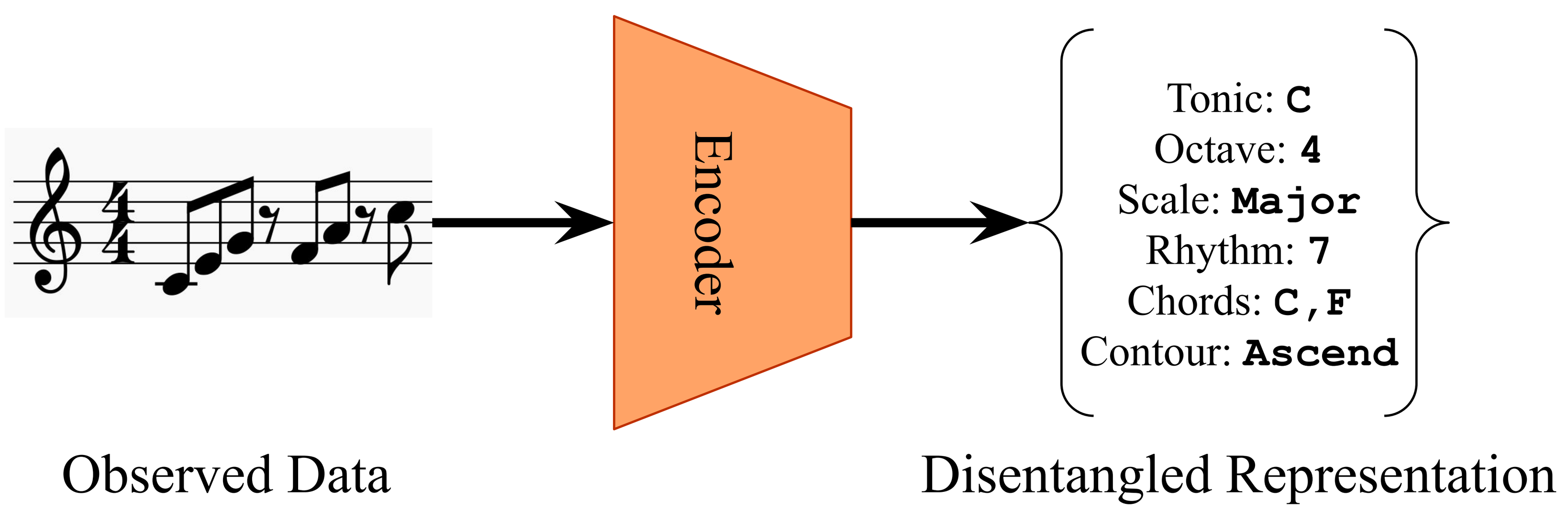
dMelodies: A Music Dataset for Disentanglement Learning

Ashis Pati | Siddharth Gururani | Alexander Lerch

Georgia Tech | Center for Music Technology
College of Design

MOTIVATION

Disentangled representations are low-dimensional representations learnt from high-dimensional data such that the underlying factors of variation are well-separated.



Lack of diversity in disentanglement studies

- majority of methods evaluated using image-based datasets
- easy availability of image-based benchmarking datasets[1]

Lack of consistency in music-based studies

- different datasets used for different studies
- no single benchmarking dataset with well-defined factors of variation

Create a simple, algorithmically generated music-based dataset with clearly defined factors of variation

KEY DESIGN PRINCIPLES

- **Homogenous**: Easy to differentiate between data-points
- **Orthogonal factors**: Changes to one factor should not affect the others. There should be a one-to-one mapping between unique combination of latent factors and the generated datapoints.
- **Diverse types of factors**: Should include categorical & ordinal attributes
- **Large size**: Sufficient to train deep neural networks

