Audio Feature Analysis for Sound Event Detection

Mark Sere

April 2025

1 Introduction

This report presents an analysis of audio features extracted from the MLPC2025 dataset for sound event detection.

2 Dataset Details

2.1 Spectrograms and Mel Spectrograms

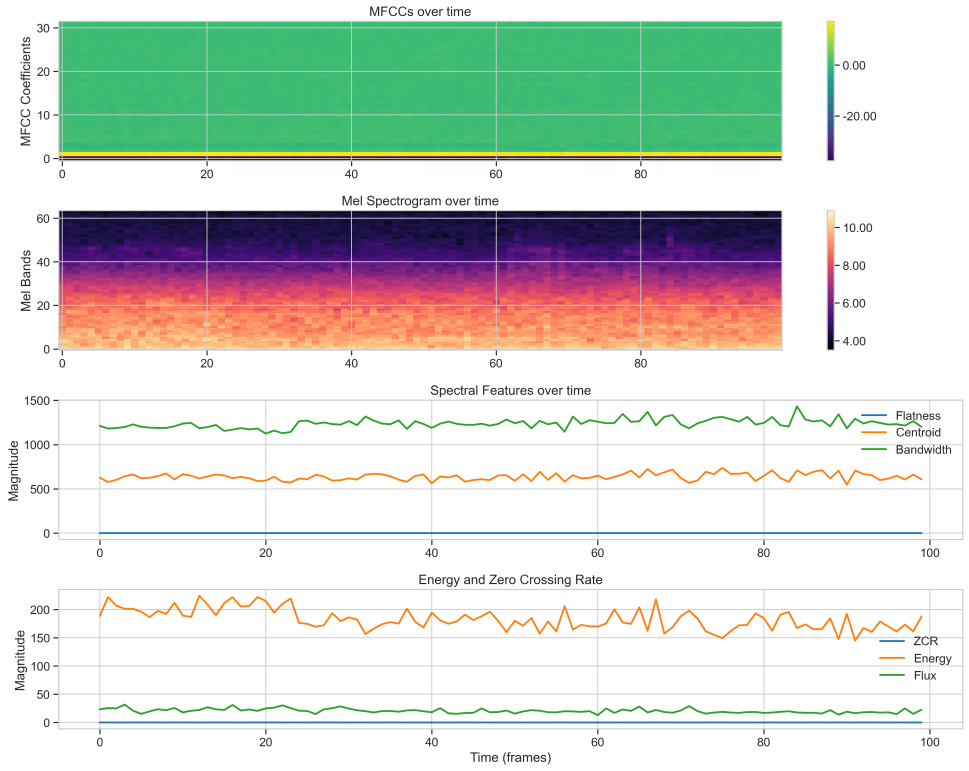

Figure 1: Visualization of different audio features over time for a sample file, including Mel spectrogram (second panel), which shows the distribution of energy across different frequency bands.

3 Feature Analysis

3.1 Dimensionality Reduction

Given the high dimensionality of the feature space (942 dimensions), Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving the most important

information.

