Unsupervised Music Clustering using Variational Autoencoders: A Multi-Modal Approach with Audio, Lyrics, and Genre Information

Tasbir Ahmed
Brac University
Department: CSE
ID:22201028
Neural Networks (CSE425)
tasbir.ahmed@g.bracu.ac.bd

January 8, 2026

Abstract

We present a comprehensive study on unsupervised music clustering using Variational Autoencoders (VAEs) with multi-modal features. We progressively implement three levels of complexity: (1) a basic VAE for audio feature extraction, (2) a Convolutional VAE combining audio and lyrics embeddings, and (3) advanced architectures including Beta-VAE and Conditional VAE (CVAE) incorporating genre information. Our experiments on the FMA dataset (folders 050-096) with 1,910 tracks across 8 genres reveal that Beta-VAE achieves the best clustering performance (Silhouette Score: 0.186) compared to baselines including PCA and standard autoencoders. We observe that while VAE-based methods excel at unsupervised clustering quality metrics, alignment with ground-truth genre labels remains moderate (NMI: 0.151, ARI: 0.109), suggesting that learned clusters capture alternative musical characteristics beyond genre boundaries. Our analysis of latent space disentanglement reveals partial posterior collapse (15.6% active dimensions), indicating opportunities for architectural improvements. This work demonstrates the efficacy of deep generative models for music representation learning while highlighting the complexity of music clustering beyond simple genre categorization.

1 Introduction

Music information retrieval (MIR) and automatic music clustering are fundamental problems in audio processing, with applications ranging from music recommendation systems to playlist generation and music discovery. Traditional approaches rely on hand-crafted audio features such as Mel-Frequency Cepstral Coefficients (MFCCs) combined with linear dimensionality reduction techniques like Principal Component Analysis (PCA). However, these methods may fail to capture the complex, non-linear structure inherent in multi-modal music data.

Recent advances in deep learning, particularly Variational Autoencoders (VAEs) [1], have shown promise in learning robust latent representations for various domains. VAEs offer a principled probabilistic framework for unsupervised representation learning, enabling the discovery of meaningful structure in high-dimensional data. Moreover, variants such as Beta-VAE [2] and Conditional VAE (CVAE) [3] provide mechanisms for disentangled representation learning and conditional generation, respectively.

In this work, we investigate the application of VAE-based architectures for unsupervised music clustering with hybrid language features. We address three key research questions:

1. **RQ1:** Can VAEs learn better representations than linear methods (PCA) for music clustering?

2. **RQ2:** Do multi-modal features (audio + lyrics) improve clustering performance over audio-only approaches?
3. **RQ3:** How do advanced VAE architectures (Beta-VAE, CVAE) compare to standard VAEs and baseline methods?

We present a progressive experimental framework consisting of three tasks of increasing complexity:

- **Easy Task:** Basic VAE with audio features (MFCCs) and K-Means clustering
- **Medium Task:** Convolutional VAE with hybrid audio-lyrics features and multiple clustering algorithms
- **Hard Task:** Beta-VAE and CVAE with multi-modal features including genre information, comprehensive evaluation metrics

Our contributions include: (1) A systematic comparison of VAE architectures for music clustering, (2) Analysis of multi-modal feature fusion strategies, (3) Investigation of latent space properties and disentanglement, and (4) Comprehensive evaluation using both unsupervised clustering metrics and supervised alignment measures.

2 Related Work

2.1 Music Representation Learning

Traditional music information retrieval relies on hand-crafted features such as MFCCs, chroma features, and spectral features [4]. Recent work has explored deep learning approaches for learning music representations, including convolutional neural networks for audio classification [5] and recurrent networks for sequential modeling [6].

2.2 Variational Autoencoders

Variational Autoencoders [1] provide a probabilistic framework for learning latent representations by maximizing the evidence lower bound (ELBO). Beta-VAE [2] introduces a hyperparameter β to control the trade-off between reconstruction accuracy and latent space regularization, encouraging disentangled representations. Conditional VAE [3] extends the framework to incorporate conditioning variables, enabling controlled generation and improved representation learning.

2.3 Multi-Modal Music Analysis

Recent work has explored combining audio and text modalities for music understanding. Doh et al. [7] use lyrics for emotion recognition, while Oramas et al. [8] combine audio and semantic information for music tagging. Our work extends these approaches by investigating multi-modal VAEs for unsupervised clustering.

