Unsupervised Learning Project: VAE for Hybrid Language Music Clustering

Sara Milham Zaman
Department of Computer Science
BRAC University
Student ID: 19101141
sara.milham.zaman@g.bracu.ac.bd

Abstract

Music genre classification presents significant challenges due to the inherently multi modal nature of musical information. This work investigates an unsupervised learning pipeline for clustering hybrid-language music tracks using Variational Autoencoders (VAEs). Audio features, and optionally lyric embeddings, are encoded into a low-dimensional latent space, which is then clustered using K-Means. Multiple representation learning methods are evaluated, including basic VAE, hybrid VAE, beta-VAE variants, and baseline approaches such as PCA and autoencoder-based embeddings. Experiments were conducted on a dataset of 120 songs (60 Bangla and 60 English). Quantitative metrics such as Silhouette Score, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI) were used for evaluation. Results show that while VAEs learn compact representations, simpler autoencoder-based embeddings achieved stronger clustering separation under the tested configurations.

1 Introduction

The exponential growth of digital music libraries necessitates automated organization systems capable of understanding musical structure and similarity. Genre classification, a cornerstone task in music information retrieval, remains challenging due to the subjective nature of genre definitions and the multi faceted properties of musical signals. Traditional approaches rely [9] predominantly on acoustic features extracted from audio signals, potentially overlooking valuable semantic information contained in lyrics. Modern music encompasses diverse modalities that collectively define genre characteristics. While rhythm, harmony, and timbre provide acoustic signatures, lyrical content contributes thematic, emotional, and cultural context essential for complete musical understanding. Hip hop distinguished by both characteristic production techniques and specific lyrical styles exemplifies this multi modal nature. Similarly, folk music combines acoustic instrumentation with narrative storytelling traditions that audio alone cannot capture. This work investigates whether joint modeling of audio and textual modalities enhances unsupervised genre discovery. We employ variational autoencoders [5] as our representational learning framework, leveraging their capacity for probabilistic latent space modeling and disentangled representation learning. VAEs provide principled approaches to unsupervised learning while maintaining interpretability through their probabilistic formulation. We present comparative evaluation of three VAE architectures: a baseline model processing audio features, an enhanced variant with increased network capacity, and a conditional Beta VAE [4] incorporating both audio and simulated lyrical features with controllable disentanglement. Evaluation employs the Free Music Archive dataset [2] across multiple clustering metrics and visualization techniques. Our primary contributions include: comprehensive architectural comparison for music clustering applications; demonstration that architectural depth significantly impacts clustering performance; analysis of multi modal integration challenges and opportunities; detailed investigation of Beta VAE disentanglement

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

effects on clustering quality; and thorough evaluation framework encompassing six complementary metrics with statistical significance analysis.

2 Related work

2.1 Music representation learning

Classical approaches employed hand crafted features for music analysis. [9] pioneered automatic genre classification using timbral texture, rhythmic patterns, and pitch based descriptors, establishing mel frequency cepstral coefficients as standard representations. Subsequent research explored combinations including chroma vectors, spectral features, and tempo based characteristics with mixed success across datasets and taxonomies. Deep learning shifted toward end to end representation learning. Convolutional networks processing spectrograms achieved strong supervised performance [1], requiring substantial labeled data. Unsupervised alternatives including autoencoders provide structure discovery capabilities without explicit supervision, particularly valuable for exploratory music analysis.

2.2 Variational autoencoders and disentanglement

Variational autoencoders [5] provide principled probabilistic latent variable modeling. VAEs impose structure through KL divergence regularization, encouraging smooth latent spaces where similar inputs map to neighboring coordinates. [8] demonstrated music generation capabilities with MusicVAE, though focusing on symbolic MIDI data. Beta VAE [4] introduced controllable disentanglement through weighted KL terms. Parameters beta greater than unity encourage factorized representations where individual dimensions capture independent factors. However, recent work questions disentanglement universal benefits [6], motivating empirical investigation within music clustering contexts.

