Towards Robust Unsupervised Disentanglement of Sequential Data — A Case Study Using Music Audio

Yin-Jyun Luo[1*], **Sebastian Ewert**[2] and **Simon Dixon**[1]
[1]Centre for Digital Music, Queen Mary University of London
[2]Spotify
yin-jyun.luo@qmul.ac.uk, sewert@spotify.com, s.e.dixon@qmul.ac.uk

Abstract

Disentangled sequential autoencoders (DSAEs) represent a class of probabilistic graphical models that describes an observed sequence with dynamic latent variables and a static latent variable. The former encode information at a frame rate identical to the observation, while the latter globally governs the entire sequence. This introduces an inductive bias and facilitates unsupervised disentanglement of the underlying local and global factors. In this paper, we show that the vanilla DSAE suffers from being sensitive to the choice of model architecture and capacity of the dynamic latent variables, and is prone to collapse the static latent variable. As a countermeasure, we propose TS-DSAE, a two-stage training framework that first learns sequence-level prior distributions, which are subsequently employed to regularise the model and facilitate auxiliary objectives to promote disentanglement. The proposed framework is fully unsupervised and robust against the global factor collapse problem across a wide range of model configurations. It also avoids typical solutions such as adversarial training which usually involves laborious parameter tuning, and domain-specific data augmentation. We conduct quantitative and qualitative evaluations to demonstrate its robustness in terms of disentanglement on both artificial and real-world music audio datasets.

1 Introduction

From a probabilistic point of view, representation learning involves a data generating process governed by multiple explanatory factors of variation [Bengio, 2013]. The goal of learning a disentangled representation is to extract the underlying factors such that perturbations of one factor only change certain attributes of the observation. In this sense, disentangled representation promotes model interpretability by exposing semantically meaningful features, and enables controllable data generation by feature manipulation.

While supervised learning simplifies training processes, label scarcity for various problems of interest leads to a need for unsupervised techniques. However, as shown by Locatello *et al.* [2019], disentanglement can only be achieved with either supervision or inductive biases – and hence any unsupervised system for learning disentangled representations has to involve the latter. For sequential data, we can aim to disentangle global from local information by leveraging such a structural bias. In this case, the observation is generated by a static (*global*) latent variable associated with the entire sequence, and a series of dynamic (*local*) latent variables varying over time [Hsu *et al.*, 2017; Li and Mandt, 2018; Khurana *et al.*, 2019; Zhu *et al.*, 2020; Vowels *et al.*, 2021; Han *et al.*, 2021; Bai *et al.*, 2021].

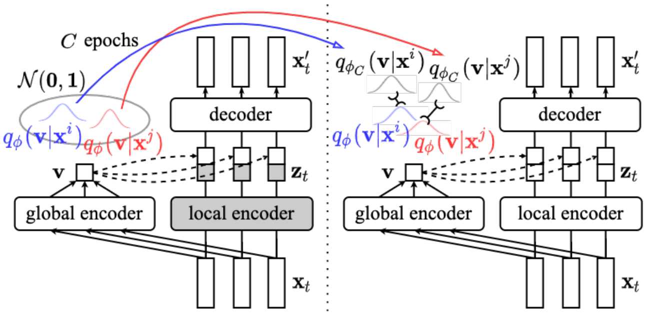
C epochs
$\mathcal{N}(\mathbf{0},\mathbf{1})$
$q_\phi(\mathbf{v}|\mathbf{x}^i)$
$q_\phi(\mathbf{v}|\mathbf{x}^j)$
$\mathbf{v}$
global encoder
local encoder
decoder
$\mathbf{z}_t$
$\mathbf{x}'_t$
$\mathbf{x}_t$
$q_{\phi_C}(\mathbf{v}|\mathbf{x}^i)$
$q_{\phi_C}(\mathbf{v}|\mathbf{x}^j)$
$q_\phi(\mathbf{v}|\mathbf{x}^i)$
$q_\phi(\mathbf{v}|\mathbf{x}^j)$
$\mathbf{v}$
global encoder
local encoder
decoder
$\mathbf{z}_t$
$\mathbf{x}'_t$
$\mathbf{x}_t$

Figure 1: System diagrams of Two-Stage DSAE. Left: The constrained training stage where the local modules are frozen. Right: The stage of informed-prior training where the global latent is regularised by the associated posterior learnt from the first stage. The dashed arrows denote broadcast along the time-axis.