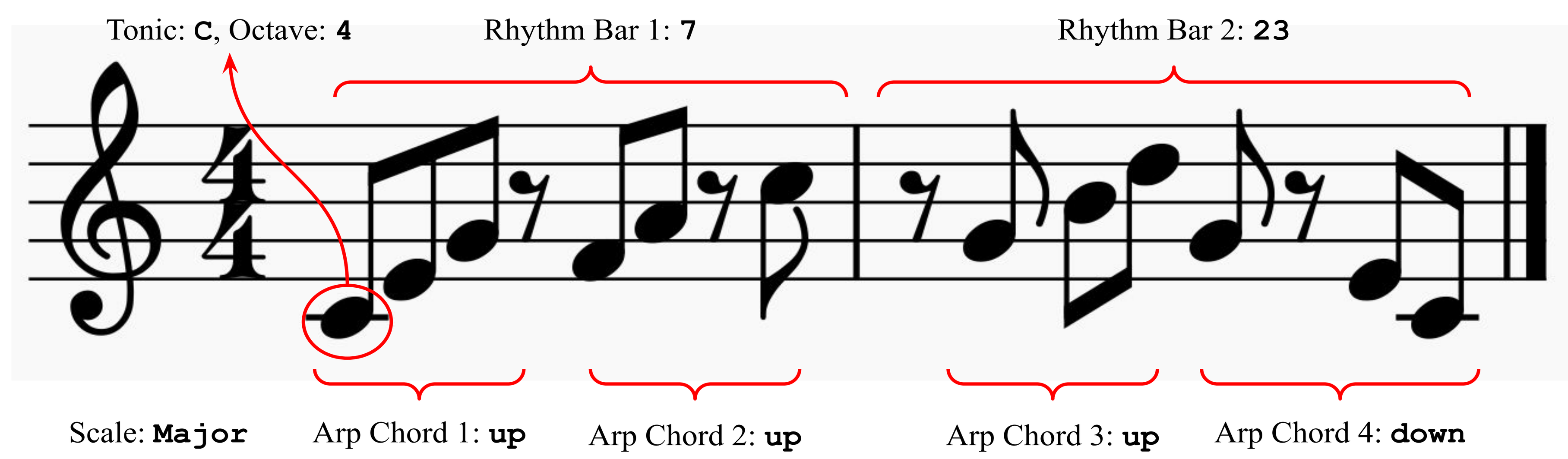
[1] For instance, dSprites, 3D-shapes, MPI3D

[13] Higgins et al., "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in ICLR, 2017
[15] Kim and Mnih, "Disentangling by Factorizing," in ICML, 2018.
[29] Burgess et al., "Understanding disentangling in β-VAE," in NIPS Workshop, 2017.
[45] Pati et al., "Learning to Traverse Latent Spaces for Musical Score Inpainting," in ISMIR, 2019.

DATASET CONSTRUCTION

- **2-bar monophonic melodies**: based on different scales
- **Arpeggios** based on the I-IV-V-I cadence chord pattern with 12 notes per melody.
- **2 chords / bar**: rhythm of each bar is varied

Factor	# Options	Notes
Tonic	12	C, C#, D, ..., through B
Octave	3	Octaves 4, 5, and 6
Scale	3	Major, Harmonic Minor, Blues
Rhythm Bar 1	28	C_6^8, based on onset locations of 6 notes
Rhythm Bar 2	28	C_6^8, based on onset locations of 6 notes
Arp Chord 1	2	up / down
Arp Chord 2	2	up / down
Arp Chord 3	2	up / down
Arp Chord 4	2	up / down



1,354,752 unique melodies

BENCHMARKING EXPERIMENTS

- **3 methods**: β-VAE [13], Annealed-VAE [29], Factor-VAE [15]
- **2 architectures**: CNN-based, Hierarchical RNN-based [45]
- Compare against CNN-based model trained on **dSprites**

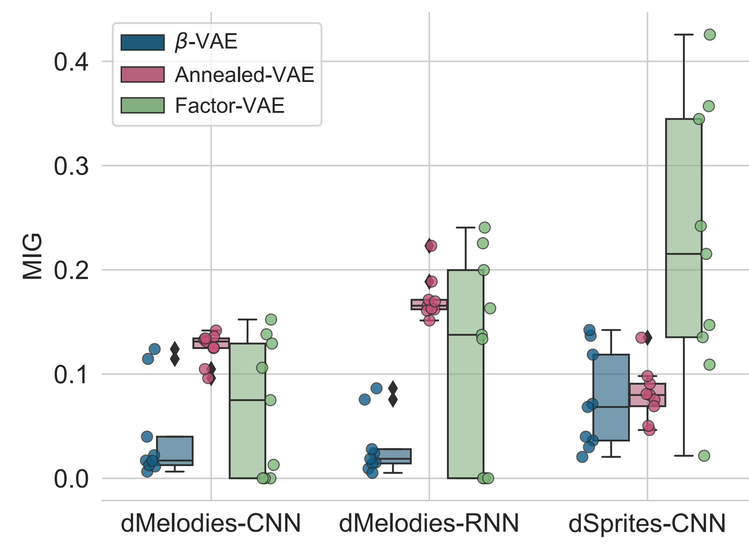


Fig .1 (higher is better)

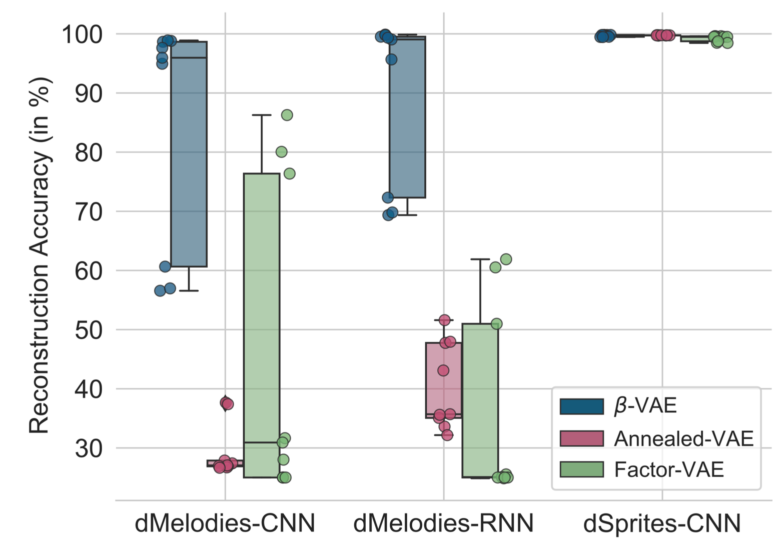


Fig. 2 (higher is better)

