Variational Autoencoder-Based Unsupervised Clustering of Music Audio Using IRMAS

Sourav Das
Student ID: 23141072
Department of Computer Science
BRAC University

Abstract

In this project, I explore unsupervised clustering of music audio using Variational Autoencoders (VAEs) in order to learn compact and structured latent representations from spectrogram features. I use the IRMAS dataset and extract log-Mel spectrograms from raw audio as input to convolutional autoencoders and a Beta-VAE. Clustering is performed on the learned latent embeddings using K-Means and Agglomerative Clustering, and evaluated using both intrinsic clustering metrics and label-alignment metrics. My results show that the Beta-VAE produces significantly better cluster separation than PCA and a standard autoencoder, achieving a Silhouette score of 0.47 compared to 0.13 for PCA and 0.11 for the autoencoder. These findings suggest that disentangled latent representations are effective for unsupervised music clustering.

1 Introduction

Unsupervised representation learning is an important problem in music information retrieval, particularly when labeled data is limited or unavailable. Instead of relying on supervised classification, clustering-based approaches aim to discover structure directly from audio signals. Traditional techniques such as Principal Component Analysis (PCA) operate linearly on handcrafted features and often fail to capture the complex spectral and temporal patterns present in music audio.

In this project, I investigate whether Variational Autoencoders (VAEs) can learn more meaningful latent representations for unsupervised music clustering. VAEs provide a probabilistic framework that encourages smooth and structured latent spaces. I focus on convolutional architectures applied to log-Mel spectrograms and compare three approaches: PCA, a convolutional autoencoder, and a convolutional Beta-VAE. All experiments are conducted on the IRMAS dataset.

2 Related Work

Early work in music clustering typically relied on spectral features such as MFCCs combined with clustering algorithms like K-Means or hierarchical clustering. While these approaches are simple and efficient, they often struggle to represent the nonlinear structure of audio data. Autoencoders have been proposed as nonlinear dimensionality reduction techniques, but standard autoencoders lack constraints on the latent space and may learn entangled representations.

Variational Autoencoders introduce a variational objective that enforces a prior distribution over latent variables, leading to smoother and more interpretable embeddings. Beta-VAEs extend this idea by weighting the KL-divergence term, encouraging disentanglement. These properties motivate the use of Beta-VAEs for unsupervised music representation learning.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

3 Method

3.1 Feature Extraction

I resample all audio clips to 22,050 Hz and truncate or zero-pad them to a duration of 3 seconds. For each clip, I compute a log-Mel spectrogram with 128 Mel bins. The spectrograms are globally normalized and represented as 128 × 130 time–frequency matrices, preserving both spectral and temporal information.

3.2 Representation Learning Models

I evaluate three representation learning methods:

- **PCA baseline**: PCA is applied to flattened spectrograms with 32 components.
- **Convolutional Autoencoder (AE)**: A CNN-based encoder–decoder trained using reconstruction loss.
- **Convolutional Beta-VAE**: A CNN-based VAE with a 32-dimensional latent space and $\beta = 4$, which encourages disentangled latent representations.

The Beta-VAE objective combines a reconstruction loss with a KL-divergence term weighted by β.

3.3 Clustering

After training, I apply K-Means and Agglomerative Clustering to the learned latent representations. The number of clusters is fixed at 11 to match the number of instrument classes. Ground-truth labels are not used during training and are only employed for evaluation.

4 Experiments

4.1 Dataset

I conduct experiments on the IRMAS training dataset, which contains 6,705 audio samples from 11 different instrument classes. Each sample consists of an isolated instrument recording. Folder names are used as ground-truth labels only for evaluation, ensuring that the clustering process remains fully unsupervised.

4.2 Training Details

All models are trained using PyTorch on a GPU. The Beta-VAE is trained for 35 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 128. The latent dimensionality is fixed at 32 for all learned representations to ensure a fair comparison.

5 Results

5.1 Quantitative Evaluation

Table 1 summarizes clustering performance across all methods.

Table 1: Clustering performance comparison.

Method	Silhouette	CH	DB	ARI	NMI	Purity
PCA(32)+KMeans	0.129	2321	1.87	0.056	0.110	0.241
AE(32)+KMeans	0.109	1351	2.07	0.064	0.122	0.258
BetaVAE(32)+KMeans	**0.475**	**35889**	**0.58**	0.042	0.092	0.214
BetaVAE(32)+Agglomerative	0.411	29225	0.61	0.037	0.089	0.209

The Beta-VAE achieves a substantial improvement in intrinsic clustering metrics. The Silhouette score increases dramatically, indicating much better cluster separation compared to PCA and the autoencoder. Lower Davies–Bouldin values further suggest tighter and more compact clusters.

5.2 Latent Space Visualization

Figure 1 shows a UMAP projection of the Beta-VAE latent space. The visualization reveals clear nonlinear structure and partial grouping of instrument classes, which is not observed in PCA-based embeddings.

