SOURCE-LEVEL PITCH AND TIMBRE EDITING FOR MIXTURES OF TONES USING DISENTANGLED REPRESENTATIONS

Yin-Jyun Luo[1] **Kin Wai Cheuk**[2] **Woosung Choi**[2] **Toshimitsu Uesaka**[2]
Keisuke Toyama[3] **Wei-Hsiang Liao**[2] **Simon Dixon**[1] **Yuki Mitsufuji**[2,3]
[1] C4DM, Queen Mary University of London [2] Sony AI [3] Sony Group Corporation
yin-jyun.luo@qmul.ac.uk, kinwai.cheuk@sony.com

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

ABSTRACT

We propose a model to learn latent representations of pitch and timbre of each individual source of instrument tones from a mixture of instruments. We employ variational autoencoders to train the model using a query-based inference network. Given a mixture, the model allows for precise source-level attribute editing, e.g., instrument or pitch replacement, by manipulating the pitch and timbre latents. On the synthetic audio clips of chords compiled using the JSB Chorales dataset, our quantitative evaluation protocol shows empirical success of the model on both pitch-timbre disentanglement of individual sources and source-level attribute manipulation of mixtures.

1. INTRODUCTION

Disentangled representation (DR) learning captures semantically meaningful latent features of observed data in low-dimensional latent spaces [1]. We propose a model that extracts a DR for each source from a mixture of tones of musical instruments. Each source-level DR encodes two attributes, i.e., pitch and timbre, of an instrument in separate groups of dimensions of the representation. Based on variational autoencoders (VAEs) [2], the model encodes a source-level DR given both a mixture and a query. The query specifies a target source sharing the same timbre characteristic. The source-level DRs form a set whose size is equal to the number of sources in the mixture. A decoder then takes as input the set to reconstruct the given mel spectrogram of the mixture. This study demonstrates a proof of concept towards precise and compositional music editing. It is precise because a user can explicitly edit the attributes of specific sources of interest by manipulating the source-level DRs. For example, given a mixture of piano and flute alongside a query which is an arbitrary sample of piano, one can extract the source-level DR of the piano source and swap its pitch or timbre latent for one extracted from another sample that carries the desired attributes. It is compositional because the decoder can render a novel mixture given combinations of the desired attributes.

Prior work on pitch and timbre disentanglement has mainly considered inputs of monophonic instruments [3–10], with only a few studies addressing mixtures of instruments. Hung et al. [11] and Cwitkowitz et al. [12] encode the overall timbre of a mixture and consider applications of symbolic music rearrangement and transcription, respectively. Cheuk et al. [13] propose a multi-task framework that conditions music pitch transcription on intermediate timbre information. Lin et al. [14] tackle source separation and generation by a unified framework. While their model is capable of generating a single source conditioned on a pair of learnt pitch and timbre representations, we further sample novel mixtures by a conditional decoder that takes as inputs a set of source-level DRs, thereby allowing for precise and compositional music editing.

2. METHOD

We describe a generative model as follows:

$$p(x_m|z_m)p(z_m|\mathcal{Z}_s)p(\mathcal{Z}_s|\mathcal{T}_s,\mathcal{V}_s)p(\mathcal{T}_s)p(\mathcal{V}_s|\mathcal{Y}_s),$$

where $x_m \in \mathbb{R}^{d\times t}$ and $z_m \in \mathbb{R}^l$ denote the mel spectrogram of the mixture and its corresponding latent variable. Let $\mathcal{Z}_s$ denote the set $\{z_s^{(i)} \in \mathbb{R}^l\}_{i=1}^{N_s}$, where $z_s^{(i)}$ is the i-th source-level latent out of N_s sources that make up x_m. $\mathcal{T}_s = \{\tau_s^{(i)} \in \mathbb{R}^{l_\tau}\}_{i=1}^{N_s}$ and $\mathcal{V}_s = \{\nu_s^{(i)} \in \mathbb{R}^{l_\nu}\}_{i=1}^{N_s}$ are the sets of timbre and pitch latents. $\mathcal{Y}_s = \{y_s^{(i)}\}_{i=1}^{N_s}$ where $y_s^{(i)}$ is the pitch annotation of source i.

We first sample a set of timbre latents from a prior distribution $p(\mathcal{T}_s)$ and another set of pitch latents conditioned on the specified pitch values $\mathcal{Y}_s$. The two sets of latents are combined for the source-level latents $\{z_s^{(i)}\}_{i=1}^{N_s}$ which are then used to parameterise the distribution over the mixture latent z_m. Finally, we sample the mixture x_m from $p(x_m|z_m)$. The generative process aims to disentangle the pitch and timbre of each individual source from a mixture.