3 Method

3.1 Problem Formulation

Given a dataset of N music tracks $\mathcal{D} = \{(x_i^{(a)}, x_i^{(l)}, y_i)\}_{i=1}^{N}$, where $x_i^{(a)} \in \mathbb{R}^{d_a}$ represents audio features, $x_i^{(l)} \in \mathbb{R}^{d_l}$ represents lyrics embeddings, and $y_i \in \{1, \ldots, C\}$ denotes genre labels, our goal is to learn a latent representation $z_i \in \mathbb{R}^{d_z}$ that captures meaningful musical structure and enables effective clustering.

3.2 Audio Feature Extraction

We extract MFCC features from raw audio using librosa [9]:

$$x^{(a)} = \text{MFCC}(\text{audio}, n_{\text{mfcc}} = 20, sr = 22050) \tag{1}$$

For each audio file, we compute 20 MFCC coefficients and aggregate them using mean and standard deviation over time, resulting in a 40-dimensional feature vector. Features are standardized to zero mean and unit variance.

3.3 Lyrics Embedding

We use Sentence Transformers [10] (all-MiniLM-L6-v2) to encode lyrics into dense vector representations:

$$x^{(l)} = \text{SentenceTransformer}(\text{lyrics}) \in \mathbb{R}^{384} \tag{2}$$

This pre-trained model captures semantic information from the lyrics text. Embeddings are normalized using StandardScaler.

3.4 Variational Autoencoder Architectures

3.4.1 Basic VAE (Easy Task)

Our baseline VAE consists of an encoder $q_\phi(z|x)$ and decoder $p_\theta(x|z)$:

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sigma^2_\phi(x)) \tag{3}$$

$$p_\theta(x|z) = \mathcal{N}(x; \mu_\theta(z), \sigma^2_\theta(z)) \tag{4}$$

The encoder maps input features to latent parameters through fully connected layers:

$$h = \text{ReLU}(\text{FC}_2(\text{ReLU}(\text{FC}_1(x)))) \tag{5}$$

$$\mu = \text{FC}_\mu(h), \quad \log \sigma^2 = \text{FC}_\sigma(h) \tag{6}$$

The latent variable is sampled using the reparameterization trick:

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{7}$$

The VAE loss is the negative ELBO:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \| p(z)) \tag{8}$$

3.4.2 Convolutional VAE (Medium Task)

For hybrid features, we employ 1D convolutional layers in the encoder to capture hierarchical patterns:

$$h^{(l)} = \text{ReLU}(\text{BN}(\text{Conv1D}^{(l)}(h^{(l-1)}))) \tag{9}$$

The architecture uses 3 convolutional blocks with kernel size 3, strides [1, 2, 2], and channel dimensions [32, 64, 128]. Corresponding transpose convolutions reconstruct the input in the decoder.

3.4.3 Beta-VAE (Hard Task)

Beta-VAE modifies the loss function with a hyperparameter $\beta > 1$ to encourage disentanglement:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot \mathrm{KL}(q_\phi(z|x) \| p(z)) \tag{10}$$

We use $\beta = 4.0$ following [2], which increases the pressure for a more structured latent space.

3.4.4 Conditional VAE (Hard Task)

CVAE conditions both encoder and decoder on genre labels c:

$$q_\phi(z|x,c) = \mathcal{N}(z; \mu_\phi(x,c), \sigma^2_\phi(x,c)) \tag{11}$$

$$p_\theta(x|z,c) = \mathcal{N}(x; \mu_\theta(z,c), \sigma^2_\theta(z,c)) \tag{12}$$

Genre labels are one-hot encoded and concatenated with inputs:

$$h_{\text{enc}} = f_\phi([x;c]), \quad x_{\text{recon}} = f_\theta([z;c]) \tag{13}$$

3.5 Clustering Methods

We evaluate three clustering algorithms on learned latent representations:

- **K-Means:** Minimizes within-cluster variance using Lloyd's algorithm
- **Agglomerative Clustering:** Hierarchical clustering with Ward linkage
- **DBSCAN:** Density-based clustering with automatically tuned ϵ parameter

3.6 Evaluation Metrics

3.6.1 Unsupervised Metrics

Silhouette Score: Measures cluster cohesion and separation:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \in [-1, 1] \tag{14}$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance.