The disentangled sequential autoencoder (DSAE) [Li and Mandt, 2018] is a minimalistic framework that implements the concept above using a probabilistic graphical model, as illustrated in Fig. 2. However, as we show in Section 6, DSAE does not robustly achieve disentanglement but heavily relies on a problem-specific architecture design and parameter tuning. Several works have built upon DSAE, extending it with either self-supervised learning techniques based on domain-specific data-augmentation [Bai *et al.*, 2021], alternative distance measures for the distributions involved which require extensive hyperparameter tuning or estimations susceptible to the instability resulting from adversarial training [Han *et al.*, 2021], or a rather complex parameterisation of a computationally heavy generative model [Vowels *et al.*, 2021].

In order to improve the robustness of DSAE, we propose

[*]Contact Author

TS-DSAE, a simple yet effective framework encompassing a two-stage training method as well as explicit regularisation to improve factor invariance and manifestation. The framework is completely unsupervised and free from any form of data augmentation or adversarial training (but could be combined with either in the future). We use an artificial as well as a real-world music audio dataset to verify the effectiveness of the proposed framework over a wide range of configurations, and provide both quantitative and qualitative evaluations. While the baseline models suffer from the collapse of the global latent space, TS-DSAE consistently provides reliable disentanglement (as measured by a classification metric), improves reconstruction quality with increased network capacity without compromising disentanglement, and is able to accommodate multiple global factors shared in the same latent space.

2 Disentangled Sequential Autoencoders

DSAEs [Li and Mandt, 2018; Zhu *et al.*, 2020; Bai *et al.*, 2021; Han *et al.*, 2021; Vowels *et al.*, 2021] are a family of probabilistic graphical models representing a joint distribution

$$p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathbf{v}) = p(\mathbf{v}) \prod_{t=1}^{T} p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{v}) p_\theta(\mathbf{z}_t|\mathbf{z}_{<t}), \quad (1)$$

where $\mathbf{x}_{1:T}$ denotes the observed sequence with T time frames, $\mathbf{z}_{1:T}$ is the sequence of local latent variables, and $\mathbf{v}$ refers to the global latent variable. In practice, $p_\theta(\mathbf{z}_t|\mathbf{z}_{<\mathbf{t}}) = \mathcal{N}\big(\mu_\theta(\mathbf{z}_{<t}), \mathrm{diag}(\sigma^2_\theta(\mathbf{z}_{<t}))\big)$ is parameterised by recurrent neural networks (RNNs), and $p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{v})$ is implemented using fully-connected networks (FCNs). The prior distribution of $\mathbf{v}$ follows $\mathcal{N}(\mathbf{0}, \mathbf{1})$. The model is trained to learn separate latent variables $\mathbf{z}_{1:T}$ and $\mathbf{v}$ for the local and global factors, respectively, imposing an inductive bias for unsupervised disentanglement, which is otherwise impossible [Locatello *et al.*, 2019]. The uni-modal prior $p(\mathbf{v})$, however, poses a great challenge to learning an informative latent space, evidenced by our results in Section 6.

Following the framework of variational autoencoders [Kingma and Welling, 2014], inference networks are introduced to optimise the evidence lower bound (ELBO):

$$\begin{aligned}
&\mathcal{L}(\theta, \phi; \mathbf{x}_{1:T}) \\
&= \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}, \mathbf{v}|\mathbf{x}_{1:T})}\Big[\log \frac{p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathbf{v})}{q_\phi(\mathbf{z}_{1:T}, \mathbf{v}|\mathbf{x}_{1:T})}\Big] \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}_{1:T}, \mathbf{v}) q_\phi(\mathbf{v}|\mathbf{x}_{1:T})}\big[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{v})\big] \\
&- \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{q_\phi(\mathbf{z}_{<t}|\mathbf{x}_{1:T}, \mathbf{v})}\big[\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_t|\mathbf{x}_{1:T}, \mathbf{v}) \| p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})\big)\big] \\
&- \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{v}|\mathbf{x}_{1:T}) \| p(\mathbf{v})\big).
\end{aligned} \quad (2)$$

We investigate the two configurations illustrated in Fig. 2. "full q" follows the inference networks written in Eq. (2), and $q_\phi(\mathbf{z}_t|\mathbf{x}_{1:T}, \mathbf{v})$ can be implemented via RNNs; while "factorised q" simplifies $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}, \mathbf{v}) = \prod_{t=1}^{T} q_\phi(\mathbf{z}_t|\mathbf{x}_t)$ with an FCN shared across the time-axis, which is independent of $\mathbf{v}$. In both cases, $q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$ can be parameterised by either RNNs or FCNs. We will use "factorised q" for the exposition in Section 3.