2.1 Dataset

We use a dataset compiled by Gha et al. [15] to train and evaluate our model. It is an audio dataset of 3,131 unique chords from the JSB Chorales dataset [16], rendered with sound fonts of piano, violin, and flute via FluidSynth, whereby the possible number of sources for a mixture is $N_s \in \{1, 2, 3\}$. The dataset is split into 19,719 training, 5,634 validation, and 2,826 testing samples. Each chord sample's notes are synthesized into 16kHz audio waveforms, summed to form the chord's waveform. These are converted into mel spectrograms using 128 mel-filter bands, a window size of 1024, and hop length of 512. We crop a 320ms segment from the sustain phase of each sample to focus on the steady-state spectral distribution.

Given a chord, the rendering process iterates through each composite note and synthesises it with a sound font randomly chosen from the three sound fonts. As a result, any of the sound fonts can play multiple notes in a chord. The mel spectrogram of a chord is the mixture x_m and that of the notes rendered by a single sound font is a source $x_s^{(i)}$. While the dataset has a maximum of three sources for a mixture x_m given the three selected sound fonts, in principle our model can handle arbitrary numbers of sources.

2.2 Training objectives

We employ VAEs [2] to learn our model. The objective function builds on an evidence lower bound (ELBO) to the marginal log-likelihood $\log p(x_m|\mathcal{Y}_s)$:

$$\mathcal{L}_{\text{ELBO}} = \log p_{\theta_x}\big(x_m|z_m\big) + \mathbb{E}_{q_{\phi_\tau}(\mathcal{T}_s)} \log p\big(z_m|\mathcal{T}_s, \mathcal{V}_s\big)$$
$$- \sum_{i=1}^{N_s} \mathcal{D}_{\text{KL}}\big(q_{\phi_\tau}(\tau_s^{(i)}) \| p(\tau_s^{(i)})\big) + \log p_{\theta_\nu}\big(\mathcal{V}_s|\mathcal{Y}_s\big).$$

The first term reconstructs x_m given $z_m = \mathrm{E}_{\phi_m}(x_m)$ through a decoder p_{θ_x}. The second term encourages z_m to fit a distribution parameterised by $\mathcal{T}_s$ and $\mathcal{V}_s$, defined by $p(z_m|\mathcal{T}_s, \mathcal{V}_s) = \mathcal{N}\big(\mu_m(\mathcal{Z}_s) = \sum_{i=1}^{N_s} z_s^{(i)}, \sigma_m^2 I\big)$, where $\sigma_m = 0.25$ and $z_s^{(i)} = \mathrm{FiLM}(\tau_s^{(i)}, \nu_s^{(i)})$. $\mathrm{FiLM}(\cdot, \cdot)$ is a modulation [17] used to "stylise" the pitch with timbre [18]. Specifically, $\mathrm{FiLM}(\tau_s^{(i)}, \nu_s^{(i)}) = \alpha^{(i)}\nu_s^{(i)} + \beta^{(i)}$ and $(\alpha^{(i)}, \beta^{(i)}) = \mathrm{MLP}(\tau_s^{(i)})$. By maximising this likelihood during training, we can composite a novel mixture by passing to the decoder the sum of a set of source-level latents $\sum_{i=1}^{N_s} \mathrm{FiLM}(\tau_s^{(i)}, \nu_s^{(i)})$ with the desired attributes specified by $\{\tau_s^{(i)}, \nu_s^{(i)}\}$.

The third term regularises outputs of the stochastic timbre encoder $q_{\phi_\tau}(\mathcal{T}_s) := \prod_{i=1}^{N_s} q_{\phi_\tau}(\tau_s^{(i)}|x_m, x_q^{(i)})$ to a prior $p(\mathcal{T}_s) = \prod_{i=1}^{N_s} p(\tau_s^{(i)})$, where $q_{\phi_\tau}(\tau_s^{(i)}|\cdot) = \mathcal{N}\big(\mu_{\phi_\tau}(\cdot), \sigma_{\phi_\tau}(\cdot)\big)$ and $p(\tau_s^{(i)}) = \mathcal{N}\big(0, 1\big)$. The last term maximises a likelihood of outputs of a deterministic pitch encoder $\mathcal{V}_s = \mathrm{f}_{\phi_\nu}(\hat{\mathcal{Y}}_s^{\text{bin}})$ given the ground-truth set of pitch values $\mathcal{Y}_s$. $\hat{\mathcal{Y}}_s^{\text{bin}} = \{\mathrm{SB}(\mathrm{E}_{\phi_\nu}(x_m, x_q^{(i)}))\}_{i=1}^{N_s}$ is a set of transcribed pitches, each source i corresponding to a multi-hot output [19] aligned with the fact that a source can play multiple notes, extracted from a deterministic function $\mathrm{E}_{\phi_\nu}(\cdot)$ and a stochastic binarisation layer $\mathrm{SB}(\cdot)$.

