Analysis of Clustering Techniques Applied to Anuran Calls Dataset

Executive Summary

This report analyzes the application of various clustering algorithms to the Anuran Calls Dataset, which contains Mel-frequency cepstral coefficients (MFCCs) extracted from frog call recordings. The analysis encompasses K-means, Hierarchical, and DBSCAN clustering methods, with a focus on their effectiveness and limitations in grouping frog species based on acoustic features.

1. Overall Clustering Process Analysis

1.1 Data Preprocessing

The dataset consisted of 22 MFCC features from 7,195 samples. Key preprocessing steps included:

- Feature correlation analysis to remove highly correlated features (threshold: 0.95)
- Polynomial feature engineering on the first 5 MFCCs
- Standardization using StandardScaler
- Dimensionality reduction using PCA to 10 components

1.2 Optimal Number of Clusters

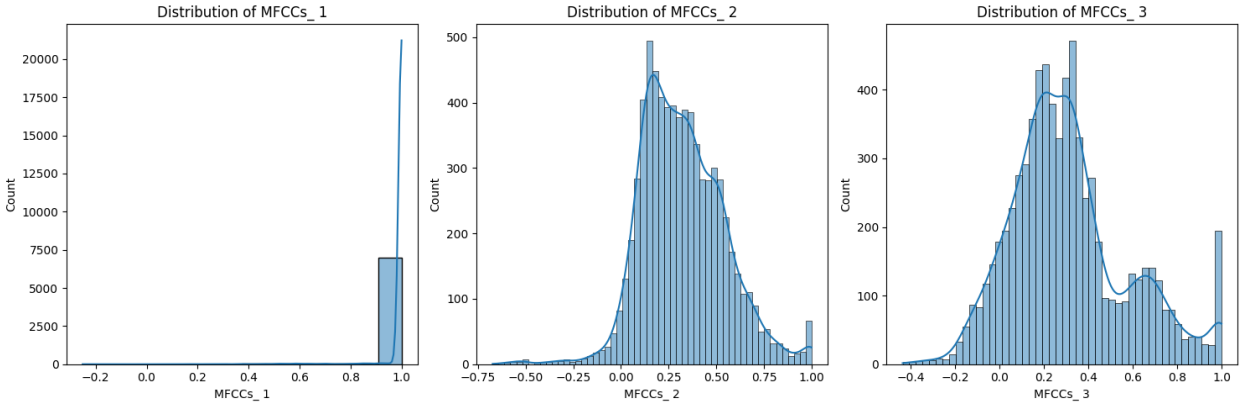
The elbow method analysis revealed:

- A distinct elbow at k=4 clusters
- This was supported by the silhouette score analysis
- The optimal number was consistently maintained across different initialization methods

