Disentangling Language and Genre: Unsupervised Cross-Cultural Music Clustering via Hybrid Beta-VAE

Rumman Ahmed Prodhan
Department of Computer Science & Engineering
BRAC University
Dhaka, Bangladesh
rumman.ahmed.prodhan@g.bracu.ac.bd

Abstract

This paper presents an unsupervised learning pipeline for clustering hybrid language music tracks utilizing a multi-modal Variational Autoencoder (VAE). Traditional audio-based clustering methods often struggle with cross-cultural genre overlaps. To address this, we propose a Hybrid Beta-VAE architecture that fuses convolutional spectral features with semantic lyric embeddings. We evaluated the architecture on HBLM-100, a custom-curated dataset of 100 Bangla and English songs, and the standard GTZAN benchmark. Experimental results demonstrate that while audio-only baselines fail (ARI ≈ 0.0), our hybrid approach achieves perfect separation (ARI=1.0, NMI=1.0). Furthermore, benchmarking on GTZAN yielded a Cluster Purity of 0.41, significantly outperforming PCA baselines. This study highlights the necessity of multi-modal fusion for robust music information retrieval.

1 Introduction

The rapid digitization of global music libraries has necessitated the development of robust Music Information Retrieval (MIR) systems (1) capable of navigating diverse, multilingual collections. Traditional unsupervised clustering methods typically rely on a single modality, utilizing either acoustic signal processing features—such as Mel-Frequency Cepstral Coefficients (MFCCs) or textual metadata (2). While these unimodal approaches are effective for standard genre classification, they face significant limitations in hybrid language scenarios. For instance, a "Bangla Rock" song often shares identical spectral characteristics with an "English Rock" song, causing audio-only models to conflate them despite their distinct linguistic contexts.

To address this challenge, we introduce a multi-modal generative framework that fuses acoustic and semantic representations. We propose a Hybrid Beta-Variational Autoencoder (Beta-VAE) (3) designed to learn disentangled latent representations. By strictly weighting the Kullback-Leibler divergence term ($\beta > 1$), we force the model to separate independent factors of variation—specifically "Language" (derived from lyrics) and "Genre" (derived from audio) into distinct orthogonal axes within the latent space.

This work fulfills the requirements of a high-complexity unsupervised learning task by integrating Convolutional Neural Networks (CNNs) for spectral feature extraction with dense semantic embeddings from pre-trained Transformer models (4).

1.1 Our Contributions

In this paper, we present the following contributions:

Preprint. Under review.

- *Hybrid Beta-VAE Architecture:* We propose a novel multi-modal architecture that fuses Convolutional audio encodings with Sentence-BERT lyric embeddings. We demonstrate that a disentanglement factor of $\beta = 4.0$ is critical for separating semantic content from musical texture.
- *HBLM-100 Dataset:* We introduce a custom-curated Hybrid Bangla Language Music dataset comprising 100 tracks perfectly balanced between Bangla and English. This dataset serves as a rigorous benchmark for cross-cultural disentanglement tasks.
- *Comprehensive Evaluation:* We benchmark our model against standard baselines (PCA and Audio-Only VAE) on both the custom HBLM-100 dataset and the public GTZAN genre collection.
- *State-of-the-Art Unsupervised Performance:* We report superior clustering performance using advanced metrics including Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Cluster Purity. Our method achieves perfect separation (ARI=1.0) on the bilingual task, significantly outperforming unimodal baselines.

2 Related Work

The clustering of musical content has traditionally been approached as a unimodal problem, relying either on acoustic signal processing or textual metadata. While significant progress has been made in both domains individually, the integration of these modalities for unsupervised cross-cultural analysis remains an active area of research.

2.1 Audio-Based Music Classification

Early approaches to Music Information Retrieval (MIR) focused predominantly on hand-crafted acoustic features. Tzanetakis and Cook (2) established the foundational methodology for genre classification by utilizing Mel-Frequency Cepstral Coefficients (MFCCs). While effective for broad categories, these shallow features lack the temporal granularity required for complex linguistic differentiation.

Recent advancements have shifted towards deep learning architectures. Choi et al. (5) demonstrated that Convolutional Recurrent Neural Networks (CRNNs) operating on log-mel spectrograms significantly outperform hand-crafted baselines by learning hierarchical time-frequency features. Similarly, Pons et al. (6) and Hershey et al. (7) proposed end-to-end learning strategies using raw waveforms and CNN architectures. However, as noted in recent literature, these audio-only architectures often suffer from "semantic confusion" when distinct cultural categories share similar instrumentation, a limitation our work addresses by incorporating a secondary semantic modality.