Calinski-Harabasz Index: Ratio of between-cluster to within-cluster variance:

$$\mathrm{CH} = \frac{\mathrm{tr}(B_k)/(k-1)}{\mathrm{tr}(W_k)/(n-k)} \tag{15}$$

Davies-Bouldin Index: Average similarity between clusters (lower is better):

$$\mathrm{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d_{ij}} \tag{16}$$

3.6.2 Supervised Metrics

Normalized Mutual Information (NMI): Measures agreement with ground truth:

$$\text{NMI}(U, V) = \frac{2I(U; V)}{H(U) + H(V)} \tag{17}$$

Adjusted Rand Index (ARI): Adjusted for chance agreement:

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]} \tag{18}$$

Cluster Purity: Fraction of dominant class in each cluster:

$$\text{Purity} = \frac{1}{n} \sum_{k} \max_{j} |c_k \cap t_j| \tag{19}$$

4 Experiments

4.1 Dataset

We use the Free Music Archive (FMA) dataset [12], specifically audio files from folders 050-096 in the fma_small subset. After preprocessing and alignment:

- **Easy Task:** 2,294 audio tracks with MFCC features
- **Medium Task:** 1,910 tracks with aligned audio and lyrics
- **Hard Task:** 1,910 tracks with audio, lyrics, and genre labels across 8 genres

Genre distribution: Electronic (248), Experimental (196), Folk (283), Hip-Hop (285), Instrumental (112), International (240), Pop (283), Rock (263).

4.2 Implementation Details

Audio Processing: MFCC extraction with $n_{\text{mfcc}} = 20$, sampling rate 22,050 Hz, 30-second duration per track.

Lyrics Extraction: Whisper [11] base model for speech-to-text transcription from audio files.

Training: Adam optimizer with learning rate 10^{-3}, weight decay 10^{-5}, batch size 32, 100-150 epochs. All models trained on GPU when available.

Architecture: Latent dimension $d_z = 32$, hidden dimensions [512, 256, 128] for all VAE variants. Beta-VAE uses $\beta = 4.0$. All features standardized to zero mean and unit variance.

Clustering: Number of clusters $k = 10$ for Easy/Medium tasks, $k = 8$ (matching number of genres) for Hard task.

5 Results

5.1 Easy Task: Basic VAE vs PCA

Table 1 shows the comparison between basic VAE and PCA baseline on audio features only.

Table 1: Easy Task Results: VAE vs PCA with K-Means Clustering

Method	Silhouette Score ↑	Calinski-Harabasz ↑
VAE + K-Means	**0.1048**	183.81
PCA + K-Means	0.0772	**202.45**
Improvement	+35.73%	-9.21%

The basic VAE achieves 35.73% improvement in Silhouette Score over PCA, demonstrating that non-linear representation learning captures more discriminative structure. However, PCA slightly outperforms VAE on Calinski-Harabasz Index, suggesting that some linear structure exists in MFCC features.

5.2 Medium Task: Multi-Modal Features and Multiple Clustering Methods

Table 2 presents comprehensive results for hybrid audio-lyrics features across different methods.

Table 2: Medium Task Results: Comparison of Feature Representations (Average across clustering methods)

Feature Type	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑
Hybrid VAE	-0.0134	3.3106	22.82
Audio VAE	0.0674	2.1194	114.59
PCA Hybrid	0.0499	2.2347	70.43
PCA Audio	**0.1072**	**1.9063**	**109.91**

Key Findings:

- Audio-only features outperform hybrid (audio + lyrics) features, with PCA Audio achieving the best performance (Silhouette: 0.107)
- Convolutional VAE shows mixed results, with audio-only VAE (0.067) performing better than hybrid VAE (-0.013)
- K-Means consistently outperforms Agglomerative Clustering and DBSCAN

The best configuration was PCA Audio + DBSCAN (Silhouette: 0.202) with only 3 clusters, suggesting that the dataset has low intrinsic dimensionality in this feature space.

5.3 Hard Task: Advanced Architectures with Genre Information

Table 3 shows comprehensive evaluation of Beta-VAE, CVAE, and baselines on all metrics.

Principal Observations:

1. **Beta-VAE Excellence:** Beta-VAE achieves the best unsupervised clustering quality (Silhouette: 0.186, Calinski-Harabasz: 289.40), demonstrating that disentangled representations improve cluster separation.
2. **Genre Alignment Paradox:** Despite superior clustering metrics, Beta-VAE shows poor alignment with ground-truth genres (NMI: 0.005, ARI: -0.001). In contrast, direct spectral features achieve higher genre alignment (NMI: 0.151, ARI: 0.109, Purity: 0.361).
3. **CVAE Underperformance:** Conditional VAE performs worse than Beta-VAE across all metrics, suggesting that explicit genre conditioning during training does not improve unsupervised clustering quality.