1.3 Key Insights from Visualizations

Distribution Analysis

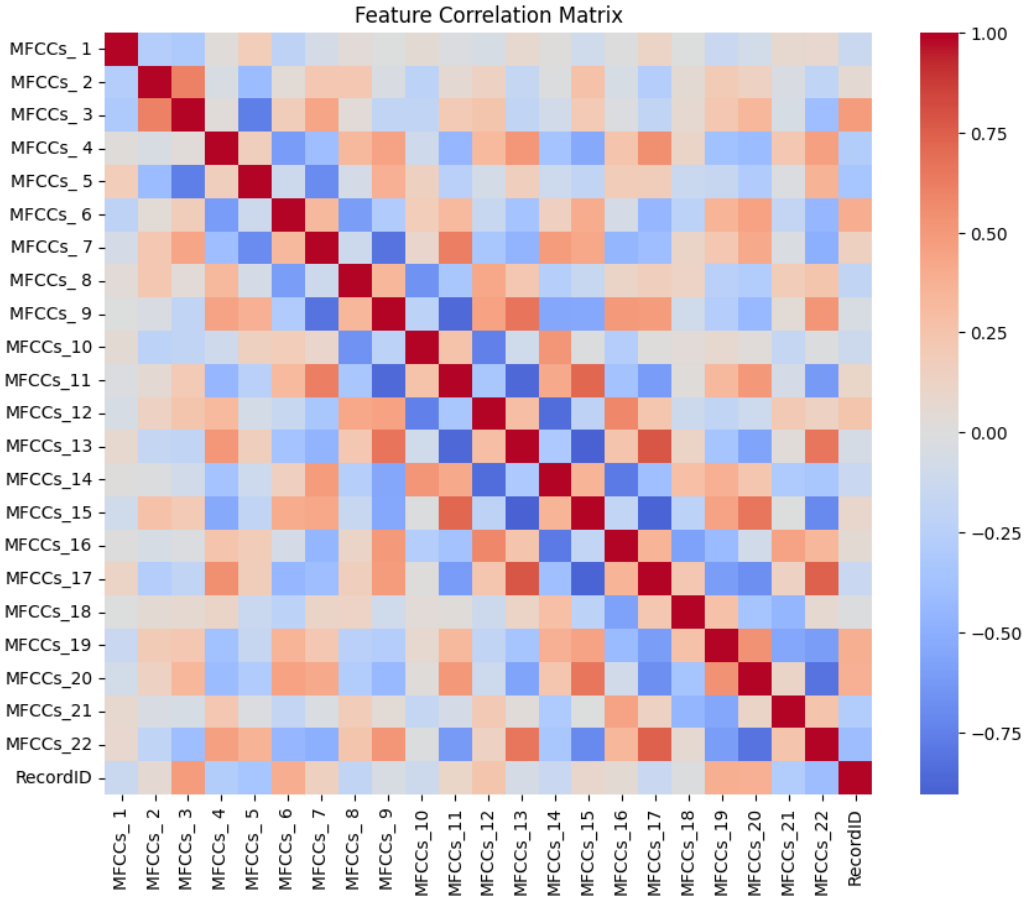
- The first few MFCCs showed varying distributions
- MFCC_1 displayed a strong right-skewed distribution with most values near 1.0
- MFCC_2 and MFCC_3 showed more balanced, roughly normal distributions

