Unsupervised Music Clustering with Variational Autoencoders and Multimodal Fusion

Mushfiqur Rahman Khan
CSE / Brac University
mushfiqur.rahman.khan2@g.bracu.ac.bd

Abstract

Unsupervised music clustering was investigated using Variational Autoencoders (VAEs) to learn compact representations of music tracks. I have focused on multimodal learning by combining audio features (e.g. MFCCs and mel spectrograms) with lyric embeddings derived from automatic transcription using Whisper [Radford et al., 2022]. I have compared multiple VAEs architectures (MLP-based, convolutional, and conditional VAEs), clustering approaches (K-Means and Gaussian Mixture Models) and evaluated using intrinsic metrics (Silhouette, Davies–Bouldin, Calinski–Harabasz) and extrinsic metrics (ARI, NMI, Purity) when labels are available. Experiments on the GTZAN dataset [Tzanetakis and Cook, 2002] show that learned embeddings significantly improve clustering quality over raw feature baselines, with the MLP-VAE often achieving the best overall performance.

1 Introduction

Music clustering is a fundamental task in music information retrieval, enabling organization, playlisting, and discovery without requiring labels. Traditional pipelines rely on hand-crafted features, while modern approaches learn representations directly from data.

1.1 Motivation

Unsupervised clustering is valuable when labels are missing, noisy, or expensive. Deep latent-variable models like VAEs [Kingma and Welling, 2014] can learn compact representations that make clustering easier.

1.2 Problem Statement

Given a collection of music tracks, our goal is to:

1. Learn meaningful latent representations from audio and lyric modalities.
2. Cluster tracks in latent space using unsupervised clustering methods.
3. Evaluate cluster quality with intrinsic and (when available) label-based metrics.
4. Analyze the role of model architecture and multimodal fusion on clustering performance.

1.3 Contributions

Our key contributions include:

- An end-to-end pipeline for audio + lyrics unsupervised clustering using VAEs.
- Comparison of MLP-VAE, Conv-VAE, and Conditional VAE variants.

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

- Extensive evaluation across clustering methods and metrics.
- Analysis of multimodal fusion challenges and representation trade-offs.

2 Related Work

Traditional approaches use hand-crafted features such as:

- **MFCC**: compact timbral descriptors.
- **Mel spectrograms**: time–frequency representations of audio.
- **Chroma**: pitch class profiles emphasizing harmony.

Deep representation learning for audio includes end-to-end models (e.g., CNN-based learning for music audio) [Dieleman and Schrauwen, 2014] and discrete latent approaches like VQ-VAE [van den Oord et al., 2017]. For clustering, Deep Embedded Clustering (DEC) [Xie et al., 2016] demonstrates that learned embeddings can improve cluster structure.

3 Method

3.1 Dataset

I have used the GTZAN dataset [Tzanetakis and Cook, 2002], converting tracks into fixed 10-second clips. I used the original genre labels only for evaluation, not for training.

3.2 Feature Extraction

Audio: I have computed mel spectrograms and MFCCs using `librosa` [McFee et al., 2015]. **Lyrics:** I have generated transcripts using Whisper ASR [Radford et al., 2022] and embedded lyrics into a fixed vector (e.g., TF-IDF or sentence embeddings, depending on implementation level).

3.3 VAE Architectures

All VAEs are trained to reconstruct inputs while regularizing the latent space with a KL divergence term:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x) \,\|\, p(z)). \tag{1}$$

MLP-VAE: Fully connected encoder/decoder for vector features.
Conv-VAE: Convolutional encoder/decoder for spectrogram inputs.
Conditional VAE (CVAE): Conditioned on modality indicators or fused features to encourage structured latent learning.

3.4 Clustering

I have applied K-Means and Gaussian Mixture Models in latent space. For baselines, I have also clustered directly in the raw feature space.

3.5 Evaluation Metrics

Intrinsic: Silhouette [Rousseeuw, 1987], Davies–Bouldin, and Calinski–Harabasz.
Extrinsic: Adjusted Rand Index (ARI) [Hubert and Arabie, 1985], Normalized Mutual Information (NMI), and Purity.

4 Experiments

4.1 Training Configurations

I have compared three implementation levels (Easy/Medium/Hard) and multiple model variants.

Table 1: Training configuration for representative model variants.

Model	Level	Input Type	Latent Dim	Epochs	Batch Size
MLP-VAE (Audio)	Easy	MFCC (40-d)	16	50	64
MLP-VAE (Audio+Lyrics)	Medium	MFCC + Text Emb	32	80	64
Conv-VAE (MelSpec)	Medium	$1\times128\times431$	32	80	32
CVAE (Fusion)	Hard	Audio + Lyrics	64	100	32

5 Results

5.1 Quantitative Results

Table 2: Clustering performance across representative models and baselines. Higher is better for Silhouette/CH/ARI/NMI/Purity; lower is better for Davies–Bouldin (DB).

Method	Silhouette↑	DB↓	CH↑	ARI↑	NMI↑	Purity↑	K	Clustering	Input
Raw MFCC + KMeans	0.12	2.31	145.0	0.08	0.10	0.21	10	KMeans	MFCC
Raw MelSpec + KMeans	0.10	2.55	120.7	0.05	0.09	0.18	10	KMeans	MelSpec
MLP-VAE (Audio) + KMeans	**0.24**	**1.71**	**310.4**	**0.19**	**0.23**	**0.34**	10	KMeans	Latent
Conv-VAE + KMeans	0.20	1.95	265.2	0.15	0.19	0.30	10	KMeans	Latent
CVAE (Fusion) + GMM	0.22	1.83	290.0	0.17	0.22	0.33	10	GMM	Latent

5.2 Discussion

Overall, learned VAE embeddings improve clustering over raw features. MLP-VAEs are strong on MFCC-based vectors; Conv-VAEs can leverage richer time–frequency structure but may need careful tuning. Multimodal fusion can help but introduces alignment and noise issues (ASR errors, lyric variability).

6 Conclusion

I have presented a NeurIPS-style study of unsupervised music clustering using VAE representations and multimodal fusion. Across multiple architectures and clustering strategies, latent embeddings provide better cluster structure than raw feature baselines. Future work includes stronger text encoders, contrastive multimodal pretraining, and improved fusion architectures.

References

Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968, 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. URL `https://arxiv.org/abs/1312.6114`.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. URL https://arxiv.org/abs/2212.04356.

Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1711.00937.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. *arXiv preprint arXiv:1511.06335*, 2016. URL https://arxiv.org/abs/1511.06335.