Importantly, both $\mathrm{E}_{\phi_\nu}(\cdot)$ and $q_{\phi_\tau}(\cdot)$ take as inputs the mixture x_m and a query $x_q^{(i)}$ to derive $\nu_s^{(i)}$ and $\tau_s^{(i)}$, respectively. This is motivated by query-based source separation [14, 20, 21]. $x_q^{(i)}$ is another mel spectrogram rendered by the same sound font as $x_s^{(i)}$.

	Disentanglement		Mixture Editing	
	Pitch	Inst.	Pitch	Inst.
The proposed	93.39%	100.00%	90.69%	100.00%
- $\mathcal{L}_{\text{BT}}$	93.18%	99.92%	87.92%	100.00%
- KLD	69.41%	100.00%	35.10%	100.00%
- SB	93.46%	46.71%	40.23%	98.91%

Table 1. Accuracy for various loss configurations.

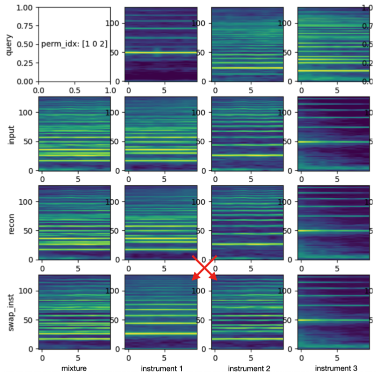

Figure 1. Source-level attribute swapping. The first two sources exchange their timbre while preserving their pitch.

We also include auxiliary loss terms to reconstruct the individual sources and classify their pitches:

$$\mathcal{L}_{\text{aux}} = \sum_{i=1}^{N_s} \mathbb{E}_{q_{\phi_\tau}(\tau_s^{(i)})} \log p_{\theta_x}\big(x_s^{(i)}|\tau_s^{(i)}, \nu_s^{(i)}\big)$$
$$+ \sum_{i=1}^{N_s} \log p(\hat{y}_s^{(i)} = \mathrm{E}_{\phi_\nu}(x_m, x_q^{(i)})|y_s^{(i)}). \tag{1}$$

Finally, we include a variant of Barlow Twins [22]:

$$\mathcal{L}_{\text{BT}} = \sum_{i=1}^{N_s} \sum_{d=1}^{l_\tau} (1 - \mathcal{C}_{dd}(z_q^{(i)}, \tau_s^{(i)}))^2, \tag{2}$$

where $\mathcal{C}$ is a cross-correlation matrix, to promote invariance between $z_q^{(i)} = \mathrm{E}_q(x_q^{(i)})$ and $\tau_s^{(i)}$, as the query and the corresponding timbre latent are supposed to carry highly correlated information. In summary, we maximise:

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{aux}} - \mathcal{L}_{\text{BT}}. \tag{3}$$

3. RESULT

We employ pre-trained pitch and instrument classifiers to quantitatively assess pitch-timbre disentanglement and source-level attribute editing given mixtures. Tab. 1 shows that the success of disentanglement relies on the Kullback-Leibler divergence and the stochastic binarisation layer, which impose critical bottleneck to the timbre and the pitch latents, respectively. We also illustrate in Fig. 1 successful attribute swapping between the first two sources.

4. CONCLUSION

We have proposed a framework that disentangles pitch and timbre from mixtures and demonstrated its application for music attribute editing. Future work will focus on extending the framework beyond the synthetic dataset.

5. REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” in *Trans. on Pattern Analysis and Machine Intelligence*, 2013.

[2] D. P. Kingma and M. Welling, “Auto-encoding Variational Bayes,” in *Int. Conf. on Learning Representations*, 2014.

