
Unsupervised Clustering of Hybrid-Language Music via Variational Autoencoders and Audio–Lyrics Fusion

MD Tanzim Qaiyum

md.tanzim.qaiyum@g.bracu.ac.bd

Abstract

We study unsupervised clustering of hybrid-language music by learning compact representations from audio and lyrics and applying standard clustering algorithms in the learned space. We train (i) an MLP variational autoencoder (VAE) on MFCC summary statistics for a combined BanglaBeats+GTZAN collection, and (ii) a convolutional VAE (ConvVAE) on fixed-size log-mel spectrograms from JamendoLyrics. We further evaluate a β -VAE variant and a simple fusion strategy that concatenates audio latents with TF-IDF lyric embeddings. Clustering is assessed using internal metrics (Silhouette, Calinski–Harabasz, Davies–Bouldin) and, when labels are available, external agreement (NMI, ARI, Purity) against language and genre. Empirically, nonlinear VAE latents improve clustering over PCA on MFCC features (Silhouette 0.237 vs. 0.183; Calinski–Harabasz 2927 vs. 2403), while ConvVAE audio latents yield strong geometric separation on JamendoLyrics (Silhouette up to 0.558 under DBSCAN with moderate noise). Fusion improves some label-alignment metrics but remains limited under simple concatenation, suggesting the need for more principled multi-modal alignment objectives.

1 Introduction

Music similarity is shaped by multiple interacting factors: timbre, rhythm, instrumentation, production characteristics, vocal delivery, and lyrical semantics. In hybrid-language settings, language identity can correlate with some attributes (e.g., phonetics, recording conventions), but musical style often overlaps across languages. This motivates unsupervised representation learning that can discover structure without relying on curated labels. We investigate whether VAE-style latent spaces are more cluster-friendly than linear baselines, and whether adding lyrics yields representations that better align with language and genre categories.

2 Related Work

Variational Autoencoders. VAEs learn probabilistic latent representations by maximizing an evidence lower bound with a reconstruction term and a KL regularizer (1).

Disentanglement and β -VAE. β -VAE increases the weight of the KL term to encourage factorized latents and disentanglement (2).

Deep clustering. Deep Embedded Clustering (DEC) jointly refines representations and cluster assignments by optimizing a clustering objective in latent space (3). Related iterative clustering-based representation learning approaches include DeepCluster (4).

Manifold visualization. t-SNE is widely used for visual inspection of embedding structure (5). UMAP is a popular alternative with different trade-offs for scalability and global structure (6).

Music datasets and benchmarks. GTZAN is a classic genre benchmark introduced in early work on automatic genre classification (7). JamendoLyrics provides multilingual songs with lyrics and alignments that enable joint audio–text analysis (9).

3 Datasets

BanglaBeats + GTZAN (MFCC experiments). We combine BanglaBeats (Bengali songs) (8) with GTZAN (7). For this study, we construct a balanced subset of 10,000 samples (5,000 Bangla and 5,000 English/Western proxy) and represent each track with MFCC summary statistics.

JamendoLyrics (spectrogram and lyrics experiments). We use JamendoLyrics for experiments that incorporate log-mel spectrograms and lyric text (9). We evaluate on two extracted subsets: (i) $N = 2924$ usable segments for clustering sweeps, and (ii) $N = 945$ samples with language and genre labels for external evaluation.

4 Method

4.1 Audio and lyrics representations

MFCC summary features. Each track is represented as an 80-D vector: MFCC means (40) concatenated with MFCC standard deviations (40). Features are standardized with z-score normalization.

Log-mel spectrograms. We compute log-mel spectrograms with $N_{mels} = 128$, $n_{fft} = 2048$, hop length = 512, and fix the time dimension to $T = 256$ via padding/truncation, yielding tensors of shape (1, 128, 256).

Lyrics embeddings. Lyrics are embedded with TF-IDF (max features 5000; 1–2 grams) and reduced via PCA to 64 dimensions. Embeddings are standardized before fusion.