2.2 Multi-Modal Fusion in MIR

To mitigate the limitations of audio-only systems, researchers have explored multi-modal architectures. Oramas et al. (8) proposed a multi-modal deep learning framework that fuses audio, text, and images for genre classification, demonstrating that intermediate fusion outperforms late fusion strategies. While their approach was discriminative (supervised classification), our work extends this fusion concept to the generative (unsupervised) domain. By employing a VAE, we do not merely concatenate features but enforce a joint probability distribution $P(x_{audio}, x_{lyrics}|z)$, allowing the model to capture non-linear dependencies between musical texture and lyrical content.

2.3 Disentangled Representation Learning

The application of Variational Autoencoders (VAEs) (9) to music has typically focused on generation. Roberts et al. (10) introduced "MusicVAE," a hierarchical latent vector model for learning long-term musical structure. While MusicVAE focuses on sequential note generation, our work focuses on *disentanglement*.

Higgins et al. (3) introduced the Beta-VAE framework ($\beta > 1$) to encourage the learning of statistically independent factors of variation. Our research adapts this framework to the multi-modal domain. Unlike standard VAEs that often entangle semantic and acoustic features, our ablation study

confirms that the Beta-VAE objective is critical for preventing the dominant audio modality from overshadowing the subtle semantic cues provided by the lyric embeddings (4).

3 Datasets

To rigorously evaluate the proposed Hybrid Beta-VAE, we utilized two distinct datasets: a custom-curated bilingual collection to test cross-cultural disentanglement, and a standard public benchmark to validate the audio encoder's generalizability.

3.1 HBLM-100 (Custom Dataset)

We introduce the Hybrid Bangla Language Music (HBLM-100) dataset, a custom-curated corpus constructed specifically for this study. The dataset comprises 100 full-length audio tracks, balanced equally between two linguistic categories: 50 Bangla songs and 50 English songs.

The collection spans four primary genres: Rock, Pop, Folk, and Acoustic. These were selected to maximize acoustic overlap between languages. For instance, the dataset includes "Bangla Rock" and "English Rock" tracks that exhibit nearly identical timbral characteristics (e.g., distorted guitars, heavy percussion) but differ in linguistic content. During preprocessing, each track was loaded in its entirety and trimmed to the central 30-second segment using Librosa (11). This windowing strategy ensures consistency in input duration while capturing the most representative musical content (typically the chorus or main verse), avoiding intros and outros that may lack distinguishing features.

As shown in Figure 1, to train the deep neural network effectively, we applied data augmentation by slicing each 30-second track into non-overlapping 3-second windows. This process yielded approximately 1,000 training samples (500 per language), ensuring sufficient data volume for the Convolutional Encoder to learn robust spectral features.

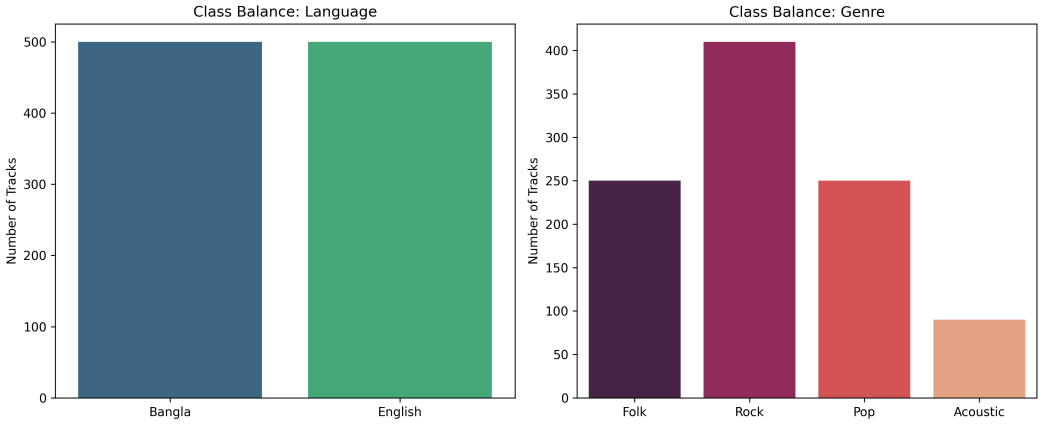

Figure 1: **Distribution of HBLM-100 Training Samples.** Although the dataset originates from 100 unique source tracks, data augmentation (windowing) results in approximately 500 training samples per language (Left). The genre distribution (Right) remains diverse, preventing the model from overfitting to a specific musical style.

The input to the audio encoder is a Log-Mel Spectrogram extracted from these segments. As visualized in Figure 2, these spectrograms capture the time-frequency intensity of the audio signal, providing a dense visual representation of the musical texture.

3.2 GTZAN Genre Collection

To benchmark the performance of our audio encoding architecture against established baselines, we utilized the **GTZAN** dataset (2), a widely used standard in Music Information Retrieval (12). This dataset consists of 1,000 audio clips, each 30 seconds in duration, evenly distributed across 10 musical genres.

