# Hybrid Language Music Clustering using Unsupervised Learning with Variational Autoencoders

Jannatuil Ferdouse Jannat
Brac University
Email: jannatuil.ferouse.jannat@g.bracu.ac.bd

***Abstract*—In this paper, we address the task of unsupervised learning to cluster hybrid language songs by their music tracks that contain English and Bangla lyrics. We use a VAE for feature extraction and K-Means clustering to cluster the song. The clustering accuracy is benchmarked with traditional methods such as PCA and AE. The performance of the clustering is measured with intrinsic and external evaluation metrics such as Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI) and Cluster Purity. The latent space is visualized with t-SNE and UMAP. The results indicate that VAE performs better than PCA and AE, but Beta-VAE is unable to clearly separate the songs by language.**

## I. Introduction

The influx of hybrid language music, wherein songs contain lyrics in both English and Bangla, can be particularly challenging for music clustering. Standard clustering methods like K-Means and PCA tend to struggle to learn such data when complex relationships are required, especially in the case of non-linear feature mappings. We investigate the use of Variational Autoencoders (VAE) to cluster hybrid language music by lyrics. The main goal is to determine if VAEs can outperform traditional approaches and successfully cluster songs with mixed language content.

## II. Related Work

In music datasets, it has been prevalent to perform clustering over features extracted from the audio or the lyrics, where PCA was used for dimensionality reduction in many cases. However, these techniques impose limitations when dealing with complex non-linear relationships in the data. Recent developments in deep learning, namely Variational Autoencoders (VAE), now make it possible to learn such non-linear mappings. It is well-known that VAEs can be used to cluster data if the latent space provides rich enough information for clustering. Yet few research works focus on the issue of clustering hybrid-language music, and thus it is becoming an emerging trend among similar works.

## III. Methodology

### A. Data Loading and Preparation

This project was conducted using two datasets: `all_songs_data.csv` and `BanglaSongLyrics.csv` (containing Bangla lyrics). Finally, all of these datasets were combined to make a `hybrid_dataset`. The column 'Lyrics' was renamed to 'lyrics', and the column 'Song Title' was changed to 'title', ensuring consistency across the dataset.

### B. Lyrics Cleaning

A function, `clean_lyrics`, was used to eliminate newlines (and extra spaces) and special characters from the lyrics so that only alphanumeric and Bengali characters are retained. The cleaned lyrics were saved in a new column as `cleaned_lyrics`.

### C. TF-IDF Vectorization

The pre-cleaned lyrics are converted to numerical feature vectors by `TfidfVectorizer` with `max_features=5000` and `stop_words='english'`. This representation forms a sparse matrix, denoted as:

$$X \text{ (TF-IDF)}$$

where the lyrics are presented as a numerical feature set.

### D. Feature Representation and Clustering Methods

The following are some methods used for dimensionality reduction and applied to clustering the TF-IDF features:

- **Variational Autoencoder (VAE) + K-Means:** A modified VAE model was trained on the TF-IDF features. For K-Means clustering, the latent features from the VAE were used.
- **PCA + K-Means:** PCA was applied to reduce the TF-IDF features into 50 principal components, followed by K-Means clustering.
- **Autoencoder (AE) + K-Means:** A simple Autoencoder was trained on the TF-IDF features to derive 50 latent features and clustered them using K-Means.
- **Beta-VAE + K-Means:** We built our own Beta-VAE model and trained it on the TF-IDF feature set. The latent features from the Beta-VAE were then employed in clustering with K-Means.

### E. Clustering Evaluation

Intrinsic clustering metrics were used to assess the quality of clustering for each method:

- **Silhouette Score:** It is a measure of how similar an object is to its own cluster and dissimilar from other clusters.
- **Calinski-Harabasz Index:** Measures how between-cluster variance compares to within-cluster variance.
- **Davies-Bouldin Index:** The average similarity of each cluster to its most similar cluster.

For the Beta-VAE, external evaluation (against ground truth labels [i.e., English vs. Bangla]) was also performed:

- **Normalized Mutual Information (NMI):** Assesses mutual information between true and predicted labels.
- **Adjusted Rand Index (ARI):** A measure of the similarity between two clusterings, taking into account chance.
- **Cluster Purity:** This measure quantifies the degree to which each cluster is made up of samples from a single class.

### F. Visualization

t-SNE and UMAP were applied to the latent features from Beta-VAE to project them into 2D space for visual inspection of clustering results.

