# Investigation on the Impact of Preprocessing Methods and Parameter Selection in Acoustic Scene Classification Based on K-means Clustering Algorithm

Yuanyao Zuo

Computer Science, University of Ottawa, Ottawa, K1N 6N5, Canada

yzuo023@uottawa.ca

**Abstract.** This research investigates the effectiveness of various preprocessing methods and parameters on Acoustic Scene Classification (ASC) using the K-means clustering algorithm. Utilizing the ESC-50 dataset, a combination of Principal Component Analysis (PCA) and StandardScaler was employed for preprocessing. The study's key findings include the identification of an optimal number of PCA components, around 30, which maximized the accuracy of the K-means algorithm. Additionally, the results revealed an unexpected phenomenon where increasing the number of clusters beyond the actual class count improved the model's accuracy, indicating potential nuanced sub-groupings within classes. These insights highlight the significance of preprocessing methods and the choice of parameters on the performance of ASC models. However, the findings may not be universally applicable across other datasets or feature sets. The study offers potential directions for future research, suggesting the exploration of other machine learning algorithms and further investigation into the potential sub-groupings within classes.

**Keywords:** Acoustic Scene Classification, K-means Clustering Algorithm, Machine Learning Algorithms

## 1. Introduction

In an ever-evolving world saturated with audio stimuli, the ability to discern and classify acoustic scenes has become a paramount challenge in the field of audio processing. From bustling city streets and serene nature reserves to lively concerts and tranquil libraries, each environment possesses a distinct soundscape that not only defines its character but also offers valuable insights into the dynamics of our surroundings. Acoustic Scene Classification (ASC) emerges as a fascinating discipline aimed at deciphering the complex symphony of audio signals, facilitating the comprehension, differentiation, and ultimate comprehension of the diverse acoustic environments that profoundly influence our daily existence.

Despite the extensive array of approaches put forward in this domain, the pursuit of achieving robust and dependable surveillance of acoustic environments continues to encounter significant impediments [1]. These challenges encompass a range of factors, including the presence of various overlapping audio sounds, persistent background


P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_30

noises, and the absence of comprehensive and universally applicable multimodal datasets [1]. The intricate nature of these complexities renders the accurate identification and classification of audio scenes and sound events an inherently challenging task, thereby emphasizing the urgent necessity for efficacious solutions within the domain of environmental audio surveillance.

The domain of deep learning in machine learning has witnessed remarkable progress across diverse fields, including natural language processing and computer vision, owing to its rapid advancement. Especially, It has been considered an effective approach to audio recognition tasks [2]. One of the key advantages of AI-based audio recognition is its adaptability and generalization capabilities. Artificial intelligence models can be trained on large-scale datasets containing a wide range of acoustic scenarios, allowing them to learn and generalize from different audio environments [3]. This flexibility allows the models to perform well even in previously unseen or challenging scenarios, making them highly applicable in real-world audio surveillance applications. The integration of Artificial Intelligence of Things (AIoT) with Low-Power Wide Area Network (LPWAN) technologies has led to a groundbreaking framework known as LPWAN-based AIoT (LPAI) [4]. This new development has revolutionized acoustic scene classification, providing efficient and precise categorization [4]. LPAI utilizes the power-saving and wide coverage advantages of LPWAN as its underlying infrastructure, improving the overall effectiveness of AIoT applications [4]. This pioneering framework, being the first of its kind, opens doors for leveraging AIOT potential in understanding and utilizing audio signals in real-world applications. It also demonstrates the potential in solving the challenges associated with accurate acoustic scene classification. Furthermore, researchers have used class hierarchy construction methods for classification errors to improve effectiveness in terms of ASC performance, and fusing similarity relations and multi-task learning frameworks offers great potential for solving inter-class similarity challenges and improving the accuracy of ASC systems [5]. While many articles primarily highlight the high accuracy achieved in their experimental outcomes, the influence of specific parameters on these results is seldom explored. In this research, the emphasis is placed on investigating the effects of both model parameters and preprocessing parameters and measures on the performance of ASC models. Specifically, the focus is on investigating the impact of different techniques, including Principal Component Analysis (PCA), normalization, and their combinations, on enhancing the accuracy of the k-means clustering algorithm in ASC.