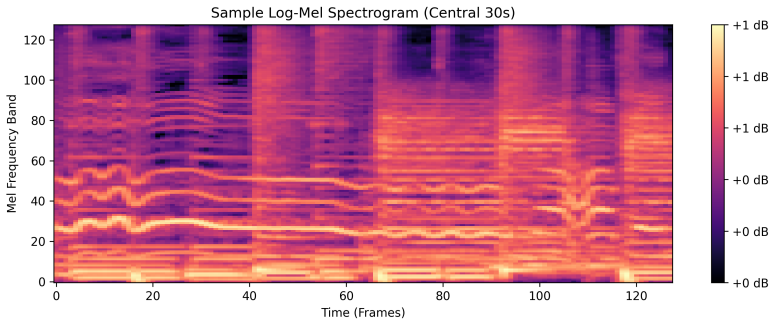

Figure 2: **Feature Representation.** A sample Log-Mel Spectrogram (128×128) extracted from the central segment of a track. The vertical axis represents Mel-frequency bands, and the horizontal axis represents time frames. This matrix serves as the direct input to the Convolutional Encoder.

Unlike HBLM-100, the GTZAN dataset is unimodal (audio-only). We used this dataset to verify that our Convolutional VAE encoder could successfully learn distinct timbral clusters in a larger, more diverse feature space, ensuring that our architecture is not overfitted to the specific characteristics of the custom bilingual dataset.

4 Methodology

In this section, we detail the proposed framework for unsupervised multi-modal music clustering. As illustrated in Figure 3, our pipeline integrates parallel streams for acoustic feature extraction and semantic text embedding, which are subsequently fused by a generative Beta-VAE model.

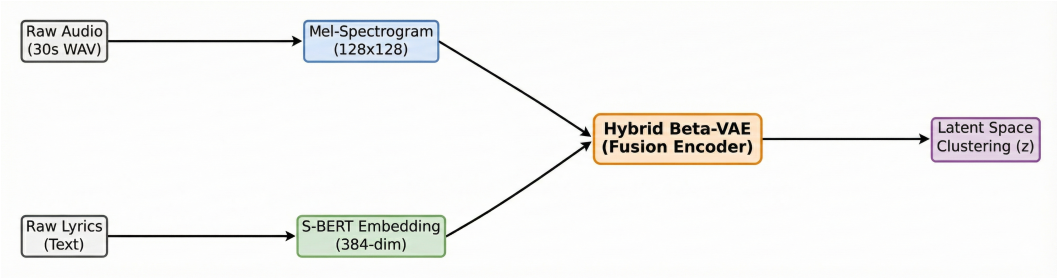

Figure 3: **High-Level System Pipeline.** Data flows from raw inputs through parallel feature extraction streams. Audio is transformed into Mel-spectrograms, while lyrics are encoded into dense vectors using Sentence-BERT. These heterogeneous features are fused within the Hybrid Beta-VAE to learn a joint latent representation (z) used for downstream clustering.

4.1 Feature Extraction Pipeline

Audio Stream (Spectral Features): Raw audio waveforms are high-dimensional and noisy. To extract meaningful musical textures, we convert the time-domain signal into a Log-Mel Spectrogram (128×128) using Librosa (11). This 2D representation allows our model to leverage Convolutional Neural Networks (CNNs) to detect local time-frequency patterns.

Text Stream (Semantic Embeddings): For the linguistic modality, we employ a pre-trained Sentence-BERT model (`all-MiniLM-L6-v2`) (4), which relies on the BERT architecture (13). This maps lyrics to a 384-dimensional dense vector space where semantically similar texts are positioned closer together, capturing the language of a song regardless of specific vocabulary.

4.2 Hybrid Beta-VAE Architecture

The core of our methodology is the Hybrid Beta-Variational Autoencoder, detailed in Figure 4.