4.2 VAE and ConvVAE

Let x denote an input feature (MFCC vector or spectrogram tensor). The encoder outputs parameters $(\mu_\phi(x), \log \sigma_\phi^2(x))$ defining a diagonal Gaussian posterior $q_\phi(z | x) = \mathcal{N}(\mu_\phi(x), \text{diag}(\sigma_\phi^2(x)))$. Sampling uses the reparameterization trick $z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$. The decoder reconstructs \hat{x} . We optimize the standard VAE objective:

$$\mathcal{L}(x) = \underbrace{\|x - \hat{x}\|_2^2}_{\text{reconstruction}} + \underbrace{D_{KL}(q_\phi(z|x) \parallel \mathcal{N}(0, I))}_{\text{regularization}}. \quad (1)$$

MLP-VAE (MFCC). We use an MLP encoder/decoder with hidden sizes (256, 128) and latent dimension 16. Training uses Adam (LR 10^{-3}), batch size 64, 50 epochs.

ConvVAE (spectrogram). We use strided convolutions to encode (1, 128, 256) to a 32-D latent and decode back to the input shape. Training uses Adam (LR 10^{-3}), batch size 32, 25 epochs.

β -VAE variant. We also evaluate a β -VAE-style loss:

$$\mathcal{L}_\beta(x) = \|x - \hat{x}\|_2^2 + \beta D_{KL}(q_\phi(z|x) \parallel \mathcal{N}(0, I)), \quad (2)$$

with $\beta = 2.0$ (latent dimension 32; 20 epochs).

4.3 Multi-modal fusion

We fuse audio latents $z_a \in \mathbb{R}^{32}$ with lyric embeddings $e_\ell \in \mathbb{R}^{64}$ by concatenation:

$$z_{\text{fused}} = [z_a, \alpha \cdot e_\ell] \in \mathbb{R}^{96}, \quad (3)$$

where α is a scaling hyperparameter (swept in the JamendoLyrics experiments; $\alpha = 2.0$ in the β -VAE fused evaluation).

4.4 Clustering and evaluation

We cluster embeddings using K-Means, agglomerative clustering (Ward/Euclidean), and DBSCAN. We report:

- **Internal metrics:** Silhouette (higher better), Calinski–Harabasz (higher), Davies–Bouldin (lower).
- **External agreement (when labels exist):** NMI, ARI, and Purity w.r.t. language or genre labels.

5 Experiments and Results

5.1 MFCC: MLP-VAE vs. PCA baseline (BanglaBeats+GTZAN)

We compare clustering in learned latent space versus PCA-reduced space under K-Means with $k = 2$. Table 1 shows that VAE latents improve both Silhouette and Calinski–Harabasz.

Table 1: MFCC clustering on BanglaBeats+GTZAN ($k = 2$): VAE latents vs. PCA baseline.

Method	Silhouette \uparrow	Calinski–Harabasz \uparrow
VAE latent (16D) + K-Means	0.2367	2927.37
PCA (16D) + K-Means	0.1833	2402.93



Figure 1: t-SNE visualization of MFCC VAE latents. Left: colored by K-Means cluster. Right: colored by dataset/language source.

5.2 Spectrogram: ConvVAE embeddings and lyrics fusion (JamendoLyrics, $k = 4$)

We evaluate audio-only ConvVAE latents against fused audio+lyrics embeddings. Table 2 reports representative full-coverage settings and a DBSCAN configuration that trades off coverage for compactness.

Table 2: JamendoLyrics clustering (N=2924 for full coverage): audio-only vs. fused audio+lyrics.

Representation	Algorithm	Params	Sil \uparrow	DB \downarrow	Noise
Audio latent (32D)	K-Means	$k = 4$	0.4562	0.9365	0.00
Audio latent (32D)	Agglo	$k = 4$	0.4393	0.8935	0.00
Audio latent (32D)	DBSCAN	eps=1.5, ms=20	0.5582	0.5387	0.14
Fused ($\alpha = 0.25$)	K-Means	$k = 4$	0.3908	1.0319	0.00
Fused ($\alpha = 0.25$)	Agglo	$k = 4$	0.3712	0.9925	0.00

5.3 β -VAE + fusion: evaluation against language and genre (JamendoLyrics, N=945)

We evaluate fused embeddings from a β -VAE audio encoder ($\beta = 2.0$) combined with lyric embeddings ($\alpha = 2.0$). We report both internal metrics and agreement with language ($k = 4$) and genre ($k = 12$).