## 2. Method

### 2.1. Dataset Description

In this study, the ESC-50 dataset, a publicly available collection of 2, 000 environmental audio recordings was considered. This dataset is evenly arranged across 50 categories [6]. The ESC-50 provides annotated data, aiding in the examination and interpretation of various environmental audio aspects [6]. The audio clips in the dataset are five seconds long, and they are drawn from the Freesound project, a collaborative database of Creative Commons Licensed sounds. These recordings are subsequently organized into 50 different categories. The categories

include animal sounds, natural soundscapes, and human non-speech sounds, among others, providing a broad scope for audio analysis. The main objective of the ESC-50 dataset is to provide a benchmark for environmental sound classification, primarily focusing on the robustness of machine learning algorithms in the face of real-world conditions [6]. Each category consists of 40 audio clips, resulting in a total of 2000 clips throughout the dataset.

### 2.2. Preprocessing

In this research, a comprehensive multi-step preprocessing protocol was implemented to prepare the ESC-50 dataset for further stages of our analysis. This process incorporated two key aspects: PCA and normalization, both of which are vital elements in the processing of machine learning data.

PCA [7], is a statistical procedure that orthogonally transforms the original 'n' dimensions of a dataset into a new set of 'n' dimensions known as Principal Components. These components are ordered according to the proportion of the dataset's variance that they contain, with the first principal component accounting for the largest variance and each succeeding component accounting for less [8]. By focusing on the components that account for the most variance, PCA allows for the reduction of a dataset's dimensionality without losing significant information, which is particularly beneficial in high-dimensional data such as the audio signals that are being used in this study.

Normalization, also known as feature scaling, was another essential preprocessing step employed in our study. As various features in a dataset can differ significantly in their magnitudes, units, and range, models trained on unscaled data can lead to weights that are too large or too small, causing the model to underperform [9]. By applying the StandardScaler method for normalization, which standardizes features to have a mean of 0 and a standard deviation of 1, it can ensure that the developed model did not become skewed or biased towards larger values, thus leading to more accurate and reliable model training [10].

After these preprocessing procedures, feature extraction was performed on the audio files using Librosa to obtain the Mel-frequency Cepstral Coefficients (MFCCs). The MFCCs features were then averaged across frames to produce a fixed-size input for our classifiers. Class labels were converted to integer indices using the LabelEncoder from the sci-kit-learn library. The dataset was then split into training and testing sets with an 80:20 ratio, providing a robust basis for further analysis and model training.

### 2.3. Proposed Approach - K-means

In the vast array of unsupervised machine learning methods, clustering algorithms serve a critical role, especially in the field of acoustic scene classification. In this study, the K-means algorithm, a well-respected and widely employed clustering technique, was used. The k-means algorithm segments the dataset into 'K' distinct non-overlapping subgroups (clusters) such that each data point belongs to the group with the closest mean value, thereby forming coherent clusters [11, 12].

In terms of the preprocessing stage, the features called MFCCs were extracted first, followed by standardization. Subsequently, this study used PCA to reduce the dimensionality of the dataset. With PCA's aid, the dataset's dimensionality was reduced to 20 components while retaining the most significant information. The sci-

kit-learn library's PCA implementation made it possible to perform this operation with ease [13].

Following preprocessing and dimensionality reduction, the K-means clustering algorithm from the sci-kit-learn library was implemented on the transformed data. The number of clusters was set equal to the number of unique classes present in the training set. In addition, this study used a 'cluster-to-class' mapping where each cluster was associated with the most common class in the cluster [11]. The performance of the K-means clustering approach was evaluated using an accuracy score, one of the simplest yet most robust metrics to assess the performance of a classification model.

## 3. Results and Discussion

Table 1 reveals the accuracy outcomes from various configurations of the K-means algorithm on the ESC-50 dataset, spotlighting the effects of different preprocessing methods and parameters on the performance. Initially, this study tested the K-means algorithm using only PCA or StandardScaler for preprocessing. The resultant accuracy scores revealed a substantial effect of feature scaling, proving it to be more impactful than dimensionality reduction alone in enhancing the model's accuracy.

**Table 1.** Performance Evaluation of K-means Algorithm with Variations in Preprocessing and Parameters.

| Methods | Number of Principal Components | Number of Clusters | Accuracy |
|---|---|---|---|
| PCA Only | - | - | 0.1125 |
| StandardScaler Only | - | - | 0.1475 |
| PCA + StandardScaler | 20 | 50 (default) | 0.1625 |
| PCA + StandardScaler | 30 | 50 (default) | 0.1725 |
| PCA + StandardScaler | 40 | 50 (default) | 0.1400 |
| PCA(30) + StandardScaler + KMeans | 30 | 45 | 0.1350 |
| PCA(30) + StandardScaler + KMeans | 30 | 50 | 0.1525 |
| PCA(30) + StandardScaler + KMeans | 30 | 55 | 0.1850 |

The influence of applying both feature scaling and PCA with varying component numbers was further investigated. The accuracy intriguingly increased from 0.1625 with 20 components to 0.1725 with 30 components but fell to 0.14 with 40 components. This suggests an optimal number of PCA components, approximately 30, that maximizes the K-means algorithm's accuracy for this context.