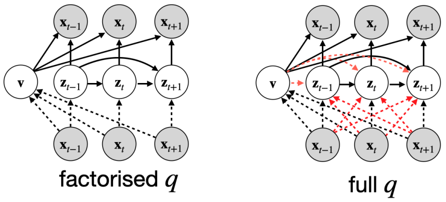

Figure 2: The two models proposed in the original DSAE. The red arrows highlight the enriched inference networks $q_\phi(\cdot)$.

A major challenge is that optimising Eq. (2) does not prevent the local latent $\mathbf{z}_{1:T}$ from capturing all the necessary information for reconstructing the given input sequence $\mathbf{x}_{1:T}$. This is referred to as the "shortcut problem" [Lezama, 2019], where the model completely ignores some latent variables (the global in this case) and only utilises the rest. In Section 6, we show that, without carefully tuning the hyperparameters, the vanilla DSAE is prone to only exploit $\mathbf{z}_{1:T}$ and ignore $\mathbf{v}$.

3 Method

We propose TS-DSAE, which constitutes a two-stage training framework and explicitly imposes regularisation for factor invariance as well as factor rendering in order to encourage disentanglement, as illustrated in Fig. 1 which depicts the simplified inference network (factorised q) to avoid clutter.

3.1 Two-Stage Training Framework

The shortcut problem mentioned in Section 2 can be ascribed to the simplicity of the uni-modal prior $p(\mathbf{v})$ which is not expressive enough to capture the multi-modal global factors, i.e. $q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$ is over-regularised. The issue is further exaggerated by the relatively capacity-rich local latent $\mathbf{z}_{1:T}$ which are allowed to carry information at the frame resolution identical to $\mathbf{x}_{1:T}$. To mitigate the problem, we divide the training into two stages, *constrained training* and *informed-prior training*.

Constrained training. During constrained training, we freeze some parameters of the local module after initialization including the local encoder and the transition network. This way, the local latents $\mathbf{z}_t$ resemble random projections from the input and thus are not optimised to hold the most important information to encode the input. That means, we strongly encourage the decoder to focus on the global latent $\mathbf{v}$ for reconstruction. As a result, $q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$ is biased to capture the global factors that are shared across the entire sequence. From an optimisation perspective, this is equivalent to eliminating the second term (the KL terms for $\mathbf{z}_t$) from Eq. (2).

Informed-prior training. The training proceeds to the second stage after C epochs of constrained training. During this stage, all the model parameters are unfrozen and trained regularly using the full objective (Eq. (2)) with a modification. In particular, instead of setting the global prior to $\mathcal{N}(\mathbf{0}, \mathbf{1})$ as in constrained training, we set:

$$p(\mathbf{v}^i) = q_{\phi_C}(\mathbf{v}^i|\mathbf{x}^i_{1:T}), \quad (3)$$

where ϕ_C denotes the parameters of the global encoder at the C-th epoch. That is, we have for *each* input sequence i a corresponding sequence-level prior that has been learnt from constrained training, whereby the last KL term in Eq. (2) is replaced by $\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{v}^i|\mathbf{x}^i_{1:T})\|q_{\phi_C}(\mathbf{v}^i|\mathbf{x}^i_{1:T})\big)$. Note that we differentiate q_ϕ from q_{ϕ_C} to emphasise that we take a "snapshot" of the global encoder $q_{\phi_C}(\cdot)$ at the C-th epoch, use the network to parameterise the sequence-specific prior, and continue training the global encoder $q_\phi(\cdot)$ which is initialised by ϕ_C. In other words, we keep training the posterior but "anchor" the distribution of each sequence i to its associated prior which is the posterior obtained from constrained training and is supposed to capture the sequence-level global factors.

This way, although the local module is introduced over the training, the global latent variables of sequences no longer commonly share the uni-modal prior, thereby mitigating the effect of over-regularisation.

In the next section, we further propose four additional loss terms to encourage disentanglement of the global and local latent variables.