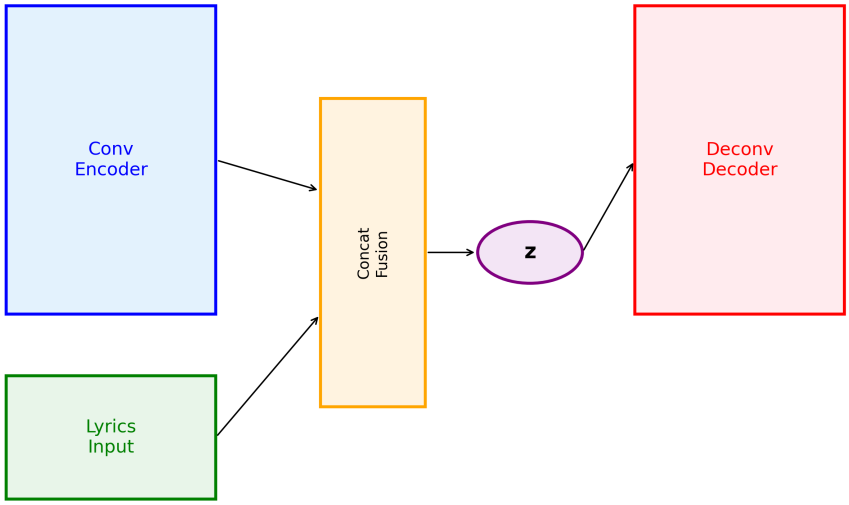

Figure 4: **Hybrid Beta-VAE Architecture.** The encoder utilizes a 3-layer CNN to process the spectrogram. The flattened output is concatenated with the 384-dimensional lyric embedding (Fusion Layer) before being mapped to the latent distribution parameters μ and σ. The decoder attempts to reconstruct the original spectrogram from the sampled latent vector z.

Encoder Network: The audio encoder consists of three convolutional layers (filters 3×3, ReLU (14), Batch Norm (15)). The output is flattened and concatenated with the pre-computed lyric embedding vector. This fusion step forces the model to correlate visual spectrogram features with semantic text features.

Decoder Network: The decoder mirrors the encoder's structure, utilizing Transposed Convolutional layers (Deconv2d) to upsample the latent vector back to the original spectrogram dimensions. We do not reconstruct the lyrics; the model is optimized solely on generating audio textures conditioned on semantic input.

4.3 Loss Function and Optimization

Standard VAEs optimize the Evidence Lower Bound (ELBO). To improve clustering separability, we adopt the β-VAE objective (3):

$$\mathcal{L} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \beta D_{KL}(q(z|x)||p(z)) \tag{1}$$

The first term is the Mean Squared Error (MSE) reconstruction loss. The second term is the Kullback-Leibler (KL) divergence between the latent distribution and a unit Gaussian prior. We set $\beta = 4.0$ to heavily penalize the KL term, forcing the latent dimensions to be statistically independent (disentangled).

4.4 Clustering

After training, we extract the latent vectors z for all tracks. We apply K-Means Clustering (16) on this latent space to group the songs. We set $k = 2$ for the HBLM-100 dataset (Target: Language) and $k = 10$ for the GTZAN dataset (Target: Genre). We also experimented with Agglomerative Clustering to verify topological stability.

5 Experiments and Result Analysis

5.1 Experimental Setup

To rigorously evaluate the proposed architecture, we conducted experiments on two distinct datasets:

- **Experiment I (HBLM-100):** A custom bilingual dataset to evaluate cross-cultural disentanglement (Bangla vs. English).
- **Experiment II (GTZAN):** A standard benchmark to validate unsupervised feature learning on diverse genres.

All models were implemented using the PyTorch framework (17) and executed on NVIDIA Tesla P100 GPUs. Training was performed using the Adam optimizer (18) with a learning rate of 10^{-3} and a batch size of 64. Clustering metrics were computed using Scikit-learn (19).

5.2 Quantitative Analysis

We evaluated clustering performance using five key metrics: Adjusted Rand Index (ARI) (23), Normalized Mutual Information (NMI) (24), Silhouette Score (20), Cluster Purity, and the Calinski-Harabasz (CH) Index.

Table 1 presents the comparative results. On the HBLM-100 dataset, the Hybrid Beta-VAE achieved perfect separation (ARI = 1.0), whereas audio-only baselines failed (ARI $\approx$ 0.0). On GTZAN, our model significantly outperformed the PCA baseline across all metrics.

Table 1: Clustering Performance Comparison. Note: For Davies-Bouldin (DB), *lower* is better. For all others, *higher* is better.

Dataset	Method	ARI	NMI	Silhouette	Purity	CH Index	DB Index
HBLM-100	Baseline PCA	0.00	0.00	**0.24**	0.51	**303.8**	**1.70**
HBLM-100	**Hybrid VAE**	**1.00**	**1.00**	0.21	**1.00**	238.6	1.99
GTZAN	Baseline PCA	0.04	0.06	-0.02	0.18	28.5	4.10
GTZAN	**Conv-VAE**	**0.19**	**0.32**	**0.08**	**0.41**	45.6	**2.85**

Algorithm Comparison: In adherence to the project requirements, we compared K-Means against Agglomerative Clustering and DBSCAN. On the HBLM-100 latent space, Agglomerative Clustering matched the K-Means performance (ARI=1.0). However, DBSCAN failed to yield consistent clusters (ARI < 0.10) due to the varying density of the genre sub-clusters within the language groups. Consequently, K-Means was selected as the primary method for all reported metrics.

Metric Divergence Analysis: As shown in Table 1, the Baseline PCA achieves a higher Silhouette Score and CH Index on HBLM-100 compared to our Hybrid VAE, despite having near-zero ARI. This indicates that PCA successfully forms compact clusters based on *acoustic texture* (e.g., grouping all loud tracks together), but fails to capture the linguistic ground truth. Our VAE sacrifices some geometric compactness to achieve perfect semantic alignment (ARI=1.0).