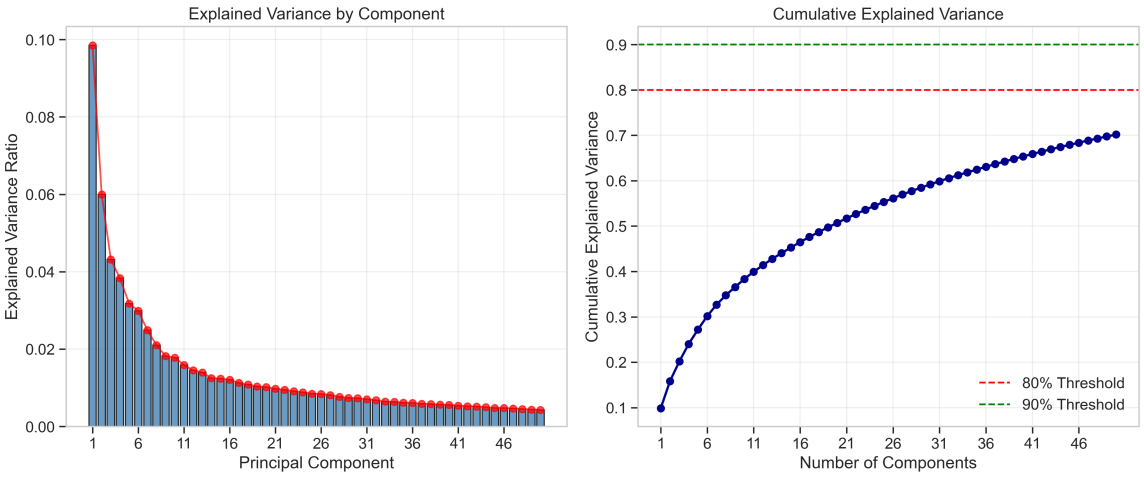

Figure 2: Explained variance by PCA components. Left: Individual variance contribution of each component. Right: Cumulative explained variance with 80% and 90% thresholds marked.

Analysis of the cumulative explained variance revealed that:

- 82 principal components are sufficient to explain 80% of the variance
- 146 components are needed to explain 90% of the variance

This represents a significant dimensionality reduction (from 942 to 82 dimensions, or 91.3% reduction) while maintaining most of the information content. The reduced feature set was used for subsequent clustering analysis.

3.2 Feature Importance

To understand which features contribute most to the principal components, the feature loadings for the top three components were analyzed.

Figure 3: Top 15 features contributing to the first three principal components, showing the relative importance of different feature types.

The analysis revealed:

- **First principal component**: Dominated by Mel spectrogram features with MFCC_0 (the DC component) having the highest loading. This suggests that energy distribution across frequency bands is the most significant factor for distinguishing audio signals.
- **Second and third principal components**: Heavily influenced by embedding features from the pre-trained neural network. These components likely capture higher-level semantic information about the audio content.

Across the top three components, the most important feature types were:

- Embeddings: 30 occurrences
- Mel spectrogram features: 14 occurrences
- MFCCs: 1 occurrence

This suggests that while traditional spectral features like MFCCs are valuable, the learned embeddings capture additional information that contributes significantly to the variance in the data.

3.3 Clustering Analysis

K-means clustering was applied to the dimensionally-reduced feature vectors to identify natural groupings in the audio data. To determine the optimal number of clusters, silhouette scores were calculated for different values of k.

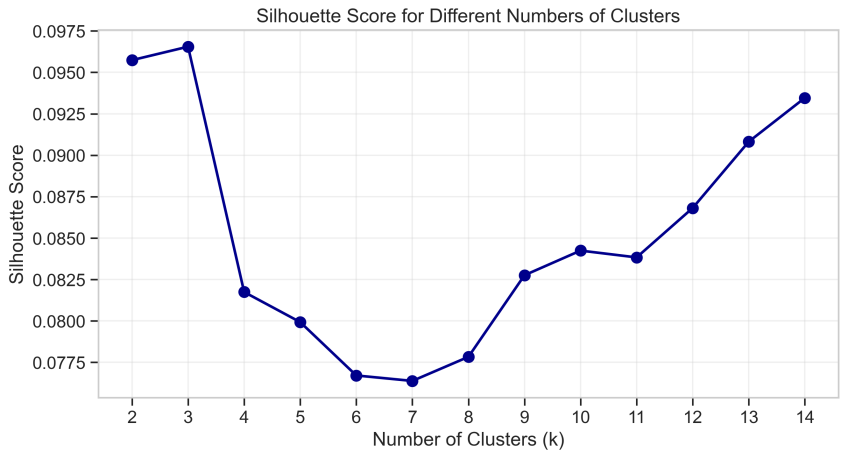

Figure 4: Silhouette scores for different numbers of clusters (k), showing that k=3 provides the best cluster separation.

The silhouette analysis indicated that 3 clusters provided the optimal balance between cluster separation and cohesion. The dataset was subsequently partitioned into these three clusters.

