Unsupervised Disentanglement of Pitch and Timbre for Isolated Musical Instrument Sounds

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

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 \( x \) as being generated by:
      - a pitch (discrete \( c \)) and
      - a timbre (continuous \( z \))
    - \( p_h(x, z, c) = p_h(x | z, c)p(z|c)p(c) \)
    - \( p(z) = N(0, 1) \)
    - \( p_h(x | z, c) = N(p_h(z, c), 1) \), decoder \((D)\)
  - **Inference**
    - Follow the framework of variational inference, introducing a factorized approximated posterior to approximate the true posterior
    - Approximated posterior \( q_l(z, c | x) = q_l(z | x)q_l(c | x) \)
      - \( q_l(z | x) = \mathcal{N}(\mu_l(z, x), \sigma_l(z, x)^2) \), timbre encoder
    - \( q_l(c | x) = \text{Cat}(c | \phi_l(x)) \), pitch encoder
  - **Learning**
    - Reparameterization tricks allow for stochastic gradient descent
      - Gaussian for \( z \) [Kingma et al., ICLR 2014]
      - Hard Gumbel-softmax for \( c \) (one-hot vectors) [Jang et al., ICLR 2017]
    - Evaluate Maximum Evidence Lower Bound (ELBO)
      - \( \mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_l} [\log p_h(z, c | x)] - D_{KL}(q_l(z, c | x) || p_h(z, c)) \)
  - **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

- Assume: Moderate pitch-shiftings \( p_h(\cdot) \) do not change timbre
  - \( p_h(x, z, c) \rightarrow x' \) where \( x' \) denotes \( x \) pitch-shifted by \( \delta \)
  - \( \mathcal{L}_{\text{regression}} = \|x - x'|^2 \)
  - \( \mathcal{L}_{\text{contrast}} = -\log \frac{\sum \exp(\text{sim}(x, z))}{\exp(\text{sim}(x, z'))} \) [Chen et al., ICLR 2020]
  - \( \mathcal{L}_{\text{cycle}} = \|z \|_2^2 + \|z' \|_2^2 + CE(z, k') + CE(z', k) \)
    - where \( k = \text{arg max}_c (\text{Zhu et al., ICCV 2017}) \)
    - \( \mathcal{L}_{\text{surrogate}} = CE(c', y') \)
      - \( y' = \text{arg max}_c (\delta) \)
      - \( \mathcal{L} = \mathcal{L}_{\text{ELBO}} - (\lambda_1 \mathcal{L}_{\text{regression}} + \lambda_2 \mathcal{L}_{\text{contrast}} + \lambda_3 \mathcal{L}_{\text{cycle}} + \lambda_4 \mathcal{L}_{\text{surrogate}}) \)

Evaluation

- Pitch Variable
  - Pitch classification accuracy (ACC and pitch mapping, need labels)
  - Consistency-Diversity Score (CDS) \((\text{CDS}) = \frac{1}{|D|} \sum_{D \in \text{data}} |p_h(y | x, \hat{c})| \)
- Timbre Variable
  - Pitch and Instrument classification accuracy (need labels)
  - Fréchet Inception Distance (FID) [Heusel et al., NeurIPS 2017]
    - \( \text{FID}_\text{true} \): FID between true and reconstructed data (upper-bound)
    - \( \text{FID}_\text{true} \): FID between true and randomly sampled data

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-conditioning without referring to pitch labels
  - Trade off between capacity and constraint for pitch representation \( c \)
  - Model larger time scale (temporal variable)