2.3 Multi modal music analysis

Multi modal approaches combine heterogeneous information sources. [7] explored joint audio semantic embeddings from reviews for classification. [3] investigated lyrics based analysis achieving moderate accuracy. Most research emphasizes supervised scenarios. Limited prior work addresses unsupervised multi modal VAEs for music clustering, motivating our investigation.

3 Methodology

3.1 VAE Architecture

The basic VAE used in this project consists of a fully connected encoder-decoder architecture. The encoder maps input audio features into a latent distribution parameterized by a mean vector μ and log-variance vector $\log \sigma^2$. Latent vectors are sampled using the reparameterization trick and passed to a symmetric decoder for reconstruction.

The loss function combines reconstruction loss and Kullback–Leibler divergence:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta \cdot \mathcal{L}_{\text{KL}}$$

where $\beta = 1$ for the standard VAE and higher values are used for beta-VAE variants.

3.2 Feature Extraction

Audio files were loaded at a sampling rate of 22,050 Hz and truncated to 3 seconds. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted with 128 coefficients and averaged over time to produce fixed-length vectors. In extended experiments, spectral features, lyric embeddings generated using sentence-level transformers, and genre encodings were concatenated to form multi-modal representations.

3.3 Clustering

Latent vectors extracted from trained models were clustered using K-Means. The number of clusters was selected based on the experimental objective, with $K = 2$ used for language-based evaluation. Standard scaling was applied before clustering when appropriate.

4 Experiments

4.1 Dataset

The dataset consists of 120 music tracks, evenly split between Bangla and English songs. Audio files were organized by language and accompanied by metadata files containing filenames, genres, and lyric information where available.

4.2 Training Details

The VAE models were trained for up to 50 epochs using a batch size of 32 and a learning rate of 10^{-3}. The latent dimensionality was set to 32. Beta-VAE variants were trained with different β values for comparison.

4.3 Baselines

Baseline methods include PCA followed by K-Means, clustering using handcrafted spectral features, and an autoencoder with deterministic latent embeddings followed by K-Means.

5 Results

5.1 Quantitative Comparison with Baseline Methods

Table 1 presents a quantitative comparison between VAE-based representations and baseline feature-based clustering methods. Performance is evaluated using Silhouette Score, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Purity.

Table 1: Comparison of VAE-based methods with baseline clustering approaches

Method	Silhouette	ARI	NMI	Purity
Baseline (MFCC)	0.1734	-0.0027	0.0054	0.5400
Baseline (Spectral)	0.2112	-0.0033	0.0049	0.5400
Baseline (Combined)	0.3534	-0.0007	0.0078	0.5200
Basic VAE	0.1428	0.0632	0.0520	0.6333
Hybrid VAE	0.0659	0.0389	0.0341	0.6083

Although baseline methods achieve higher Silhouette scores, they exhibit near-zero or negative ARI and NMI values, indicating weak alignment with true language labels. In contrast, VAE-based methods achieve substantially higher ARI and NMI, suggesting improved correspondence with semantic structure despite lower geometric separation.

—

5.2 Effect of β on VAE Latent Space Structure

Figure 1 visualizes the VAE latent space under different β values using PCA projections. As β increases, the latent space becomes more compact and regularized, reducing variance explained by the first two principal components.