Table 3: Fused embedding vs. **language** labels ($k = 4$, $N=945$).

Algorithm	Params	Sil \uparrow	DB \downarrow	NMI \uparrow	ARI \uparrow	Purity \uparrow
K-Means	$k = 4$	0.0299	3.8111	0.0863	0.0108	0.3429
Agglo	$k = 4$	0.1216	2.2800	0.0231	0.0002	0.2730
DBSCAN	eps=1.2, ms=10	–	–	0.0000	0.0000	0.2540

Table 4: Fused embedding vs. **genre** labels ($k = 12$, $N=945$).

Algorithm	Params	Sil \uparrow	DB \downarrow	NMI \uparrow	ARI \uparrow	Purity \uparrow
K-Means	$k = 12$	0.0287	3.0691	0.0782	0.0070	0.2931
Agglo	$k = 12$	0.1072	1.0069	0.0963	0.0022	0.3185
DBSCAN	eps=1.2, ms=10	–	–	0.0000	0.0000	0.2794

6 Discussion

Why nonlinear latents help on MFCC. Compared to PCA, the MLP-VAE learns a nonlinear mapping that can reshape the MFCC feature space into a geometry that favors cluster separation, reflected by improved Silhouette and Calinski–Harabasz.

Spectrogram ConvVAE yields highly clusterable audio embeddings. On JamendoLyrics, audio-only ConvVAE latents are consistently strong across K-Means and agglomerative clustering, and DBSCAN can further increase Silhouette when allowing a moderate noise fraction.

Fusion remains challenging under simple concatenation. Concatenating TF-IDF lyric embeddings can dilute distance structure from strong audio latents unless scaling is carefully tuned; even then, external alignment metrics (NMI/ARI/Purity) remain modest. This suggests exploring learned fusion (e.g., modality-specific projections), contrastive alignment, or deep clustering objectives (3; 4).

7 Limitations

Our experiments are limited by (i) reliance on TF-IDF as the lyric representation (which may not capture cross-lingual semantics well), (ii) a simple concatenation-based fusion rather than learned multi-modal alignment, and (iii) evaluation on fixed-size subsets, where results can vary with segment selection and clustering hyperparameters (especially DBSCAN). Future work should incorporate stronger text encoders, learned fusion, and repeated-run reporting with variance estimates.

8 Conclusion

We presented VAE-based unsupervised clustering pipelines for hybrid-language music using MFCC summaries and log-mel spectrograms, and studied audio–lyrics fusion. VAEs improved clustering over PCA on MFCC features, while ConvVAE embeddings on JamendoLyrics produced strong geometric clustering. Simple fusion provides limited gains in label alignment, motivating more principled multi-modal representation learning.

References

References

- [1] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013. :contentReference[oaicite:0]index=0
- [2] 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. In *ICLR*, 2017. :contentReference[oaicite:1]index=1
- [3] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016. :contentReference[oaicite:2]index=2

- [4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. :contentReference[oaicite:3]index=3
- [5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 2008.
- [6] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018. :contentReference[oaicite:4]index=4
- [7] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 2002. :contentReference[oaicite:5]index=5
- [8] Md. Mehedi Hasan Jibon, Dewan Mahinur Alam, and Mohammad Shahidur Rahman. BanglaBeats: A comprehensive dataset of Bengali songs for music genre classification tasks. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 2023. :contentReference[oaicite:6]index=6
- [9] Simon Durand, Daniel Stoller, and Sebastian Ewert. Contrastive learning-based audio to lyrics alignment for multiple languages (JamendoLyrics Multi-Lang). *ICASSP 2023 / arXiv:2306.07744*, 2023. :contentReference[oaicite:7]index=7