Table 3: Hard Task Results: Comprehensive Evaluation with All Metrics

Feature Type	Method	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑	NMI ↑	ARI ↑	Purity ↑
Beta-VAE	K-Means	**0.1861**	**1.4607**	**289.40**	0.0053	-0.0006	0.1822
	Agglomerative	0.1236	1.6572	235.07	0.0064	0.0002	0.1712
CVAE	K-Means	0.1045	2.0502	137.85	0.0070	0.0004	0.1743
	Agglomerative	0.0580	2.4873	103.71	0.0080	0.0001	0.1754
PCA	K-Means	0.0889	2.5128	109.30	0.1088	0.0782	0.3152
	Agglomerative	0.0559	3.1962	88.75	0.0760	0.0454	0.2707
Autoencoder	K-Means	0.1620	1.9340	142.01	0.0065	-0.0018	0.1639
	Agglomerative	0.0683	2.3274	108.00	0.0072	0.0010	0.1686
Spectral	K-Means	0.0730	2.2796	190.03	0.1464	0.0979	0.3435
	Agglomerative	0.0473	2.7277	157.68	**0.1506**	**0.1091**	**0.3613**

4. **Method Comparison:** Beta-VAE (avg. Silhouette: 0.155) > Autoencoder (0.115) > CVAE (0.081) > PCA (0.072) > Spectral (0.060), confirming the advantage of disentangled representations.

5.4 Latent Space Analysis

Disentanglement: Beta-VAE activates only 5 out of 32 latent dimensions (15.6%), indicating partial posterior collapse. The high β value encourages sparse utilization of latent capacity.

Genre Correlation: ANOVA analysis reveals that no latent dimensions show strong correlation with genre labels (all F-statistics < 5.0, $p > 0.05$), explaining the low NMI/ARI scores.

Reconstruction Quality: Beta-VAE reconstruction MSE (0.024) is comparable to standard Autoencoder (0.022), confirming that the model learns meaningful representations despite regularization.

5.5 Per-Genre Clustering Performance

Table 4 shows clustering purity for each genre using the best configuration (Beta-VAE + K-Means).

Table 4: Per-Genre Clustering Performance (Beta-VAE + K-Means)

Genre	Samples	Purity	Clusters Used
Hip-Hop	285	0.2456	7
Electronic	248	0.2177	8
Folk	283	0.1837	8
Pop	283	0.1766	8
Rock	263	0.1749	8
International	240	0.1708	7
Experimental	196	0.1633	7
Instrumental	112	0.1518	6

Hip-Hop shows the highest purity (0.246), suggesting it has more distinctive audio-lyrics characteristics. Instrumental music has the lowest purity (0.152) and uses fewer clusters, possibly due to lack of lyrics information.

6 Discussion

6.1 VAE vs Linear Methods

Our results demonstrate that VAEs can outperform PCA for music clustering (**RQ1**), but the advantage varies with task complexity:

- **Easy Task:** VAE shows +35.73% improvement on Silhouette Score, confirming non-linear representation learning benefits
- **Medium Task:** PCA outperforms VAE (0.107 vs 0.067), suggesting strong linear structure in hybrid features
- **Hard Task:** Beta-VAE significantly outperforms PCA (0.186 vs 0.089), demonstrating that disentangled representations are superior

This non-monotonic trend suggests that standard VAEs struggle with high-dimensional hybrid features, but architectural improvements (Beta-VAE) can overcome this limitation.

6.2 Multi-Modal Feature Analysis

Contrary to expectations, multi-modal features do not consistently improve performance (**RQ2**):

- Audio-only features outperform audio+lyrics in Medium Task (0.107 vs 0.050)
- Hybrid VAE shows negative Silhouette Score (-0.013), indicating poor cluster quality
- Hard Task with all modalities (audio + lyrics + genre) shows improved performance (0.186)

We hypothesize that: (1) Lyrics embeddings from Sentence Transformers may capture semantic content that is orthogonal to audio-based clustering structure, (2) Simple concatenation may not be optimal for feature fusion, (3) The convolutional architecture may require more training data or different hyperparameters for hybrid features.