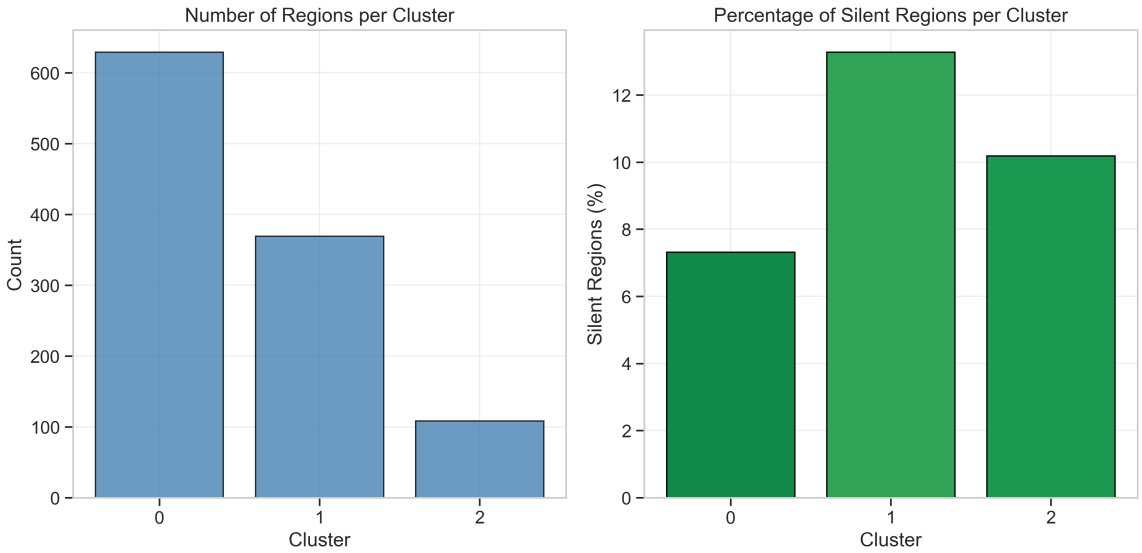

Figure 5: Left: Number of audio regions in each cluster. Right: Percentage of silent regions in each cluster.

3.4 Cluster Characteristics

Analysis of the three clusters revealed distinct audio characteristics:

Cluster	Size & Silent %	Common Annotations	Audio Characteristics
Cluster 0	629 regions (56.9%) 7.3% silent regions	Beeps, cat meows, pedestrian crossing sounds, metallic impacts	High embedding values, predominantly sharp, distinct sounds
Cluster 1	369 regions (33.4%) 13.3% silent regions	Insect buzzing, dog barking, bass drums, human vocalizations	Low embedding values, more sustained sounds with less tonal clarity
Cluster 2	108 regions (9.8%) 10.2% silent regions	Hihat sounds, guitar playing, rhythmic claps, alarm-like sounds	High embedding values, musical and rhythmic sounds

Table 1: Characteristics of the three audio feature clusters

Interestingly, silent regions were distributed across all three clusters rather than being concentrated in a single cluster. Cluster 1 had the highest percentage of silent regions (13.3%), suggesting that this cluster might represent lower-energy or background sounds.

3.5 Visualization with t-SNE

To visualize the high-dimensional feature space, t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied to project the data into two dimensions while preserving local relationships.