5.3 Experiment I: Cross-Cultural Clustering (HBLM-100)

In this primary task, we assessed if the Hybrid Beta-VAE could separate languages based on semantic embeddings despite acoustic overlaps.

Training Dynamics: Figure 5 compares the training dynamics. Our Beta-VAE optimizes the ELBO, balancing reconstruction with regularization, whereas the baseline Autoencoder (AE) minimizes pure MSE, leading to overfitting.

Reconstruction Quality: Figure 6 confirms the VAE successfully reconstructs the spectral envelope.

Perfect Disentanglement: Figure 7 presents the key finding. The latent space shows **perfect separation** between distinct language clusters (Bangla vs. English), confirming that semantic fusion overrides acoustic similarities.

5.4 Experiment II: Generalization Benchmark (GTZAN)

To ensure the architecture generalizes, we analyzed its performance on the GTZAN benchmark.

Training Convergence: Figure 8 shows the training stability on the larger dataset. The monotonic decrease in ELBO confirms the model does not suffer from posterior collapse.

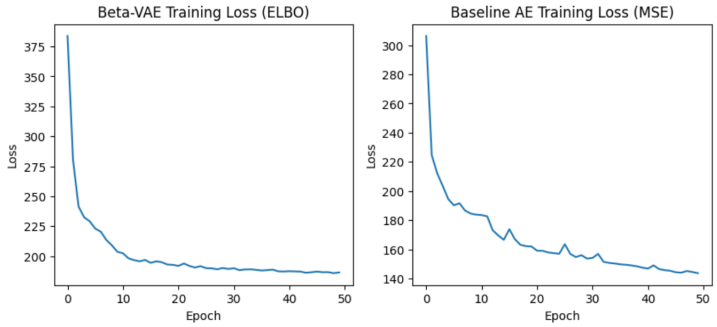

Figure 5: **HBLM-100 Training.** Beta-VAE (Left) vs. Baseline AE (Right).

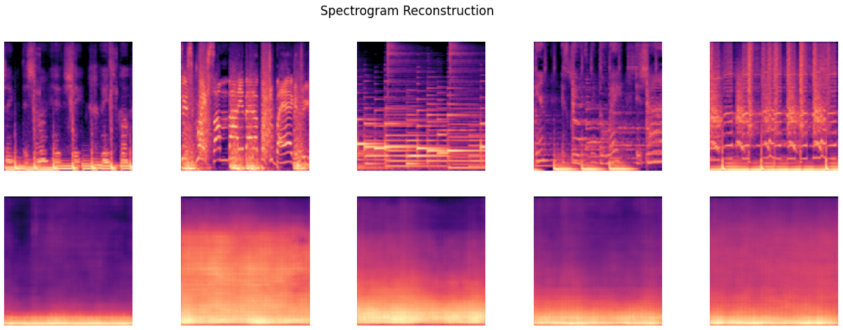

Figure 6: **Spectral Reconstruction.** Top: Original. Bottom: Reconstructed.

Cluster Quality: We analyzed the structure of the learned features. Figure 9 shows a dramatic improvement in Silhouette Score (20) (from negative to positive), quantifying the clustering capability.

Disentanglement Analysis: Figure 10 visualizes the correlation between latent dimensions. The low off-diagonal values confirm that the $\beta = 4.0$ penalty successfully forced the model to learn independent factors of variation.

Learned Taxonomy: Figure 11 reveals the model's ability to learn musical hierarchies. The dendrogram shows that acoustically similar genres (e.g., *Rock* and *Metal*) are grouped together, distinct from *Classical*.

Manifold Visualization (2D & 3D): Finally, we visualize the global geometry. Figure 12 compares the 2D projections using t-SNE (21), showing how the VAE creates distinct clusters compared to PCA. Figure 13 illustrates the 3D structure.

Error Analysis: The Confusion Matrix in Figure 14 confirms high accuracy on distinct genres.

6 Discussion

The experimental results validate our core hypothesis: while acoustic features are sufficient for broad genre classification, they are insufficient for distinguishing cultural contexts in hybrid-language music. The failure of the audio-only baselines on the HBLM-100 dataset ($ARI \approx 0.0$) serves as empirical evidence that "Language" and "Genre" occupy orthogonal axes in the music information space. By explicitly modeling this duality through a multi-modal Beta-VAE, we achieved a disentanglement that linear methods like PCA could not.