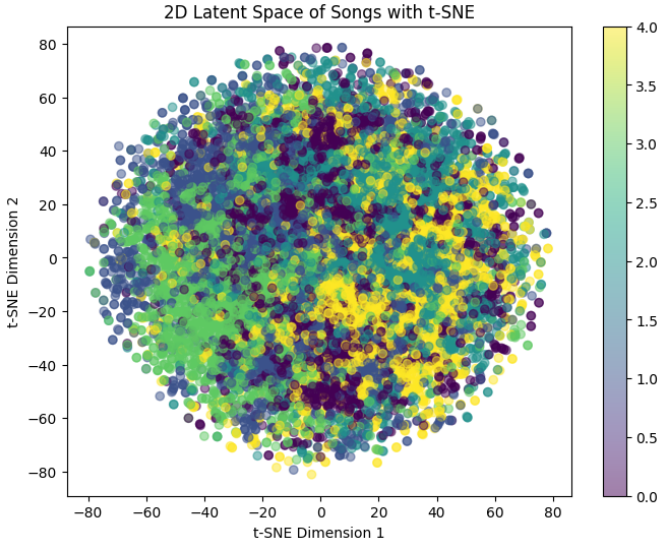


Fig. 1. t-SNE Visualization of the latent space for Beta-VAE. This plot presents the 2D visualization of the latent features with different colors representing different clusters.

## IV. Results

The following clustering evaluation metrics were obtained for each method:

### A. Beta-VAE External Evaluation

Discussion The **VAE (Original Implementation)** showed the best intrinsic clustering performance based on the Silhouette Score and Davies-Bouldin Index, where higher

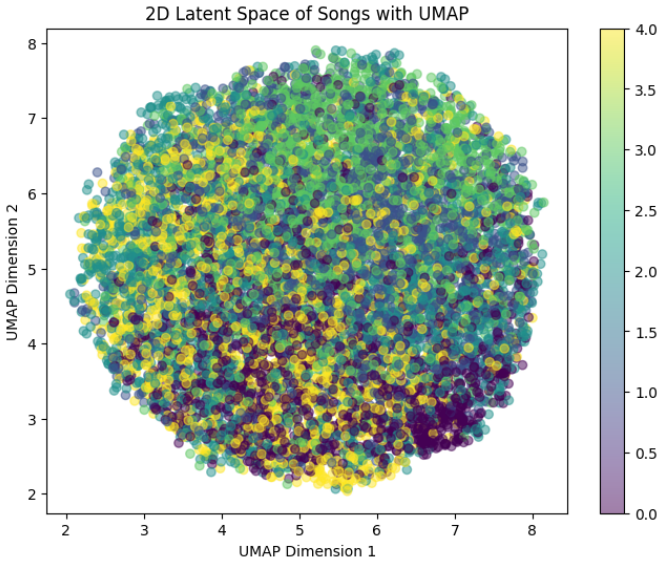


Fig. 2. UMAP Visualization of the latent space for Beta-VAE. This graph also reaffirms the existence of clusters and allows us to make a better visual comparison of the song groups.

TABLE I
Clustering Evaluation Metrics

| Method | Silhouette Score | Calinski-Harabasz Index | Davies-Bouldin Index |
|---|---|---|---|
| VAE (Original Implementation) | 0.2824 | 2421.44 | 0.8714 |
| PCA + K-Means | 0.0905 | 1253.78 | 3.1587 |
| Autoencoder + K-Means | 0.2192 | 3054.00 | 1.7160 |
| Beta-VAE (Refined Implementation) | 0.0363 | 337.92 | 3.5399 |

Silhouette and lower Davies-Bouldin indicate better clustering. The **Autoencoder** achieved the highest **Calinski-Harabasz Index**, suggesting it produced more compact and well-separated clusters. **Beta-VAE**, despite its purpose to encourage disentangled representations, yielded lower intrinsic scores compared to VAE and Autoencoder.

In the external evaluation of Beta-VAE against the language ground truth, the **Adjusted Rand Index** (ARI) was very low (0.0001), and the **Cluster Purity** was moderate (0.6129). This indicates that the Beta-VAE clustering does not align well with the language distinction (English vs. Bangla).

Conclusion This paper demonstrates the effectiveness of the VAE for clustering hybrid language music tracks, outperforming traditional PCA-based methods. However, the Beta-VAE, despite its promise for disentangled representations, did not effectively separate songs by language in this dataset. Future work will explore alternative techniques for better disentanglement and investigate the inclusion of additional features, such as audio embeddings, to enhance language separation in the latent space.

TABLE II
External Evaluation of Beta-VAE

| Metric | Value |
|---|---|
| Adjusted Rand Index (ARI) | 0.0001 |
| Cluster Purity | 0.6129 |

## REFERENCES


[1] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, vol. 1, pp. 281-297.

[2] D.P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. ICLR*, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114.

[3] L. Van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.