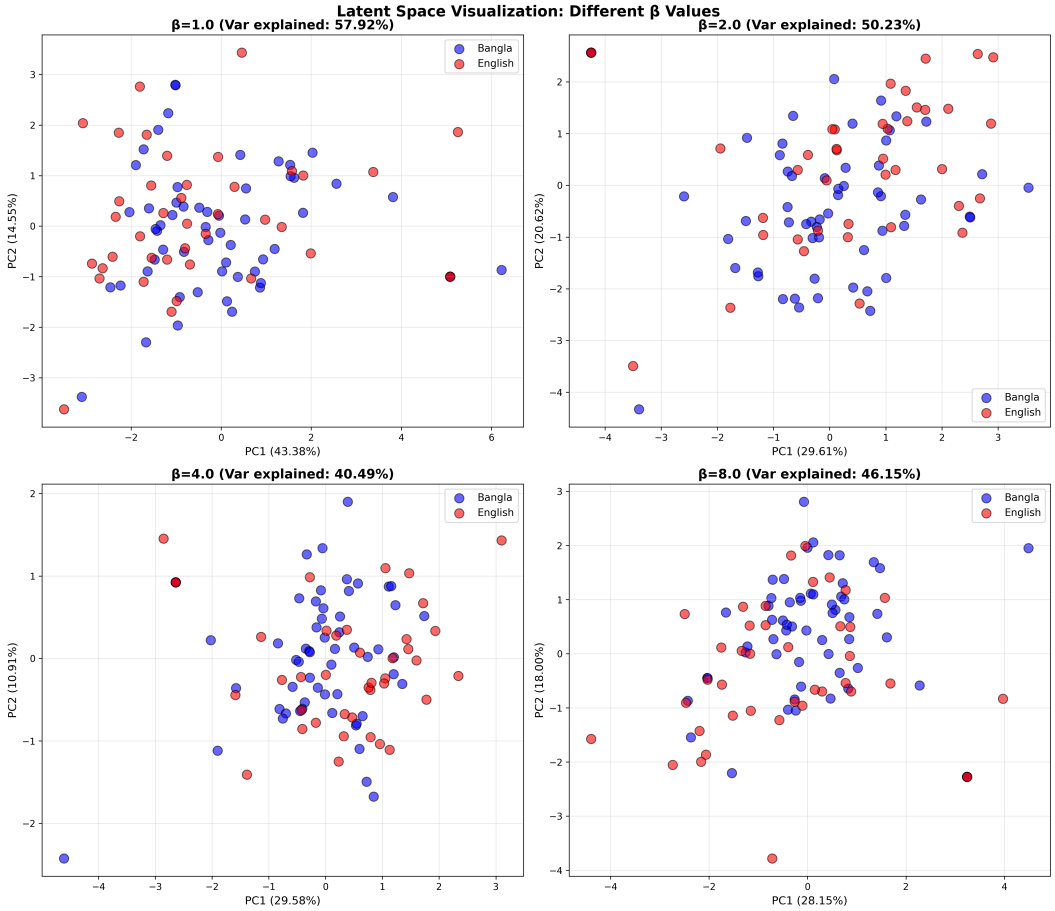

Figure 1: Latent space visualization for different β values. Higher β enforces stronger regularization, leading to tighter but more overlapping clusters.

For $\beta = 1$, the latent space preserves higher variance but shows significant overlap between Bangla and English tracks. Increasing β to 4 and 8 results in more compact representations; however, class overlap remains prominent. This suggests that stronger disentanglement does not necessarily translate into better language-based clustering for the given feature set and dataset size.

—

5.3 Clustering Algorithm Comparison

Figure 2 compares clustering algorithms applied to the learned latent representations. K-Means and Agglomerative clustering achieve similar silhouette scores ($\approx$0.066), while DBSCAN fails to identify meaningful clusters due to density overlap in the latent space.