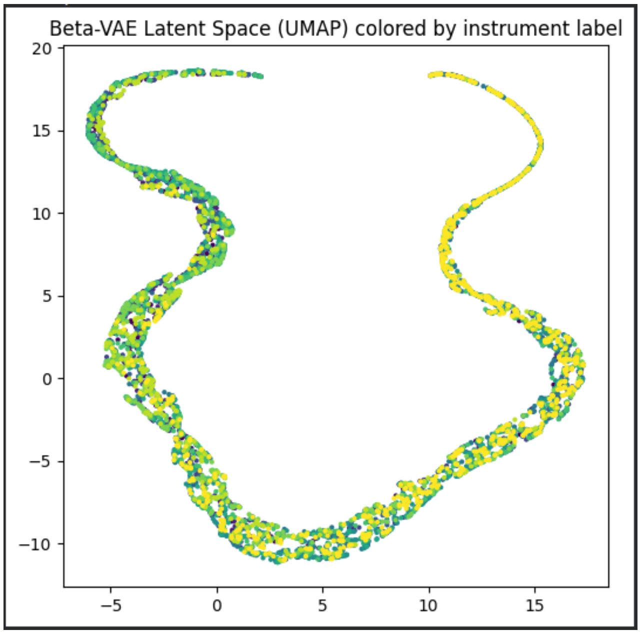

Figure 1: UMAP projection of Beta-VAE latent space colored by instrument label.

5.3 Reconstruction Analysis

Figure 2 compares original and reconstructed log-Mel spectrograms. The reconstructed outputs preserve global spectral energy patterns while smoothing fine-grained harmonic details, indicating that the model captures the most salient audio characteristics.

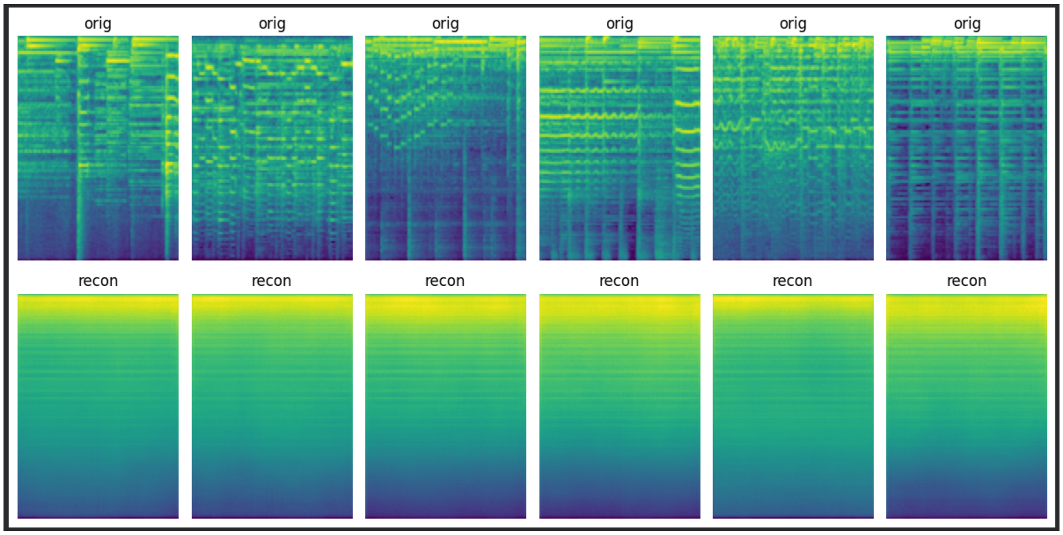

Figure 2: Original (top) and reconstructed (bottom) log-Mel spectrograms.

5.4 Cluster Composition

Figure 3 shows the cluster composition heatmap for K-Means clustering on the Beta-VAE latent space. Several clusters are dominated by specific instrument classes, while overlaps remain due to acoustic similarities between instruments.

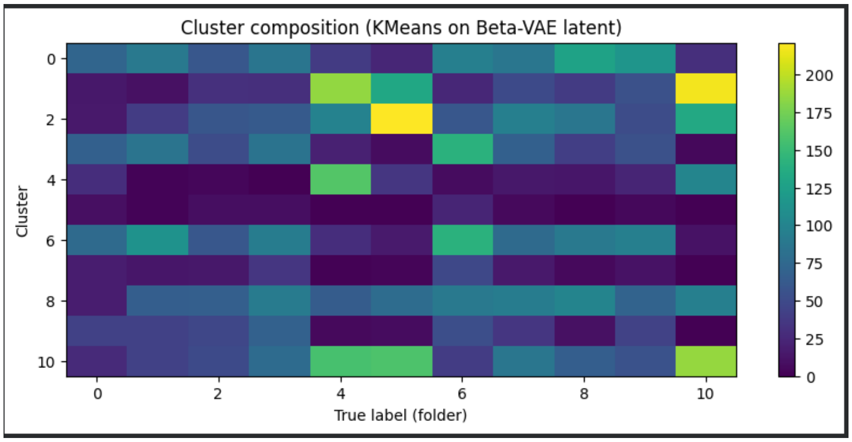

Figure 3: Cluster composition heatmap for Beta-VAE latent clustering.

6 Discussion

My results show that enforcing a structured latent prior through the Beta-VAE leads to significantly improved cluster separability. While intrinsic metrics improve substantially, ARI and NMI remain modest. This behavior is expected, as the model is trained without supervision and instrument labels do not perfectly align with acoustic similarity. Instruments from the same family often share overlapping spectral characteristics, which limits perfect label alignment.

7 Limitations

The IRMAS dataset consists of isolated instrument recordings rather than full musical compositions, which limits generalization to real-world music. Additionally, this project only considers audio features; incorporating symbolic or lyrical information could further enhance clustering performance.

8 Conclusion

In this project, I demonstrated that convolutional Beta-VAEs can learn structured and disentangled latent representations that significantly improve unsupervised music clustering. Compared to PCA and standard autoencoders, the Beta-VAE produces embeddings with much clearer cluster separation. These findings highlight the potential of variational representation learning for unsupervised music analysis.

References

[1] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *ICLR*, 2014.

[2] I. Higgins et al. beta-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

[3] IRMAS Dataset. `https://www.upf.edu/web/mtg/irmas`