VAE for Hybrid Language Music Clustering (Easy + Medium + Hard)

Tahsin (ID: 1000054859)
tahsin12zaman
vae-hybrid-music-clustering-Tahsin-ID-1000054859

Abstract

This project implements an unsupervised learning pipeline inspired by Variational Autoencoders (VAEs) to cluster a small multilingual music dataset (English + Bangla) using audio and lyric information. Following the project specification, we complete Easy, Medium, and Hard tasks: (i) a basic VAE for feature extraction and K-Means clustering with a PCA+KMeans baseline, (ii) a convolutional VAE (Conv-VAE) for log-mel spectrograms plus hybrid audio+lyrics features and multiple clustering algorithms, and (iii) a Beta-VAE (and AE baseline) with multimodal clustering using audio, lyrics, and genre metadata, evaluated using Silhouette, ARI, NMI, and Cluster Purity, with latent space and reconstruction visualizations.

1 Introduction

Unsupervised clustering of music can reveal latent structure in timbre, rhythm, language, and style without requiring labeled training data. The project goal is to learn compact representations from audio (and optionally lyrics/metadata) using VAEs, then cluster those representations and evaluate clustering quality using both intrinsic and label-based metrics when partial labels are available. The required pipeline includes comparisons with baselines such as PCA+KMeans and direct spectral feature clustering, as well as visualizations of the learned latent space. (Project requirements: Easy/Medium/Hard tasks, metrics, and report format.)

2 Related Work

VAEs learn continuous latent variables with a probabilistic objective, enabling smooth latent spaces useful for clustering and interpolation. Beta-VAE modifies the KL term weight to encourage disentanglement. For clustering, K-Means is a strong baseline, while Ward agglomerative clustering can capture hierarchical structure and DBSCAN can identify noise/outliers. Multimodal fusion via concatenation or weighted combination is common when audio and text features are both informative.

3 Method

3.1 Dataset and Preprocessing

We use a small hybrid language dataset consisting of English and Bangla tracks. Audio is processed into time-frequency representations (log-mel spectrograms). Lyrics are stored in a CSV and embedded into fixed-dimensional vectors (e.g., TF-IDF + SVD-style embedding pipeline). For the Hard task, each track is associated with a genre value (manually created for this small dataset).

Preprint. Under review.

3.2 Representations

Easy Task (Basic VAE): A lightweight VAE is trained on extracted audio features (baseline spectral features) and the latent means are used as clustering features. A PCA baseline is also computed and clustered with K-Means.

Medium Task (Conv-VAE + Hybrid Audio+Lyrics): A convolutional VAE is trained on log-mel spectrograms; latent means are used as audio embeddings. Lyrics are embedded separately. A hybrid embedding is formed by combining (concatenating or weighted mixing) audio latent vectors and lyric vectors.

Hard Task (Beta-VAE + Multimodal): A Beta-VAE is trained on log-mel spectrograms to encourage more disentangled latent variables. We include additional baselines:

- **PCA+KMeans** on spectral features
- **Autoencoder+KMeans** (AE = Conv-VAE with $\beta = 0$)
- **Direct spectral clustering** (clustering on raw spectral feature statistics)

For multimodal clustering, audio and lyric embeddings are combined; an optional genre one-hot feature can be appended for a genre-aware multimodal feature space.

3.3 Clustering Algorithms

We evaluate:

- **K-Means** (fixed k)
- **Agglomerative Clustering (Ward)**
- **DBSCAN** (density-based; may output noise points)

3.4 Evaluation Metrics

The project requires intrinsic and label-based metrics including Silhouette, Calinski–Harabasz, Davies–Bouldin, ARI, NMI, and Cluster Purity. We treat `language` as a partial label for ARI evaluation (Easy/Medium) and use `language` and `genre` for Hard-task ARI/NMI/Purity.