RESULTS

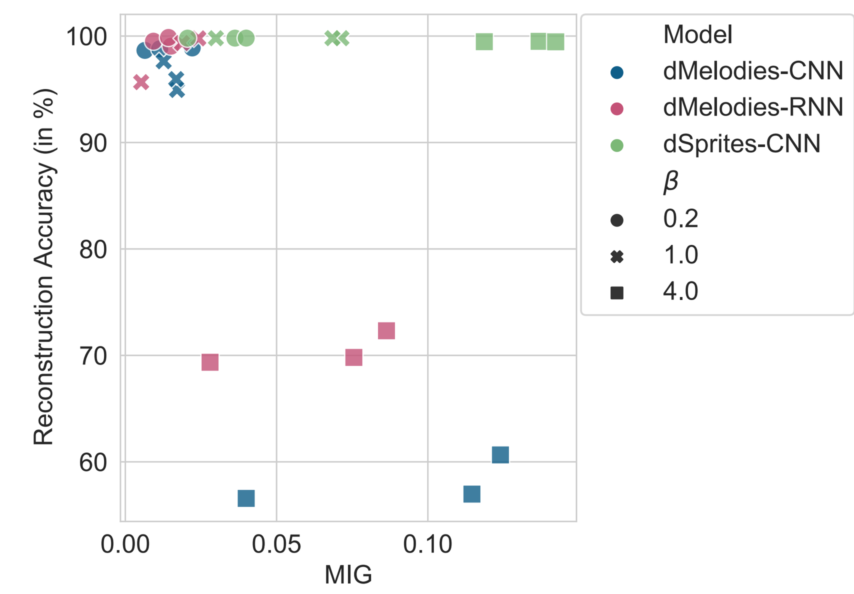


Fig. 3 (top-right is better)

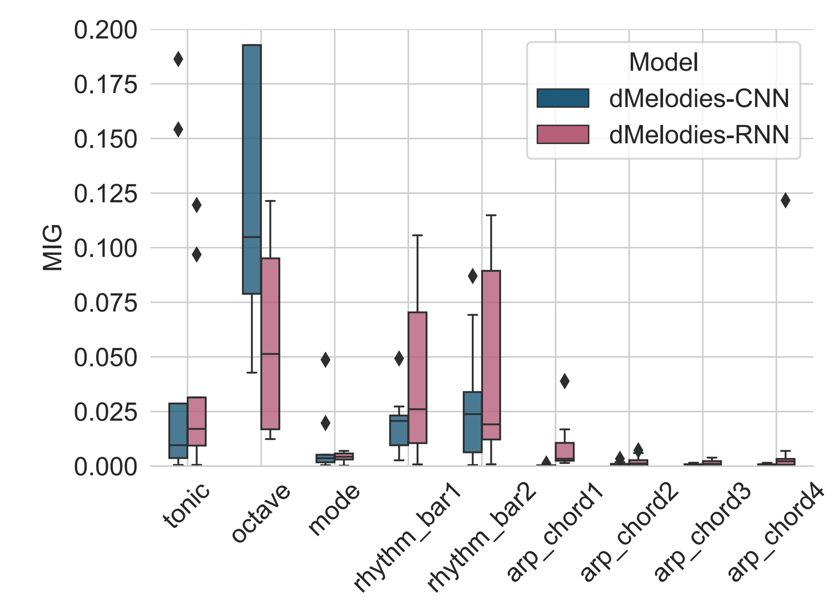


Fig. 4 (higher is better)

- **Disentanglement** (Fig. 1) is **comparable** across datasets and models
- **Reconstruction accuracy** (Fig. 2) for dMelodies is **significantly worse**
- **Sensitivity** to hyperparameters (Fig. 3) is **significantly higher** for dMelodies
- **Some factors** such as octave and rhythm are better disentangled while binary factors perform the worst (Fig. 4) .

KEY TAKEAWAYS

Unsupervised methods do not generalize across domains

Improving disentanglement while maintaining reconstruction fidelity was hard

Modeling diverse factors of variation was challenging

CONTACT

Ashis Pati
Music Informatics Group
Center for Music Technology
Georgia Tech
ashis.pati@gatech.edu

Github Repository