The final set of results in Table 1 pertains to the fluctuation in the number of clusters employed by the K-means algorithm. The data demonstrated an ascending trend in accuracy scores as the cluster count increased from 45 to 55, culminating in the highest accuracy of 0.185 with 55 clusters. This insinuates that increasing the cluster count beyond the number of actual classes might yield improved performance in certain cases.

The research offers crucial insights into employing the K-means algorithm for environmental sound classification. It confirmed the significant role of data preprocessing in algorithm accuracy, as corroborated by previous studies emphasizing feature scaling's importance [9]. Notably, the results indicated an optimal number of PCA components, underlying the need to strike a balance in capturing adequate original data information without introducing noise or overfitting through lesser significant dimensions. Moreover, the study unveiled an unexpected insight regarding the number of clusters. Counter to initial assumptions, the model's performance improved when increasing the cluster count beyond the actual classes, hinting at the presence of nuanced sub-groupings within classes.

However, this study is still confined by some limitations. The optimal parameters discerned in this study are specific to the ESC-50 dataset and the chosen features, and may not be generalized to other audio datasets or feature sets. While the K-means algorithm demonstrated promising results, future research could explore other machine-learning algorithms for potential improvements in environmental sound classification.

## 4. Conclusion

In this study, an in-depth exploration of the ASC domain was undertaken, focusing on the effectiveness of preprocessing methods and parameter selection in conjunction with the K-means clustering algorithm. Through meticulous analysis and experimentation, key insights emerged that shed light on the intricate relationship between data preprocessing, parameter tuning, and ASC performance. The combination of PCA and StandardScaler demonstrated their pivotal roles in enhancing the accuracy of the K-means algorithm. The identification of an optimal number of PCA components, approximately 30, attests to the delicate balance required to capture pertinent information while mitigating the introduction of noise and overfitting. Intriguingly, an increase in the number of clusters beyond the actual class count improved model accuracy, suggesting potential nuanced sub-groupings within classes.

While the findings underscore the substantial impact of preprocessing methods and parameter choices on ASC performance, it's imperative to acknowledge the context-specific nature of these insights. The optimal parameters discerned within this study are closely tied to the ESC-50 dataset and the chosen feature set, thus cautioning

against their universal applicability. The diverse landscape of audio datasets and features necessitates a tailored approach when extending these findings to different scenarios. Therefore, future research is encouraged to explore other machine learning algorithms and investigate the potential sub-groupings within classes for improved ASC accuracy.

## References


1. Chandrakala, S., & Jayalakshmi, S.: Environmental Audio Scene and Sound Event Recognition for Autonomous Surveillance: A Survey and Comparative Studies. ACM Computing Surveys, 52(3), 1–34 (2019).
2. Singh, M. et al.: Audio Recognition Using Deep Learning for Edge Devices, in Advances in Computing and Data Sciences. Switzerland: Springer International Publishing AG (2022).
3. Yousefpour, N., & Pouragha, M.: Prediction of the post-failure behavior of rocks: Combining artificial intelligence and acoustic emission sensing. International journal for numerical and analytical methods in geomechanics. 46 (10), 1874–1894 (2022).
4. Jing, X., et al.: LPAI—A Complete AIoT Framework Based on LPWAN Applicable to Acoustic Scene Classification Scenarios. Sensors (Basel, Switzerland). 22 (23), (2022).
5. Zheng, W., et al.: Clustering by Errors: A Self-Organized Multitask Learning Method for Acoustic Scene Classification. Sensors (Basel, Switzerland). 22 (1), (2021).
6. Piczak, K. J.: ESC: Dataset for Environmental Sound Classification. Proceedings of the 23rd Annual ACM Conference on Multimedia, 1015-1018. (2015) https://github.com/karolpiczak/ESC-50
7. Pearson, K.: On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 2(11), 559-572 (1901).
8. Abdi, H., & Williams, L. J.: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459 (2010).
9. Raschka, S.: Python Machine Learning. Packt Publishing Ltd (2014).
10. Jain, A., Zongker, D.: Feature selection: Evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2), 153-158 (1997).
11. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830 (2011).
12. Yu. Q., et al.: Clustering Analysis for Silent Telecom Customers Based on K-means++, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, pp. 1023-1027 (2020).
13. Zhang, X., et al.: StandardScaler: A scikit-learn function to center and scale the dataset. Neurocomputing, 267, 406-416 (2016).