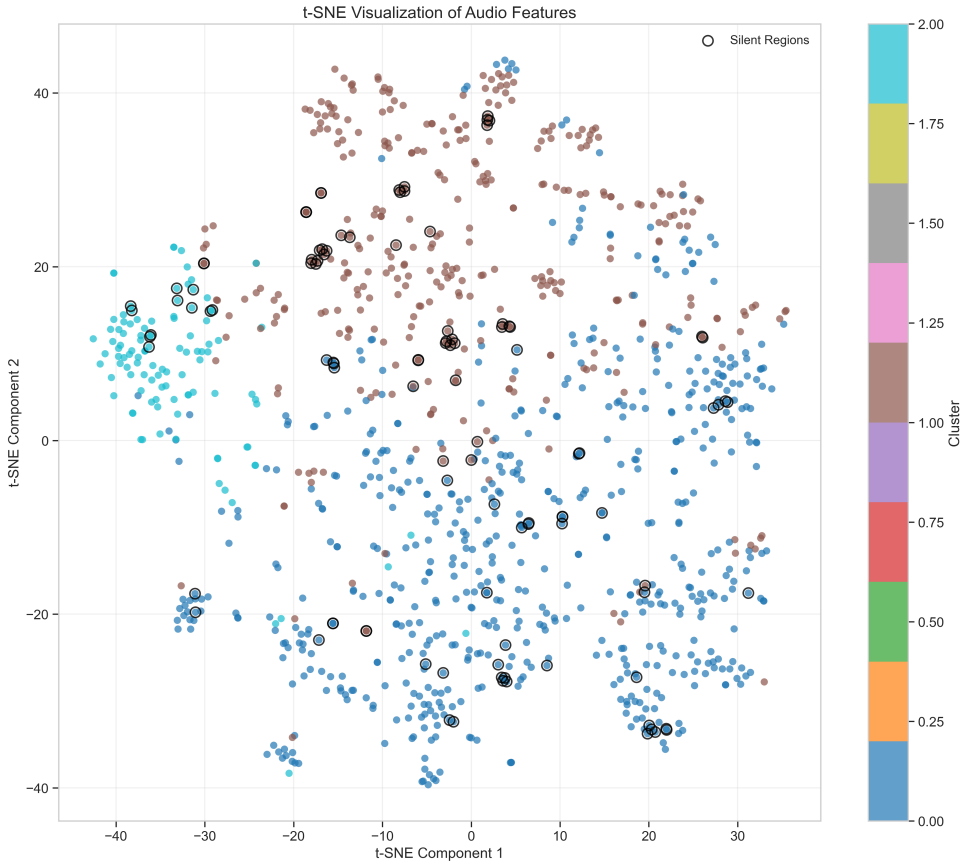

Figure 6: t-SNE visualization of audio features colored by cluster assignment. Silent regions are highlighted with black circles.

The t-SNE visualization confirms the existence of distinct clusters in the audio feature space. However, it also shows that the boundaries between clusters are not always clear-cut, with some overlap between different sound types. Silent regions (marked with black circles) appear scattered throughout the feature space rather than forming their own distinct cluster, suggesting that the absence of sound is characterized differently depending on the context.

4 Key Findings and Discussion

4.1 Dimensionality Reduction

The dimensionality of the audio feature space can be significantly reduced (by 91.3%) while preserving 80% of the information.

4.2 Important Features

The most significant features for distinguishing different audio events are:

- **Embeddings**: These learned representations capture good semantic content of the audio and account for the majority of the top features in principal components.
- **Mel spectrogram features**: Particularly in the lower frequency bands, these features provide important information about the energy distribution across frequencies.
- **MFCCs**: While fewer in number among the top features, the first MFCC (MFCC_0) is particularly important as it relates to the overall energy of the signal.

4.3 Audio Clusters

Three distinct clusters of audio features were identified:

- Cluster 0 (56.9% of regions): Characterized by short, distinct sounds like beeps, meows, and impacts
- Cluster 1 (33.4% of regions): Contains more continuous sounds like buzzing, barking, and sustained notes
- Cluster 2 (9.8% of regions): Consists primarily of musical and rhythmic sounds

4.4 Silent Regions

Silent regions did not form a distinct cluster but were distributed across all three clusters, with the highest concentration in Cluster 1 (13.3%). This suggests that silence is not uniform and may retain some characteristics of the surrounding audio context.

5 Conclusion

The audio feature analysis presented in this report provides valuable insights for developing sound event detection systems:

- The original high-dimensional feature space (942 dimensions) can be effectively reduced to 82 dimensions while preserving 80% of the variance.
- Mel spectrogram features, particularly in lower frequency bands, provide important information about energy distribution.
- Natural clustering of audio features reveals three primary groups of sounds with distinct characteristics.
- Silent regions show context-dependent properties rather than forming a single homogeneous group.