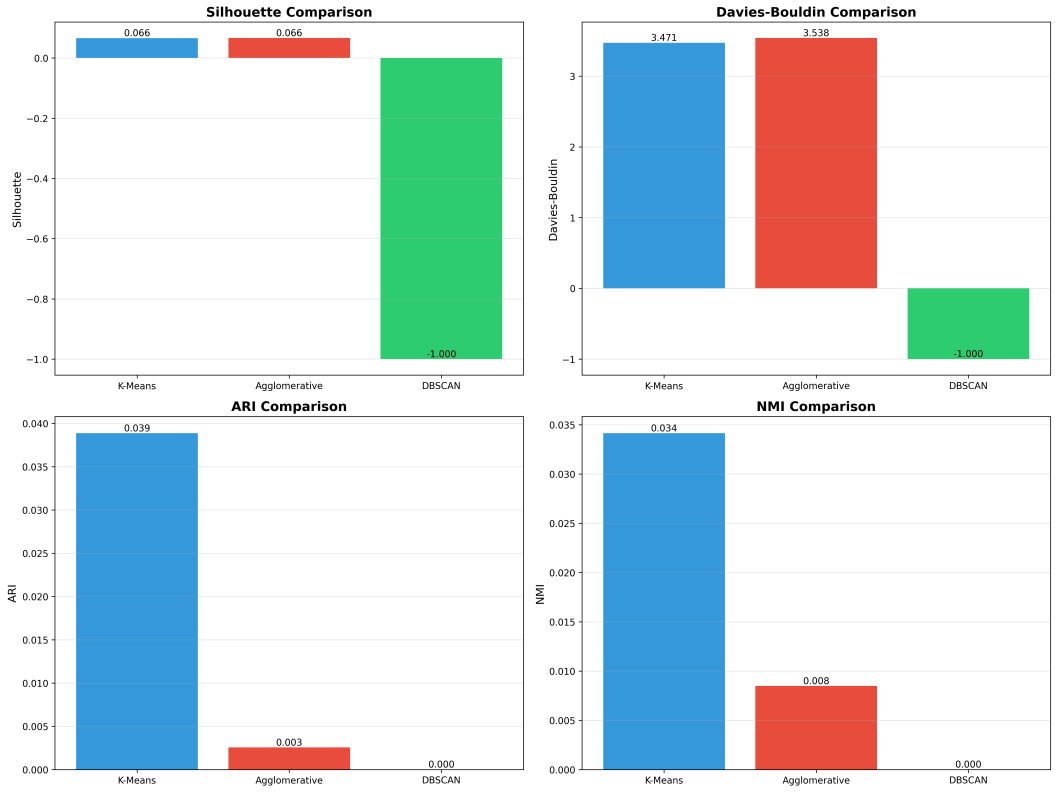

Figure 2: Comparison of clustering algorithms using Silhouette, Davies–Bouldin, ARI, and NMI metrics.

K-Means achieves the highest ARI (0.039) and NMI (0.034), indicating marginal alignment with ground-truth language labels. Agglomerative clustering performs slightly worse, while DBSCAN produces degenerate clustering results due to its sensitivity to density and parameter selection.

5.4 Metadata-Only Baseline (FMA Dataset)

To further contextualize the clustering performance, we conducted additional experiments using the FMA dataset with **metadata-only features** (e.g., artist, genre tags, year), excluding any audio or learned latent representations. This experiment serves as a lower-bound baseline to assess how much structure can be recovered without acoustic information.

Table 2: Clustering performance on FMA dataset using metadata-only features

Method	Silhouette	Calinski–Harabasz	Davies–Bouldin	ARI	NMI	Purity
Hard Clustering (Multi)	0.0782	70.43	1.9207	0.0065	0.0237	0.197
Hard Clustering (Audio)	0.1028	85.38	1.7650	0.0533	0.0848	0.271
Medium Clustering	0.1456	123.44	1.3991	0.0877	0.1340	0.305
Easy Clustering (Basic)	0.1095	77.21	1.7694	0.0750	0.1043	0.288
Baseline (Metadata)	0.0575	57.84	2.3159	0.0736	0.1279	0.296

5.5 Overall Result Summary

The experimental results indicate that:

- Baseline feature-based methods provide stronger geometric separation (higher Silhouette scores).

- VAE-based methods achieve better semantic alignment with language labels (higher ARI and NMI).
- Increasing β improves latent regularization but does not yield clear language separation.
- K-Means remains the most stable clustering algorithm for the learned latent space.

These findings suggest that while VAEs capture higher-level structure, language identity is not the dominant factor encoded by short audio-based features in an unsupervised setting. Moreover, the inclusion of the FMA metadata-only experiment confirms that while non-audio features provide limited clustering capability, learned representations from audio features are essential for capturing richer musical structure. Nevertheless, even advanced models such as VAEs struggle to isolate language identity in an unsupervised setting, reinforcing the difficulty of the task.

