Unsupervised Disentanglement of Pitch and Timbre for Isolated Musical Instrument Sounds

Yin-Jyun Luo, Kin Wai Cheuk, Tomoyasu Nakano, Masataka Goto, Dorien Herremans

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN · Agency for Science, Technology and Research SINGAPORE · AIST

Summary

- Tackle unsupervised disentanglement of pitch and timbre
- Leverage pitch-shifting to further improve disentanglement
- Design a quantitative metric that accounts for disentanglement

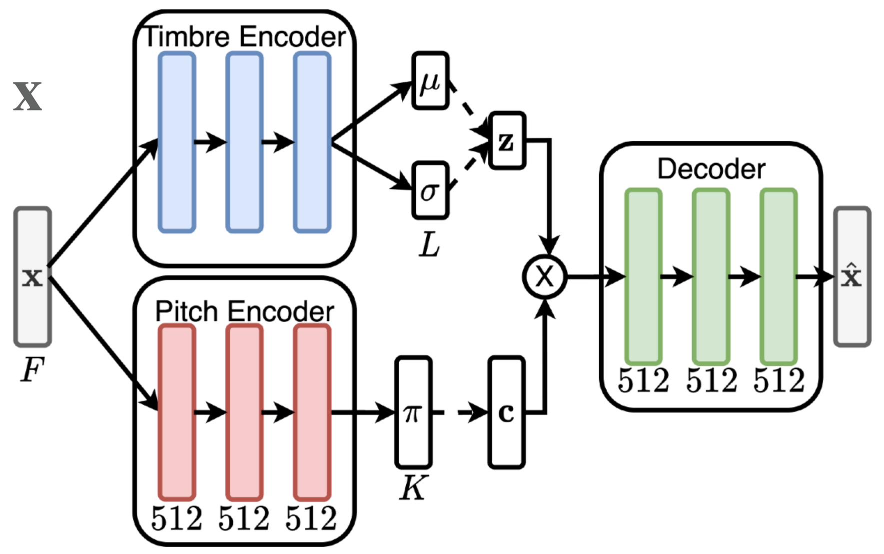
Model

Idea: Introduce inductive biases through architectural constraints

Generation

- Model a note of musical instruments $\mathbf{x}$ as being generated by
  - a pitch (discrete $\mathbf{c}$) and
  - a timbre (continuous $\mathbf{z}$)

  latent variable



- $p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{c}) = p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{c}) p(\mathbf{z}) p(\mathbf{c})$
  - $p(\mathbf{c}) = \mathbf{U}(\mathbf{0}, \mathbf{1})$
  - $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{1})$
  - $p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{c}) = \mathcal{N}(\mu_\theta(\mathbf{z}, \mathbf{c}), \mathbf{1})$, decoder ($D$)

Inference

- Follow the framework of variational inference, introducing a factorized approximated posterior to approximate the true posterior
- Approximated posterior $q_\phi(\mathbf{z}, \mathbf{c} \mid \mathbf{x}) = q_\phi(\mathbf{z} \mid \mathbf{x}) q_\phi(\mathbf{c} \mid \mathbf{x})$
  - $q_\phi(\mathbf{c} \mid \mathbf{x}) = Cat(\mathbf{c} \mid \pi_\phi(\mathbf{x}))$, pitch encoder
  - $q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mu_\phi(\mathbf{x}), diag(\sigma^2_\phi(\mathbf{x})))$, timbre encoder

Learning

- Reparameterization tricks allow for stochastic gradient descent
  - Gaussian for $\mathbf{z}$ [*Kingma et al., ICLR 2014*]
  - Hard Gumbel-softmax for $\mathbf{c}$ (one-hot vectors) [*Jang et al., ICLR 2017*]
- Maximize Evidence Lower BOund (ELBO)