4 Experiments

4.1 Training Setup

Conv-VAE/Beta-VAE models are trained on log-mel spectrogram inputs with latent dimension 16 (Hard/Medium). For the AE baseline, we set $\beta = 0$ to disable KL regularization. For Beta-VAE we use a nonzero β (e.g., $\beta = 0.1$ in the provided runs). Models are trained for 400 epochs in our main runs.

4.2 Implementation Notes

All tasks are implemented as reproducible scripts under `src/`: `easy_task_scripts/`, `medium_task_scripts/`, `hard_task_scripts/`. Outputs (metrics/plots) are written to `results_easy/`, `results_medium/`, `results_hard/`.

5 Results

5.1 Easy Task: VAE+KMeans vs PCA+KMeans

Table 1 reports the intrinsic clustering metrics required for the Easy task.

Table 1: Easy task clustering metrics (Silhouette, Calinski–Harabasz).

Method	Silhouette	Calinski–Harabasz
VAE+KMeans (dim=8, k=2)	0.4075	14.0420
PCA+KMeans (dim=8, k=2)	0.3310	9.9249

5.2 Medium Task: Conv-VAE + Hybrid Features

We evaluate multiple representations and clustering methods using Silhouette, Davies–Bouldin (lower is better), and ARI with respect to language where applicable.

Table 2: Medium task (selected best-per-representation).

Representation	Best Clustering	Silhouette	ARI (language)
Audio baseline (mean-mel)	KMeans	0.3431	-0.0322
Audio Conv-VAE latent	Ward Agglomerative	0.0526	0.0268
Lyrics SVD embedding	KMeans	0.0212	0.8619
Hybrid ($\alpha = 0.3$ audio+lyrics)	KMeans	0.0202	0.7329
Hybrid ($\alpha \approx 0.1$ audio+lyrics)	KMeans	0.0210	0.8619

5.3 Hard Task: Beta-VAE + Multimodal + Genre

The Hard task requires evaluation with ARI, NMI, and Purity using labels (language/genre) and comparisons against PCA+KMeans, AE+KMeans, and direct spectral clustering.

Table 3: Hard task (selected best-per-representation; genre-based).

Representation	Best Clustering	ARI (genre)	NMI (genre)
Spectral PCA16	KMeans	0.0244	0.3021
Lyrics embedding	Ward Agglomerative	0.2081	0.4972
Beta-VAE audio latent	Ward Agglomerative	0.0043	0.2117
AE audio latent	Ward Agglomerative	0.1279	0.3673
Hybrid audio+lyrics	Ward Agglomerative	0.1309	0.3245

5.4 Visualizations

We generate the required latent space plots (t-SNE) and reconstruction examples. Figures are forced to stay here using `[H]` and `\usepackage{float}`.