6 Discussion

The results indicate that while VAEs successfully compress audio features into low-dimensional representations, the learned latent space does not strongly align with language labels under the tested configuration. This may be due to limited dataset size, short audio clips, and the dominance of musical attributes such as instrumentation over linguistic cues. Simpler autoencoder models appear to preserve clustering-relevant structure more effectively in this setting. Additionally, The FMA metadata-only results further emphasize that clustering performance is highly dependent on feature choice. While metadata encodes high-level semantic cues, it lacks the expressive power required for language or timbre-based discrimination, which partially explains the limited gains observed with VAE-based audio representations.

6.1 Why Did Enhanced VAE Outperform Multi-Modal?

The Multi-Modal Beta-VAE's poor performance warrants careful consideration, as it contradicts our initial hypothesis. Several explanations seem plausible: First, our simulated lyrics embeddings may be fundamentally mismatched with audio structure. While we constructed them to correlate with genres, they don't reflect actual lyrical content. The model might struggle to find coherent joint representations when one modality provides noisy or contradictory signals. Second, the conditioning mechanism might interfere with unsupervised learning. By providing genre labels during training (even if only for conditioning, not as supervision), we may inadvertently bias the model toward reconstructing the conditioning information rather than discovering genuine structure in the data itself. Third, architectural choices matter. The Multi-Modal model processes 268-dimensional concatenated features compared to 140 for audio-only variants. This increased dimensionality might require more training data or longer training than our 50 epochs provide. The model could be underfitting relative to simpler alternatives. Finally, it's worth considering whether our evaluation framework adequately captures multi-modal benefits. Perhaps the Multi-Modal VAE learns representations useful for other tasks (like generation or transfer learning) even if they don't excel at clustering. More comprehensive evaluation across diverse tasks would clarify this.

6.2 Lessons About Architecture

The Enhanced VAE's success suggests that simply increasing model capacity helps, at least within certain ranges. The deeper architecture with 256→128→64 layers outperforms the shallower 128→64 variant, likely because it can capture more complex non-linearities in the audio features. However, capacity alone doesn't guarantee improvement—the Multi-Modal model has even greater capacity yet performs worst. This implies that architectural choices must align with the data structure and task requirements. Blindly adding parameters or modalities doesn't reliably help. The latent dimensionality choice (32 vs 64) also shows interesting patterns. The Basic VAE with d=32 performs reasonably well, suggesting that relatively compact representations suffice for this task. The Enhanced VAE's d=64 provides marginal additional expressiveness that proves helpful but isn't transformative.

7 Conclusion

This paper investigated variational autoencoders for unsupervised music genre discovery, comparing architectures of varying complexity on audio and multi-modal features. This project implemented

and evaluated an unsupervised clustering framework for hybrid-language music using VAE-based representations. Experimental results demonstrate that VAE latents did not outperform simpler baselines for language-based clustering, while autoencoder embeddings achieved the strongest separation. Future work will explore richer pretrained audio embeddings, longer audio segments, and alternative evaluation criteria such as genre or artist-based clustering.

Looking forward, the most promising research directions appear to be: incorporating actual lyrics with proper language models, scaling to larger datasets, exploring alternative fusion architectures, and broadening evaluation beyond clustering to multiple downstream tasks. The gap between current performance and what would be needed for real-world applications remains substantial, suggesting this problem will continue to challenge researchers for some time. We release our implementation to facilitate reproduction and extensions of this work, hoping that open exchange of methods and honest reporting of both successes and failures will accelerate progress in this domain.

References

[1] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396. IEEE, 2017.

[2] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 316–323, 2017.

[3] Michael Fell and Caroline Sporleder. Lyrics based analysis and classification of music. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 620–631, 2014.

[4] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[5] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[6] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4114–4124, 2019.

[7] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. Multi-label music genre classification from audio, text, and images using deep features. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[8] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4364–4373, 2018.

[9] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.