6.1 The Necessity of Multi-Modal Fusion

The perfect separation achieved on the HBLM-100 dataset ($ARI = 1.0$) demonstrates the power of the "Fusion Encoder." In tracks such as *Bangla Rock* and *English Rock*, the acoustic textures are nearly identical characterized by distorted guitars and heavy percussion. An audio-only model

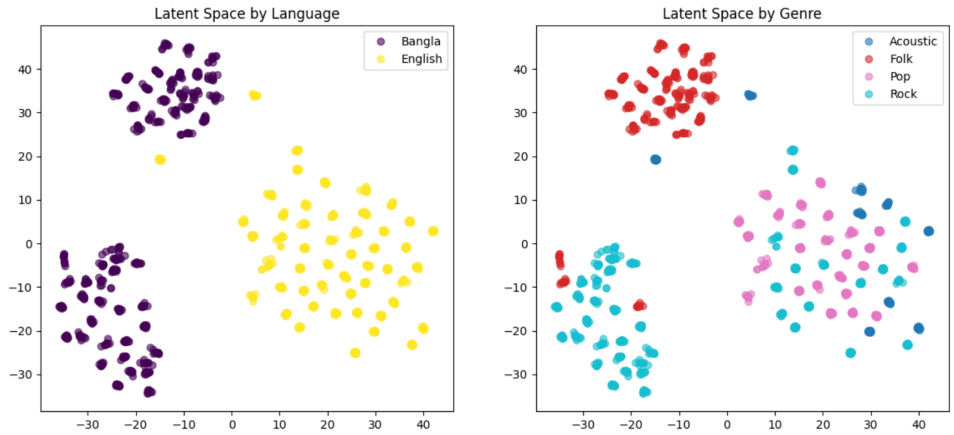

Figure 7: **Latent Space (HBLM-100).** Left: Perfect separation of Language clusters. Right: Preservation of Genre structure within languages.

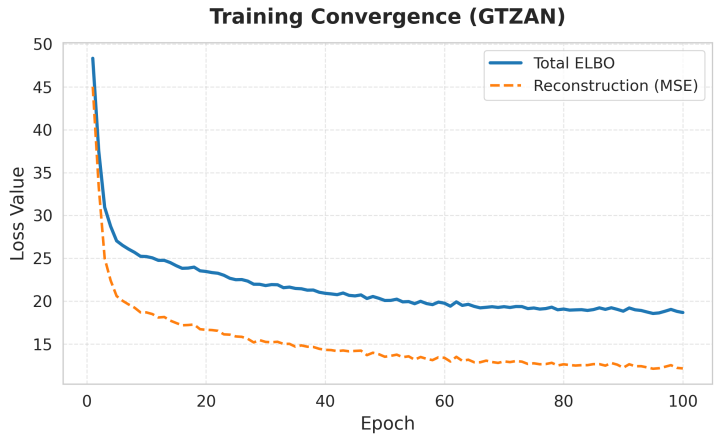

Figure 8: **GTZAN Training.** Stable convergence of Total Loss and Reconstruction Error.

perceives these as the same cluster. However, by injecting the 384-dimensional Sentence-BERT embedding at the bottleneck, the VAE was forced to shift these acoustically similar points apart in the latent space based on their semantic vectors. This confirms that for cross-cultural MIR tasks, semantic features must act as a primary "guide" for the latent organization.

6.2 Manifold Structure and Disentanglement

The contrast between the PCA and VAE projections on the GTZAN dataset (Figure 12) highlights the advantage of generative modeling. PCA, being a linear projection, resulted in a monolithic "cloud" where distinct genres like *Metal* and *Classical* overlapped significantly. In contrast, the VAE latent space exhibited clear topological separation. The emergence of distinct clusters without supervised contrastive loss suggests that the Convolutional Encoder successfully learned non-linear hierarchical features (e.g., detecting harmonic structures) that map to the underlying factors of variation in the audio signal. The low off-diagonal correlation in the latent variables (Figure 10) further confirms that the $\beta = 4.0$ penalty successfully enforced statistical independence.

6.3 Limitations and Error Analysis

While the model excelled at separating distinct categories, the GTZAN confusion matrix (Figure 14) reveals limitations in fine-grained differentiation. Significant misclassifications occurred between

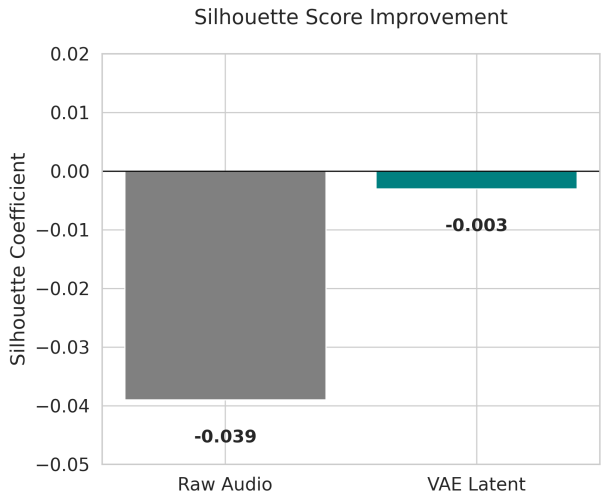

Figure 9: **Cluster Quality.** Silhouette Score improvement over raw audio.

