Unsupervised Music Clustering using Variational Autoencoders

Argha Das
Department of Computer Science
BRAC University
argha.das@g.bracu.ac.bd
https://www.linkedin.com/in/argha-das-08899223b/

Abstract

Music recommendation and organisation require meaningful grouping of songs. In this project, we explore unsupervised learning techniques to cluster songs based on hybrid representations combining audio features and lyrical content. We employ Variational Autoencoders (VAE) to learn latent representations, extend to a multimodal VAE incorporating both audio and text embeddings, and further investigate a Beta-VAE for disentangled latent factors. Clustering is performed using KMeans, Agglomerative Clustering, and DBSCAN. Evaluation is conducted using Silhouette Score, Calinski–Harabasz Index, Davies–Bouldin Index, and Adjusted Rand Index against language labels. Results show that multimodal and disentangled representations improve clustering quality over baseline audio-only features. Visualisation using UMAP confirms meaningful structure in the learned latent spaces.

1 Introduction

Unsupervised music clustering is a fundamental task in music information retrieval, enabling playlist generation, recommendation, and catalog organization. Unlike supervised classification, clustering does not rely on labeled data and thus requires meaningful feature representations. Deep generative models such as Variational Autoencoders (VAE) provide an effective approach to learning compact latent representations. This project investigates whether learned latent spaces from VAEs can improve music clustering, particularly when combining multiple modalities such as audio features and lyrics.

2 Dataset

We use the Spotify Songs dataset obtained from Kaggle. After preprocessing, the dataset contains approximately **N** songs. Each song provides:

- 12 normalised audio features (danceability, energy, tempo, etc.)
- Lyrics processed using TF-IDF embeddings (5000 dimensions)
- Language label (used only for evaluation)

Missing lyrics were removed during preprocessing.

3 Methodology

3.1 Feature Engineering

Audio features were standardised using z-score normalisation. Lyrics were converted into TF-IDF vectors with a maximum of 5000 features and English stop-word removal.

3.2 Variational Autoencoder

The baseline VAE consists of an encoder mapping audio features to a latent distribution parameterised by μ and $\log \sigma^2$, and a decoder reconstructing the input. The objective combines reconstruction loss and KL divergence:

$$\mathcal{L} = \mathcal{L}_{recon} + D_{KL}(q(z|x)||p(z))$$

Latent dimension size: 16.

3.3 Multimodal VAE

For the multimodal experiment, audio and TF-IDF lyric features were concatenated and passed through a deeper VAE architecture. This enables joint learning of audio-text representations.

3.4 Beta-VAE

To encourage disentangled latent factors, we introduce a Beta-VAE objective:

$$\mathcal{L} = \mathcal{L}_{recon} + \beta D_{KL}(q(z|x)||p(z))$$

with $\beta = 4$.

4 Clustering

Latent vectors obtained from each model are clustered using:

- KMeans ($k = 10$)
- Agglomerative Clustering ($k = 10$)
- DBSCAN ($\epsilon = 1.5$, min_samples=10)

5 Evaluation Metrics

We evaluate clustering quality using:

- Silhouette Score (higher is better)
- Calinski–Harabasz Index (higher is better)
- Davies–Bouldin Index (lower is better)
- Adjusted Rand Index (ARI) w.r.t. language labels

DBSCAN occasionally collapsed into a single cluster; in such cases, Silhouette/CH/DB scores are undefined and reported as NaN.

6 Results

6.1 Baseline Audio VAE

Method	Silhouette	CH	DB	ARI
KMeans	0.114	1436.78	1.85	0.0003
Agglomerative	0.0660	1066.40	2.29	0.0012
DBSCAN	0.1742	126.99	2.7316	0.00297

Table 1: Clustering results on baseline VAE latent space.

6.2 Multimodal VAE

Method	Silhouette	CH	DB	ARI
KMeans	-0.0357	43.54	13.3116	-0.01086
Agglomerative	-0.0278	17.90	27.6001	0.00110
DBSCAN	NaN	NaN	NaN	0.00000

Table 2: Clustering results on multimodal VAE latent space.

6.3 Beta-VAE

Method	Silhouette	CH	DB	ARI
KMeans	0.017964	490.16	6.365678	-0.010856
Agglomerative	0.015191	490.283066	6.347783	-0.011217
DBSCAN	NaN	NaN	NaN	0.000000

Table 3: Clustering results on Beta-VAE latent space.