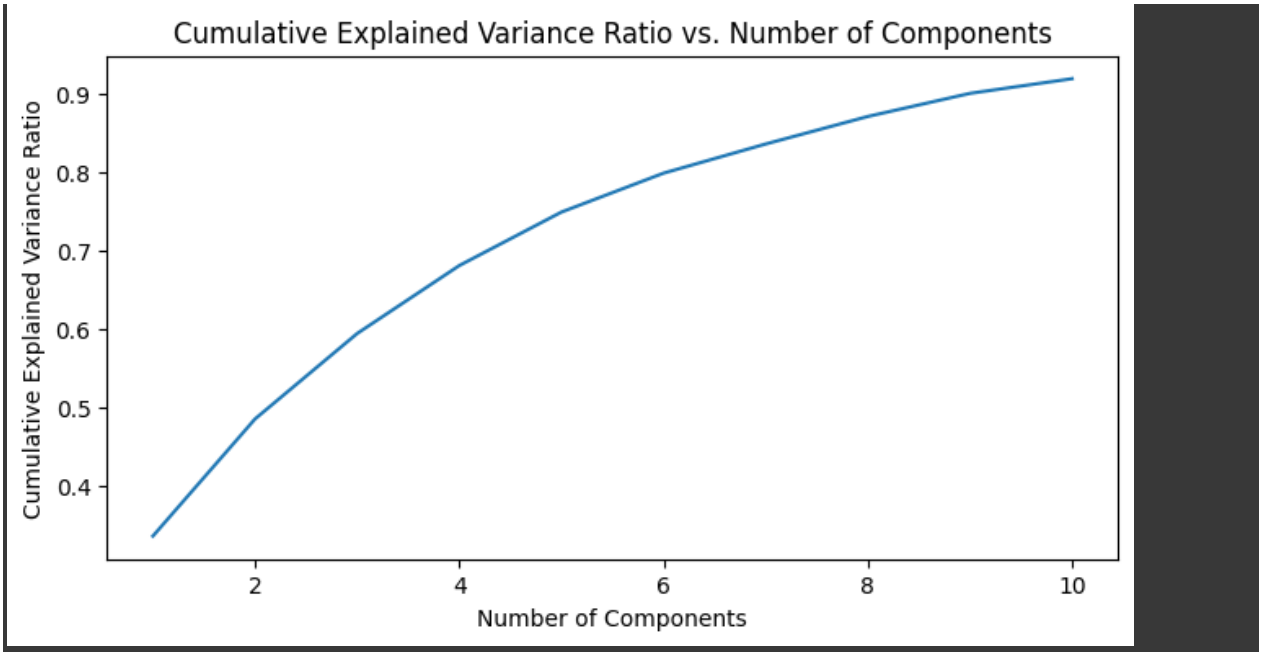


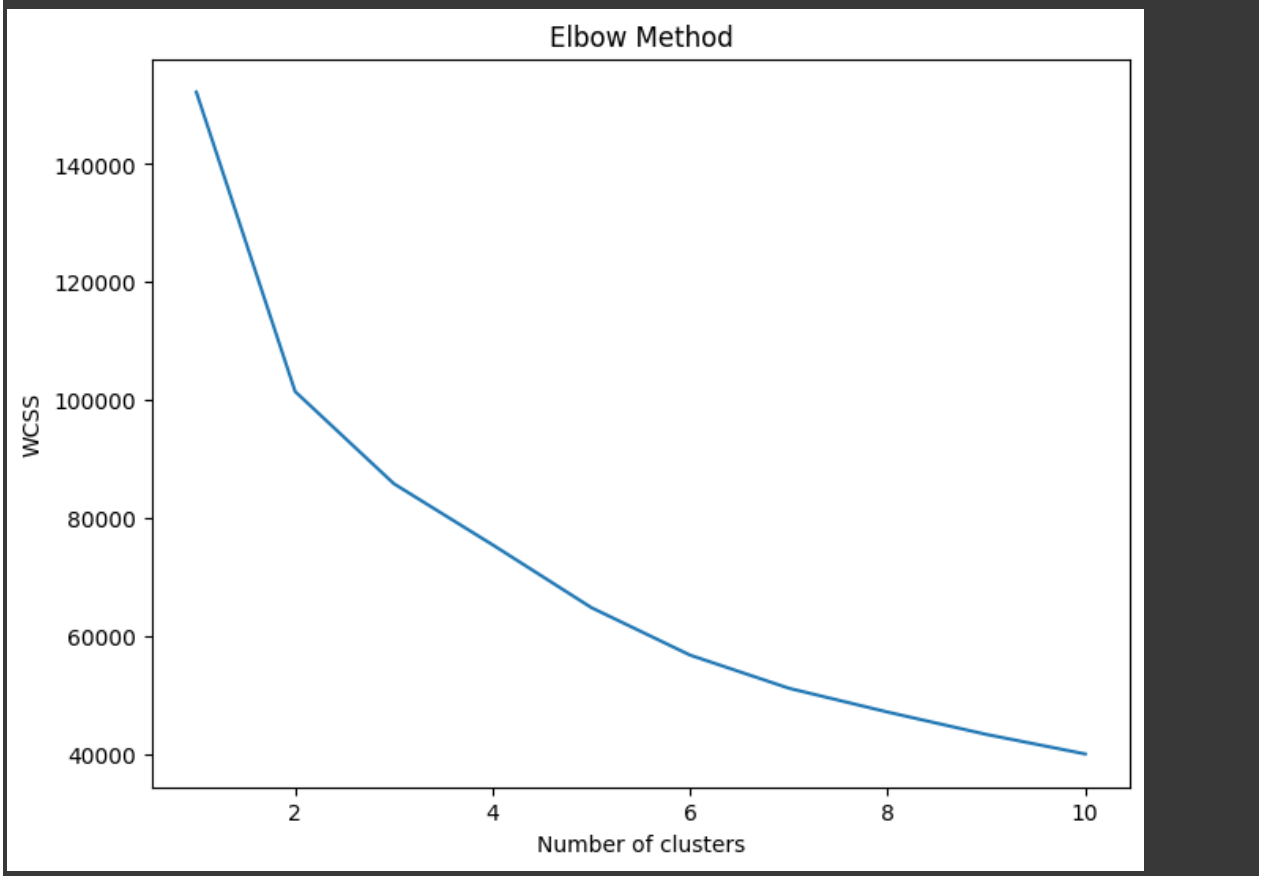
Feature Importance

The top 5 most important features for clustering were MFCCs 10, 22, 4, 8, and 6, suggesting that:

- Mid-range and high-frequency components are crucial for species differentiation
- The contribution of features is not linear with their order




Cumulative Explained Variance Ratio vs. Number of Components
Cumulative Explained Variance Ratio
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
Number of Components


Elbow Method
WCSS
140000
120000
100000
80000
60000
40000
2
4
6
8
10
Number of clusters