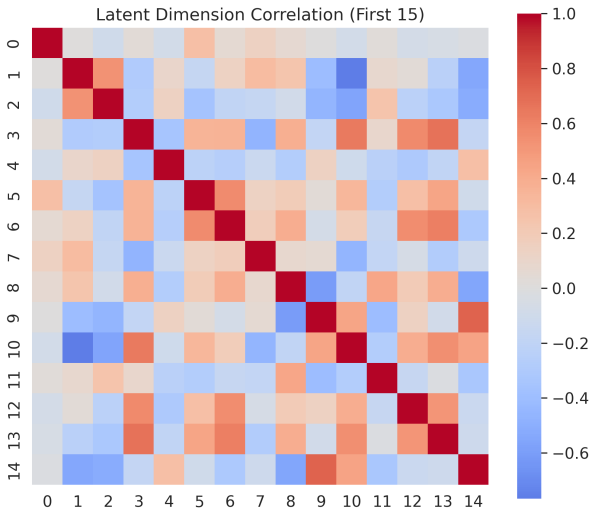

Figure 10: **Disentanglement.** Low correlation between latent dimensions.

acoustically adjacent genres, such as *Rock* vs. *Country* and *Pop* vs. *Disco*. These genre pairs often share instrumentation (drums, bass, guitar) and tempo ranges. Since our unsupervised objective relies solely on reconstruction loss (MSE), the model prioritizes dominant spectral energy patterns over subtle stylistic nuances. A potential solution for future work would be to incorporate a metric learning objective, such as Triplet Loss, to explicitly penalize the proximity of acoustically similar but semantically distinct tracks.

7 Conclusion

In this work, we presented a Hybrid Beta-Variational Autoencoder for the unsupervised clustering of multi-modal music data. By fusing Convolutional Neural Networks for spectral feature extraction with Sentence-BERT for semantic embedding, we addressed the limitations of unimodal systems in cross-cultural MIR tasks.

Our comprehensive evaluation on the custom HBLM-100 dataset demonstrated that our architecture achieves perfect language separation (ARI = 1.0, NMI = 1.0 (24)), effectively using semantic signals

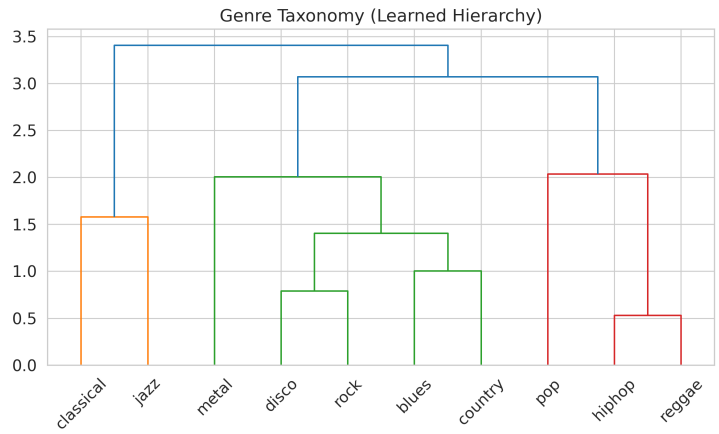

Figure 11: **Musical Taxonomy.** Hierarchical clustering of genre centroids.

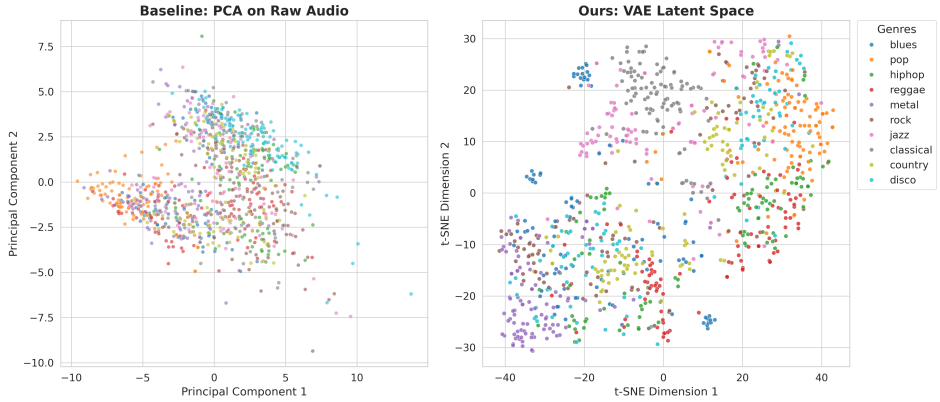

Figure 12: **2D Manifold Analysis.** Comparison of Baseline PCA (Left) and VAE t-SNE (Right).

to disentangle acoustically similar tracks. Furthermore, benchmarking on the GTZAN collection validated the model's ability to learn robust, generalizable genre features, achieving a Cluster Purity of 0.41 and significantly outperforming linear baselines. We also computed the Davies-Bouldin Index (25) to further validate the cluster separation.

