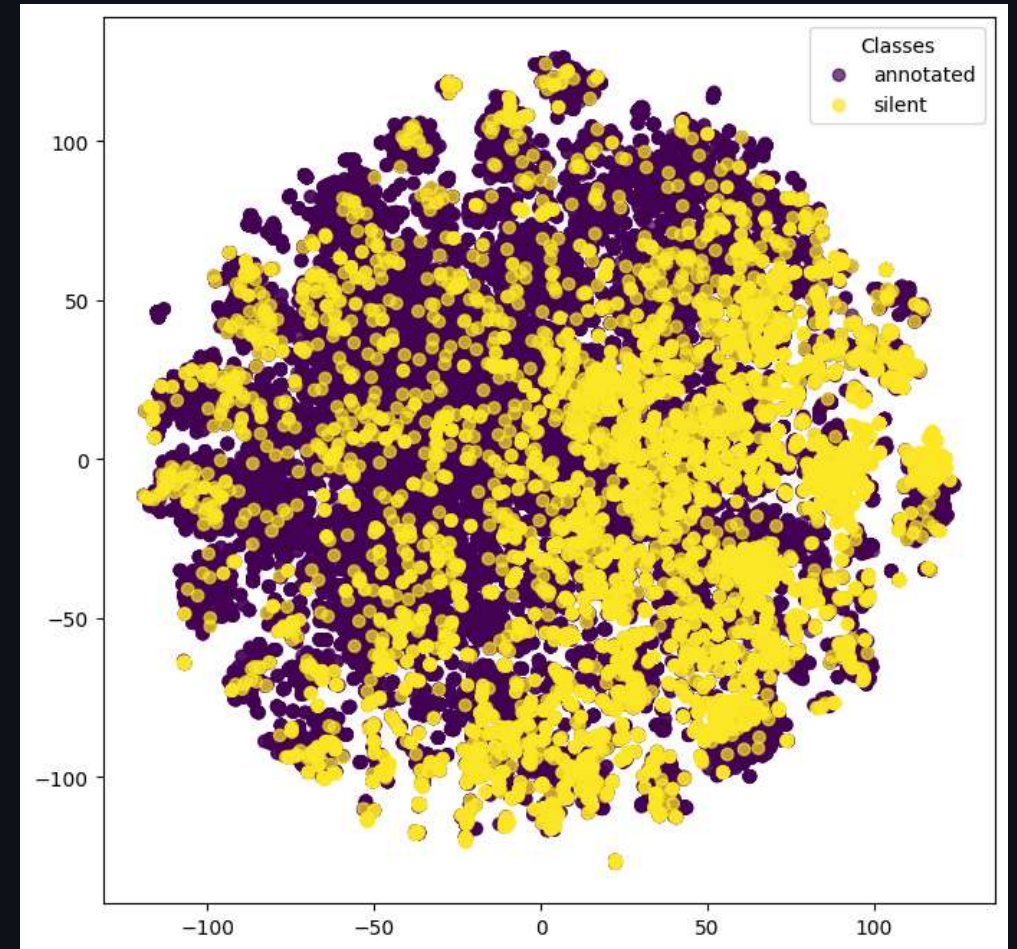


Figure 3: t-SNE Visualization of Annotated vs. Silent Samples