Cluster Visualization

The 2D PCA plots revealed:

- K-means showed well-defined, relatively balanced clusters
- Hierarchical clustering produced similar to marginally better cluster structures to K-means
- DBSCAN identified significantly more clusters (44) with less clear boundaries

1.4 Evaluation Metrics Analysis

K-means Performance:

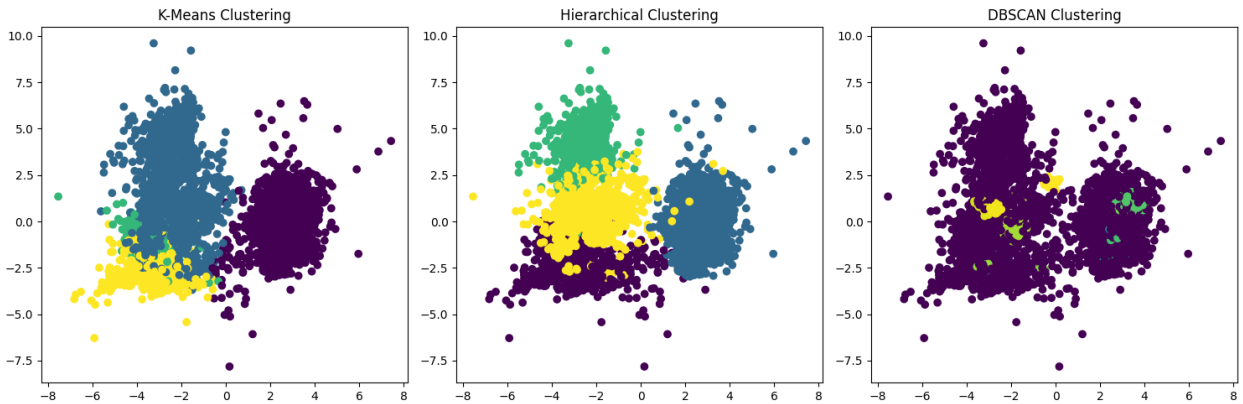
- Silhouette Score: 0.401
- Davies-Bouldin Index: 1.284
- Calinski-Harabasz Index: 2436.420

Hierarchical Clustering Performance:

- Silhouette Score: 0.366
- Davies-Bouldin Index: 1.499
- Calinski-Harabasz Index: 2366.501

DBSCAN Performance:

- Silhouette Score: -0.434
- Davies-Bouldin Index: 1.488
- Calinski-Harabasz Index: 23.272



2. Limitations and Applicability Analysis

2.1 K-means Limitations

1. **Assumption of Spherical Clusters**
    - The algorithm assumes clusters are spherical and of similar size
    - This may not reflect natural groupings in acoustic features
    - Could explain the moderate silhouette score (0.401)
2. **Sensitivity to Initialization**
    - Minimal difference between random (0.401) and k-means++ (0.401) initialization
    - Suggests robust cluster structure but potential local optima issues
3. **Fixed Number of Clusters**
    - Requires pre-specification of cluster number
    - May not capture natural groupings in frog species

2.2 Hierarchical Clustering Limitations

1. **Performance Metrics**
    - Slightly lower silhouette score (0.366) than K-means
    - Higher Davies-Bouldin index (1.499) indicating less compact clusters
2. **Computational Complexity**
    - $O(n^2)$ complexity makes it less suitable for larger datasets
    - May not scale well with additional frog species

2.3 DBSCAN Limitations

1. **Parameter Sensitivity**
    - Identified 44 clusters with current parameters
    - Negative silhouette score (-0.434) indicates potential over-segmentation
2. **Density Variations**
    - Poor performance suggests varying density patterns in the MFCC space
    - May not handle varying acoustic feature densities well

2.4 Dataset-Specific Challenges

1. **High Dimensionality**
    - 22 original MFCCs plus engineered features
    - Curse of dimensionality affects distance-based clustering
2. **Feature Interactions**
    - Complex relationships between MFCCs
    - Non-linear relationships may not be captured effectively

3. Conclusions

The analysis reveals that K-means clustering provides the most balanced and interpretable results for this dataset, despite its limitations. The moderate silhouette scores across algorithms suggest inherent complexity in the acoustic feature space. Future work could explore non-linear dimensionality reduction techniques and alternative distance metrics for improved cluster separation.