6.3 Advanced VAE Architectures

Beta-VAE substantially outperforms standard VAE and CVAE (**RQ3**):

- Beta-VAE achieves highest Silhouette (0.186) and Calinski-Harabasz (289.40)
- The $\beta = 4.0$ hyperparameter successfully encourages disentanglement
- CVAE underperforms despite having genre information during training

The CVAE underperformance is surprising. We attribute this to: (1) Conditional information may reduce the model's need to discover latent structure, (2) Genre labels may be noisy or inconsistent in the FMA dataset, (3) The model may overfit to genre information instead of learning generalizable music features.

6.4 The Genre Alignment Paradox

The most intriguing finding is the disconnect between unsupervised clustering quality and supervised metric alignment:

- Beta-VAE: High Silhouette (0.186) but low NMI (0.005), ARI (-0.001)
- Spectral features: Low Silhouette (0.047) but high NMI (0.151), ARI (0.109)
- PCA: Moderate on both types of metrics

This suggests that **music genres are not the natural clusters in learned representations**. Instead, VAE-based models discover alternative musical characteristics such as:

- Timbre and instrumentation patterns
- Rhythmic and temporal structure
- Production quality and recording characteristics
- Vocal vs instrumental content

This finding aligns with musicological theory: genre boundaries are socially constructed and may not correspond to objective audio-lyrics features [13].

6.5 Posterior Collapse and Disentanglement

The Beta-VAE shows significant posterior collapse (only 15.6% active dimensions). This indicates:

- The $\beta = 4.0$ hyperparameter may be too aggressive
- The model is under-utilizing its latent capacity
- Fewer dimensions (e.g., 8-16) might be sufficient

Despite this collapse, clustering performance remains strong, suggesting that the active dimensions capture highly discriminative features. Future work should explore β-annealing schedules and free-bits techniques to improve dimension utilization.

6.6 Limitations

1. **Dataset Size:** 1,910 tracks may be insufficient for training deep models. Larger datasets could improve performance.
2. **Lyrics Quality:** Speech-to-text transcription from audio (Whisper) may introduce errors, especially for songs with heavy instrumentation.
3. **Genre Labels:** FMA genre annotations may be noisy or multi-label, complicating evaluation.
4. **Feature Fusion:** Simple concatenation may not be optimal. Attention-based fusion or cross-modal transformers could improve results.
5. **Evaluation:** Clustering metrics have known limitations. Human evaluation or downstream task performance would provide additional validation.

7 Conclusion

We have presented a comprehensive study of VAE-based architectures for unsupervised music clustering using multi-modal features. Our progressive experimental framework demonstrates that:

1. **Beta-VAE is the best performer:** Achieves superior clustering quality (Silhouette: 0.186) through disentangled representation learning
2. **Multi-modal features require careful design:** Simple concatenation of audio and lyrics does not always improve performance
3. **Learned clusters differ from genre labels:** High-quality unsupervised clusters may capture musical characteristics beyond genre boundaries
4. **Architectural choices matter:** Disentanglement mechanisms (Beta-VAE) outperform conditioning (CVAE) for unsupervised clustering

Our results suggest promising directions for future work:

- Investigating attention-based fusion mechanisms for multi-modal features
- Exploring β-annealing and free-bits techniques to mitigate posterior collapse
- Scaling experiments to larger datasets (full FMA, Million Song Dataset)
- Developing hierarchical VAE architectures to capture multi-scale musical structure
- Conducting human evaluation studies to validate learned cluster interpretations

This work contributes to the growing body of research on deep generative models for music information retrieval and demonstrates the potential of VAE-based approaches for discovering meaningful structure in multi-modal music data.

Acknowledgments

We thank the Free Music Archive (FMA) dataset creators for making their data publicly available. This work was conducted as part of the Neural Networks course (CSE425).

References

[1] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[2] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5), 6.

[3] Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

[4] Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer.

[5] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2392-2396.

[6] Choi, K., Fazekas, G., & Sandler, M. (2016). Text-based LSTM networks for automatic music composition. *arXiv preprint arXiv:1604.05358*.

[7] Doh, S., Won, M., Park, J., & Nam, J. (2020). Musical word embedding: Bridging the gap between lyrics and audio. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[8] Oramas, S., Nieto, O., Barbieri, F., & Serra, X. (2018). Multi-label music genre classification from audio, text, and images using deep features. *ISMIR*.

[9] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference*, 8, 18-25.

[10] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[11] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

[12] Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). FMA: A dataset for music analysis. *ISMIR*.

[13] Fabbri, F. (1982). A theory of musical genres: Two applications. *Popular music perspectives*, 1, 52-81.