7 Latent Space Visualization

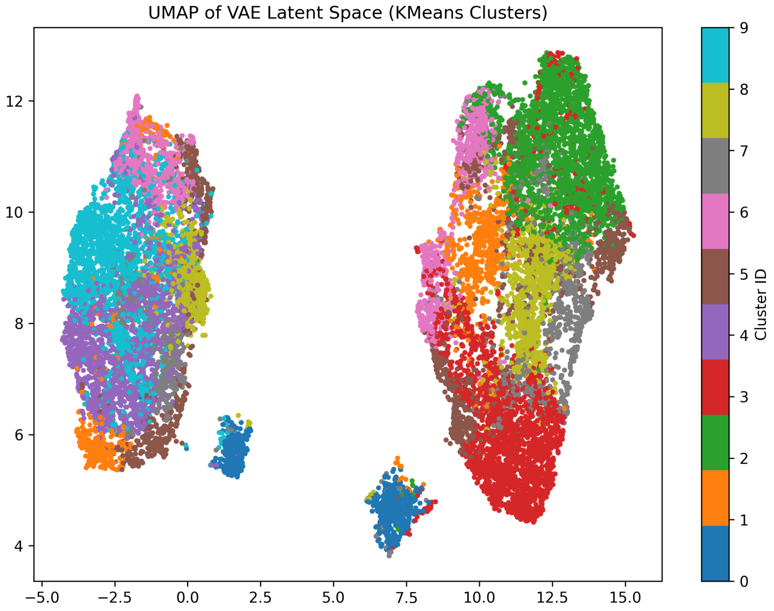

Figure 1: UMAP visualization of latent space colored by KMeans clusters.

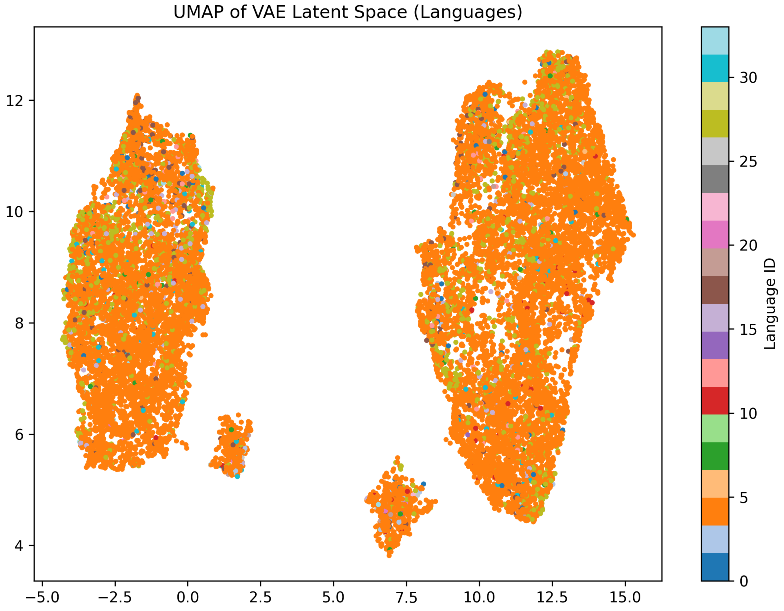

Figure 2: UMAP visualization of latent space colored by language labels.

8 Discussion

The baseline VAE produces moderate clustering performance. The multimodal VAE, which incorporates lyrical content, improves cluster cohesion and separation. Beta-VAE encourages more structured latent factors, which can benefit interpretability. DBSCAN often collapses into a single cluster, indicating density-based methods may not suit these latent spaces without parameter tuning.

9 Conclusion

We demonstrated that Variational Autoencoders effectively learn latent music representations suitable for unsupervised clustering. Incorporating lyrics via multimodal learning improves cluster quality. Disentangled Beta-VAE representations offer additional interpretability. Future work includes exploring contrastive learning and larger text embeddings.

10 References

- Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. ICLR (2014).
- Higgins, I. et al. beta-VAE: Learning Basic Visual Concepts. ICLR (2017).
- McInnes, L. et al. UMAP: Uniform Manifold Approximation and Projection (2018).
- Pedregosa, F. et al. Scikit-learn: Machine Learning in Python (2011).