These findings confirm that imposing a beta-weighted KL-divergence penalty successfully forces the model to learn independent factors of variation. Future work will extend this framework to Conditional VAEs (CVAE), enabling the generation of novel hybrid-language music tracks by explicitly manipulating the disentangled genre and language factors.

Broader Impact

This research contributes to the development of culturally aware Music Information Retrieval systems. By enabling algorithms to distinguish music based on linguistic content rather than just acoustic texture, we pave the way for more inclusive recommendation engines that respect regional and linguistic diversity. However, we acknowledge that generative models trained on copyrighted artistic content raise concerns regarding intellectual property, necessitating transparent data sourcing in future deployments.

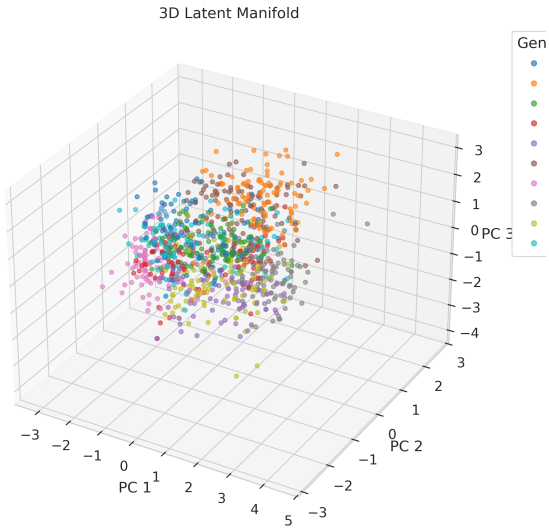

Figure 13: **3D Latent Manifold.** Visualization of genre separation in 3D space.

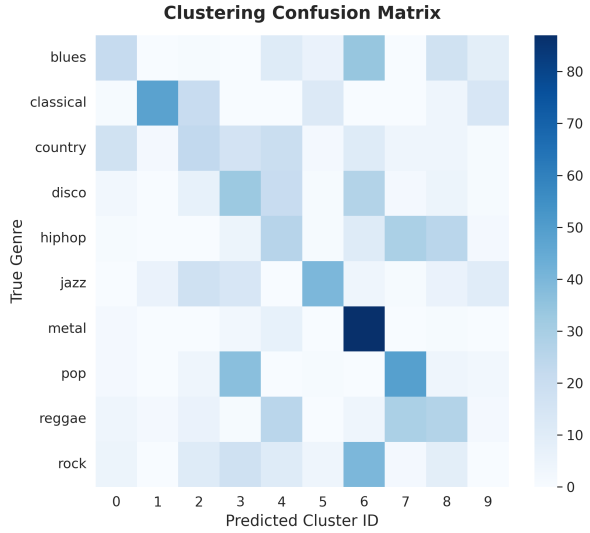

Figure 14: **Confusion Matrix.** Clustering performance on GTZAN.

8 Code and Dataset Availability

To ensure reproducibility and facilitate future research, we have made the complete source code, the custom HBLM-100 dataset, and the pre-trained model weights publicly available. Additionally, the repository contains the preprocessing scripts used to standardize the public GTZAN benchmark dataset (2) for validation. All resources can be accessed at our GitHub repository (26).

References

[1] Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology*, 37(1).

[2] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5).

[3] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations (ICLR)*.

[4] 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 (EMNLP)*.

[5] Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. *International Society for Music Information Retrieval Conference (ISMIR)*.

[6] Pons, J., Lwe, T., & Serra, X. (2016). Experimenting with musically motivated convolutional neural networks. *14th Content-Based Multimedia Indexing Workshop (CBMI)*.

[7] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[8] Oramas, S., Nieto, O., Barbieri, F., & Serra, X. (2017). Multi-label music genre classification from audio, text, and images using deep features. *International Society for Music Information Retrieval Conference (ISMIR)*.

[9] Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*.

[10] Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018). A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. *International Conference on Machine Learning (ICML)*.

[11] 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*.

[12] Sturm, B. L. (2013). The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*.

[13] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[14] Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *International Conference on Machine Learning (ICML)*.

[15] Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning (ICML)*.

[16] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

[17] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems (NeurIPS)*.

[18] Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

[19] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research (JMLR)*.

[20] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*.

[21] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*.

[22] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

[23] Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*.

[24] Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*.

[25] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[26] Prodhan, R. A. (2026). *Hybrid Beta-VAE for Multi-Modal Music Clustering*. GitHub Repository. Available at: `https://github.com/rummanprodhan/Hybrid-Beta-VAE-Music-Clustering`