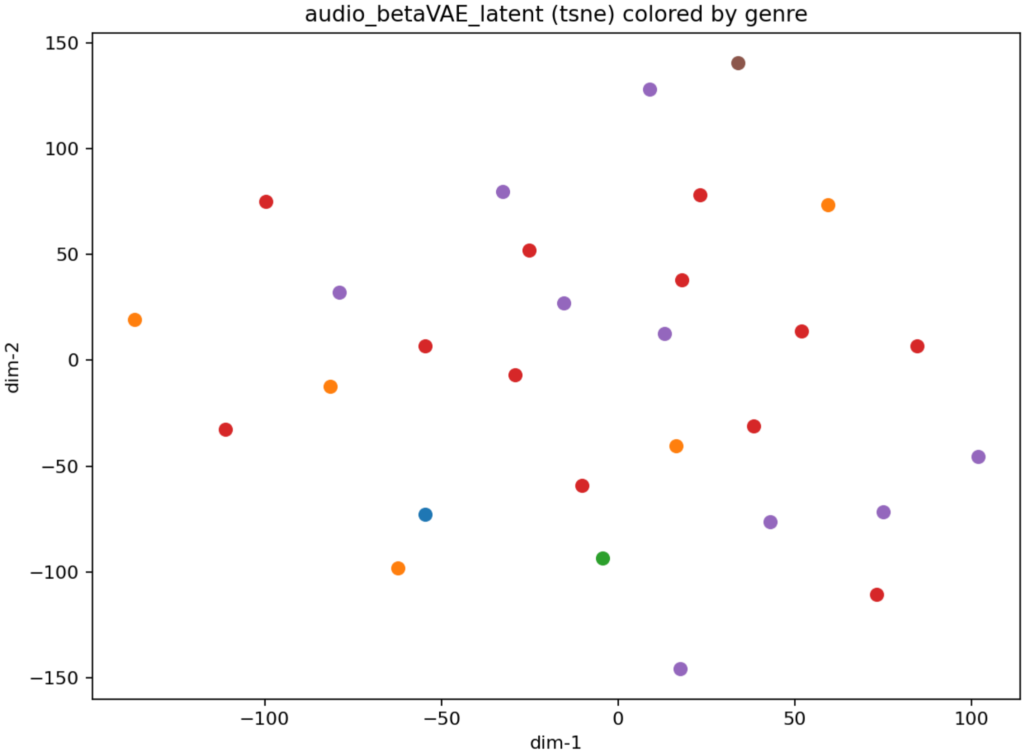

Figure 1: Example latent visualization: Beta-VAE audio latent (t-SNE) colored by genre.

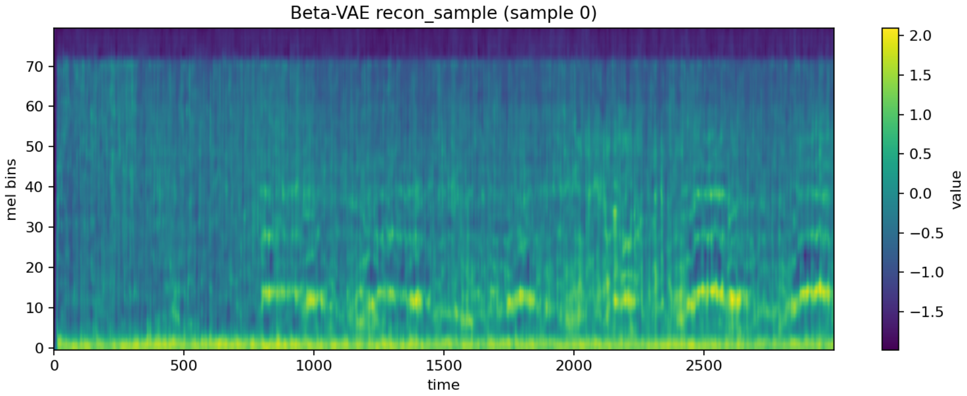

Figure 2: Example reconstruction visualization from Beta-VAE (sample).

6 Discussion

Across tasks, we observe a trade-off between intrinsic clustering structure (silhouette/DB) and label alignment (ARI/NMI). Lyrics embeddings often align strongly with language/genre labels, while audio representations can yield higher silhouette in some settings. However, the dataset is small, which increases metric variance and makes density-based methods (e.g., DBSCAN) prone to producing many noise points depending on hyperparameters. Additionally, including genre as an explicit one-hot feature improves genre alignment by construction; this is useful for genre-aware clustering but should be interpreted carefully.

7 Conclusion

We implemented the complete Easy/Medium/Hard pipelines: basic VAE and PCA baselines; Conv-VAE with hybrid audio+lyrics representations and multiple clustering algorithms; and a Hard-task Beta-VAE/AE setup with multimodal clustering and evaluation using ARI/NMI/Purity along with visualizations and baseline comparisons. Future work includes scaling to larger datasets (e.g., GTZAN/MSD), more robust lyric embeddings (transformer encoders), and systematic hyperparameter sweeps for disentanglement and clustering stability.

References

- D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. (VAE)
- I. Higgins et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. (Beta-VAE)
- L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. (t-SNE)