[3] Y.-J. Luo, K. Agres, and D. Herremans, “Learning Disentengled Representations of Timbre and Pitch for Musical Instrument Sounds Using Gaussian Mixture Variational Autoencoders,” in *Int. Soc. for Music Information Retrieval*, 2019.

[4] O. Cífka, A. Ozerov, U. Şimşekli, and G. Richard, “Self-Supervised VQ-VAE for One-Shot Music Style Transfer,” in *Int. Conf. on Acoustics, Speech and Signal Processing*, 2021.

[5] K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii, and S. Morishima, “Pitch-Timbre Disentanglement Of Musical Instrument Sounds Based On Vae-Based Metric Learning,” in *Int. Conf. on Acoustics, Speech and Signal Processing*, 2021.

[6] Y.-J. Luo, K. W. Cheuk, T. Nakano, M. Goto, and D. Herremans, “Unsupervised Disentanglement of Pitch and Timbre for Isolated Musical Instrument Sounds,” in *the Int. Soc. for Music Information Retrieval*, 2020.

[7] K. Tanaka, Y. Bando, K. Yoshii, and S. Morishima, “Unsupervised Disentanglement of Timbral, Pitch, and Variation Features From Musical Instrument Sounds With Random Perturbation,” in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022.

[8] Y.-J. Luo, S. Ewert, and S. Dixon, “Towards Robust Unsupervised Disentanglement of Sequential Data — A Case Study Using Music Audio,” in *Int. Joint Conf. on Artificial Intelligence*, 2022.

[9] X. Liu, D. Chin, Y. Huang, and G. Xia, “Learning Interpretable Low-dimensional Representation via Physical Symmetry,” in *Conf. on Neural Information Processing Systems*, 2023.

[10] Y. Wu, Z. Wang, B. Raj, and G. Xia, “Emergent Interpretable Symbols and Content-Style Disentanglement via Variance-Invariance Constraints,” *arXiv preprint arXiv:2407.03824*, 2024.

[11] Y.-N. Hung, I.-T. Chiang, Y.-A. Chen, and Y.-H. Yang, “Musical Composition Style Transfer via Disentangled Timbre Representations,” in *Int. Joint Conf. on Artificial Intelligence*, 2019.

[12] F. Cwitkowitz, K. W. Cheuk, W. Choi, M. A. Martínez-Ramírez, K. Toyama, W.-H. Liao, and Y. Mitsufuji, “Timbre-trap: A Low-resource Framework for Instrument-agnostic Music Transcription,” in *Int. Conf. on Acoustics, Speech and Signal Processing*, 2024.

[13] K. W. Cheuk, K. Choi, Q. Kong, B. Li, M. Won, J.-C. Wang, Y.-N. Hung, and D. Herremans, “Jointist: Simultaneous Improvement of Multi-Instrument Transcription and Music Source Separation via Joint Training,” *arXiv preprint arXiv:2302.00286*, 2023.

[14] L. Lin, Q. Kong, J. Jiang, and G. Xia, “A Unified Model for Zero-shot Music Source Separation, Transcription and Synthesis,” in *Int. Soc. for Music Information Retrieval*, 2021.

[15] J. Gha, V. Herrmann, B. Grewe, J. Schmidhuber, and A. Gopalakrishnan, “Unsupervised Musical Object Discovery from Audio,” in *Conf. on Neural Information Processing Systems, ML4Audio Workshop*, 2023.

[16] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, “Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription,” in *Int. Conf. on Machine Learning*, 2012.

[17] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, “FiLM: Visual Reasoning With a General Conditioning Layer,” in *AAAI Conf. on Artificial Intelligence*, 2018.

[18] J. W. Kim, R. Bittner, A. Kumar, and J. P. Bello, “Neural Music Synthesis for Flexible Timbre Control,” in *Int. Conf. on Acoustics, Speech and Signal Processing*, 2018.

[19] H.-W. Dong and Y.-H. Yang, “Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation,” in *Int. Soc. for Music Information Retrieval*, 2018.

[20] J. H. Lee, H.-S. Choi, and K. Lee, “Audio Query-based Music Source Separation,” in *Int. Soc. for Music Information Retrieval*, 2019.

[21] Y. Wang, D. Stoller, R. Bittner, and J. Bello, “Few-Shot Musical Source Separation,” in *Int. Conf. on Acoustics, Speech and Signal Processing*, 2022.

[22] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow Twins: Self-Supervised Learning via Redundancy Reduction,” in *Int. Conf. on Machine Learning*, 2021.