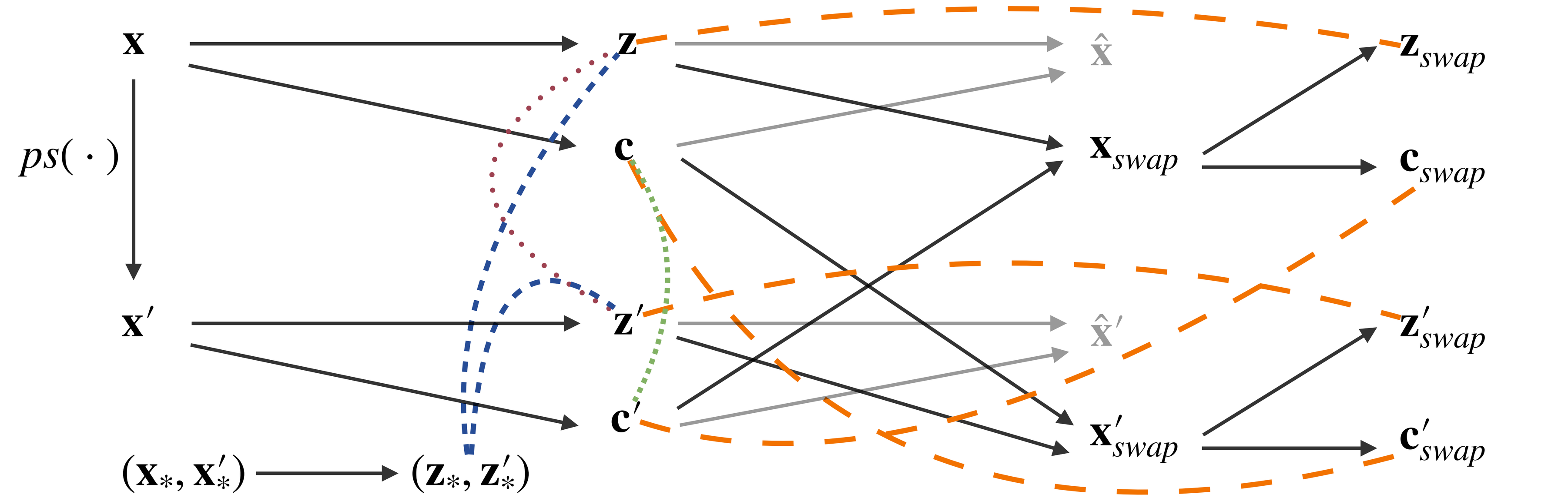
$$\mathcal{L}_{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{c}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{c})] - D_{KL}(q_\phi(\mathbf{z}, \mathbf{c} \mid \mathbf{x}) \| p_\theta(\mathbf{z}, \mathbf{c}))$$

Parameters

- Number of Mel-frequency bins $F = 256$
- Dimension of timbre latent variable $L = 8$
- Number of categories for pitch latent variable $K = 82$

Auxiliary Losses

Assumption: Moderate pitch-shiftings $ps(\cdot)$ do not change timbre



- Create pseudo data pairs $(\mathbf{x}, \mathbf{x}')$ where $\mathbf{x}'$ denotes $\mathbf{x}$ pitch-shifted by δ
- $\mathcal{L}_{regression} = \|\mathbf{z} - \mathbf{z}'\|_2^2$
- $\mathcal{L}_{contrast} = -\log \dfrac{\exp(sim(\mathbf{z}_i, \mathbf{z}'_i)/\tau)}{\sum_{\mathbf{z} \neq \mathbf{z}_i} \exp(sim(\mathbf{z}_i, \mathbf{z})/\tau)}$ [*Chen et al., ICML 2020*]
- $\mathcal{L}_{cycle} = \|\mathbf{z}_{swap} - \mathbf{z}\|_2^2 + \|\mathbf{z}'_{swap} - \mathbf{z}'\|_2^2 + CE(\mathbf{c}_{swap}, k') + CE(\mathbf{c}'_{swap}, k)$, where $k = \arg\max(\mathbf{c})$ [*Zhu et al., ICCV 2017*]
- $\mathcal{L}_{surrogate} = CE(\mathbf{c}', y')$, where pseudo pitch label $y' = \arg\max(\mathbf{c}) + \delta$
- The final objective function to be maximized becomes

$$\mathcal{L} = \mathcal{L}_{ELBO} - (\lambda_1 \mathcal{L}_{regression} + \lambda_2 \mathcal{L}_{contrast} + \lambda_3 \mathcal{L}_{cycle} + \lambda_4 \mathcal{L}_{surrogate})$$

Evaluation

Pitch Variable

- Pitch classification accuracy (ACC and pitch mapping, need labels)
- Consistency-Diversity Score (CDS) $= \mathbb{E}_k \left[D_{KL}(p_k(\mathbf{y} \mid \hat{\mathbf{x}}) \| \mathbb{E}_k[p_k(\mathbf{y} \mid \hat{\mathbf{x}})]) \right]$;

  $p_k(\mathbf{y} \mid \hat{\mathbf{x}}) = p(\mathbf{y} \mid D(\mathbf{z}, \mathbf{c}_k))$ is posterior of a pre-trained pitch classifier, where the one-hot vector $\mathbf{c}_k$ denotes $\mathbf{c} \ni k = \arg\max(\mathbf{c})$
  - $p_k(\mathbf{y} \mid \hat{\mathbf{x}})$ should be consistent and have low entropy given a fixed $\mathbf{c}_k$
  - $\mathbb{E}_k[p_k(\mathbf{y} \mid \hat{\mathbf{x}})]$ should be as uniformly distributed as possible

Timbre Variable

- Pitch and Instrument classification accuracy (need labels)
- Fréchet Inception Distance (FID) [*Heusel et al., NeurIPS 2017*]
  - FID_{recon}: FID between true and reconstructed data (upper-bound)
  - FID_{rand}: FID between true and randomly sampled data

Dataset

- Studio-On-Line [*Ballet et al., JIM 1999*]
- 1,885 samples of 12 musical instruments and 82 pitches
- waveform (22,050Hz) → STFT (w = 92ms, h = 11ms) → Mel-Spec ($F = 256$) → log-scaled → normalized to $[-1, 1]$ → $\mathbf{x}$ (200-th frame)

Quantitative Results

λ_1	λ_2	λ_3	λ_4		Pitch	Instrument	Combine	ACC	PM	FID_{recon}	FID_{rand}	CDS
				♭	8.81±3.47	87.68±1.09	89.43±1.85	95.14±0.98	96.04±0.71	21.80±1.05	23.78±1.47	24.33±0.71
0	0	0	0	♯	33.78±7.38	80.90±4.41	73.55±5.77	72.65±4.82	74.46±4.06	24.86±2.27	25.27±1.80	8.49±1.96
				M0	16.38±7.65	86.44±2.20	85.02±4.03	78.53±5.68	80.22±6.01	23.93±1.97	26.40±2.39	11.45±2.34
1	0	0	0	M1	17.85±4.52	87.34±1.26	84.74±2.53	77.28±3.47	78.75±3.60	18.86±1.77	21.53±1.10	9.15±1.28
0	1	0	0	M2	20.45±7.98	84.74±2.67	82.14±5.17	77.40±5.01	79.09±6.08	26.00±1.78	26.90±2.28	9.20±1.55
0	0	1	0	M3	32.54±6.28	84.18±1.92	75.81±4.08	**80.45±1.58**	**82.71±1.26**	18.68±2.36	20.82±1.67	10.79±2.37
0	0	0	1	M4	17.06±3.83	84.18±1.38	83.55±1.84	74.35±2.75	75.59±3.32	22.36±2.36	24.74±2.17	11.99±2.67
1	1	1	0	M5	18.19±4.79	**87.90±1.62**	84.85±2.48	78.19±2.35	79.66±2.81	16.73±2.13	21.39±2.49	9.35±2.81
1	1	1	1	M6	**14.57±2.29**	86.44±2.55	**85.93±2.06**	79.88±1.84	80.90±2.18	**13.76±1.07**	**19.18±1.90**	**13.46±1.64**

♭: Supervised model trained with pitch labels

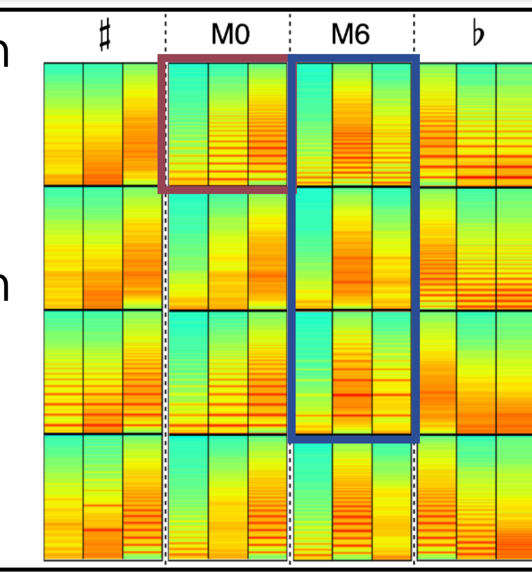
♯: Unsupervised model trained without pitch-shifting

M0 - M6: Proposed unsupervised models with different losses activated

- Supervised model does not yield good generation quality (FID)
- Pitch-shifting alone improves disentanglement
- No auxiliary loss alone yields consistent improvement for all metrics
- Activating $\mathcal{L}_{surrogate}$ on top of the rest reaches the best-performing model (M5→M6)

Qualitative Results

- Perform pitch-conditioning spectrum generation
  - Last row: seeds (three seeds per model)
  - First to third rows: three different k's
- Spectral distribution stays consistent per column
- Spectrums generated given a k are expected to have a consistent pitch (consistency)
- Different k's render different pitches (diversity)



Future Works

- Perform pitch mapping without referring to pitch labels
- Trade off between capacity and constraint for pitch representation $\mathbf{c}$
- Model larger time scale (temporal variable)