3.2 Factor Invariance and Manifestation

Consider the following scheme of inference, replacement, decoding, and inference: given the *inferred* variables $\mathbf{z}^i_{1:T} \sim q_\phi(\mathbf{z}_{1:T}|\mathbf{x}^i_{1:T})$ and $\mathbf{v}^i \sim q_\phi(\mathbf{v}|\mathbf{x}^i_{1:T})$, we can *replace* $\mathbf{v}^i$ with $\mathbf{v}^j$ inferred from another sequence j, and *decode* $\mathbf{x}^{\mathbf{v}^i\to\mathbf{v}^j}_{1:T} \sim p_\theta(\mathbf{x}_{1:T}|\mathbf{z}^i_{1:T}, \mathbf{v}^j)$. We can then *infer* $\mathbf{z}^{\mathbf{v}^i\to\mathbf{v}^j}_{1:T} \sim q_\phi(\mathbf{z}_{1:T}|\mathbf{x}^{\mathbf{v}^i\to\mathbf{v}^j}_{1:T})$ and $\mathbf{v}^{\mathbf{v}^i\to\mathbf{v}^j} \sim q_\phi(\mathbf{v}|\mathbf{x}^{\mathbf{v}^i\to\mathbf{v}^j}_{1:T})$.

If $\mathbf{z}_{1:T}$ and $\mathbf{v}$ have been successfully disentangled, the difference between $\mathbf{z}^{\mathbf{v}^i\to\mathbf{v}^j}_{1:T}$ and $\mathbf{z}^i_{1:T}$ would be minimal because replacing the global factor should not affect the subsequently inferred local factor; and $\mathbf{v}^{\mathbf{v}^i\to\mathbf{v}^j}$ should be close to $\mathbf{v}^j$ in order to faithfully manifest the swapping. Similarly, if we replace $\mathbf{z}_{1:T}$ instead, difference between $\mathbf{v}^{\mathbf{z}^i_{1:T}\to\mathbf{z}^j_{1:T}}$ and $\mathbf{v}^i$ is expected to be small; and $\mathbf{z}^{\mathbf{z}^i_{1:T}\to\mathbf{z}^j_{1:T}}_{1:T}$ should be close to $\mathbf{z}^j_{1:T}$.

We can impose the desired properties of factor invariance as well as the rendering of the target factors by introducing the following terms to Eq. (2):

$$-\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{v}|\mathbf{x}^{\mathbf{v}^i\to\mathbf{v}^j}_{1:T})\|q_\phi(\mathbf{v}|\mathbf{x}^j_{1:T})\big), \tag{4}$$

$$-\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{1:T}|\mathbf{x}^{\mathbf{v}^i\to\mathbf{v}^j}_{1:T})\|q_\phi(\mathbf{z}_{1:T}|\mathbf{x}^i_{1:T})\big), \tag{5}$$

$$-\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{v}|\mathbf{x}^{\mathbf{z}^i_{1:T}\to\mathbf{z}^j_{1:T}}_{1:T})\|q_\phi(\mathbf{v}|\mathbf{x}^i_{1:T})\big), \text{and} \tag{6}$$

$$-\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{1:T}|\mathbf{x}^{\mathbf{z}^i_{1:T}\to\mathbf{z}^j_{1:T}}_{1:T})\|q_\phi(\mathbf{z}_{1:T}|\mathbf{x}^j_{1:T})\big). \tag{7}$$

By maximising these terms, we encourage invariance of the local and global latent variables through Eq. (5) and Eq. (6), respectively. Meanwhile, posteriors of the replaced factors are regularised to follow the target posteriors through Eq. (4) and Eq. (7).

In practice, we pair each input sequence i in a mini-batch with a randomly sampled input sequence j from the same mini-batch, and perform the above-mentioned scheme of inference, replacement, decoding, and inference. Note that we do not require any form of supervision or data-augmentation. While the above terms encourage meaningful behaviour, they can still be minimised with a trivial global latent space, which is undesired. Thus, the two-stage training plays a crucial role in obtaining robust disentanglement. Further, note that the individual terms above vary in terms of magnitude and thus importance to the gradient and so could benefit from balancing. However, we found scaling them unnecessary for the success of disentanglement, and leave this study for future work.

To summarise, TS-DSAE constitutes a two-stage training framework that facilitates the exploitation of additional divergences to achieve robust unsupervised disentanglement, which we empirically verify in Section 6.

4 Related Work

The assumption of a sequence being generated by a stationary global factor and a temporally changing local factor to achieve unsupervised disentanglement was used before. FHVAE [Hsu *et al.*, 2017] constructs a hierarchical prior where each input is governed by a sequence-level prior on top of a segment-level prior. Our two-stage training framework shares the spirit, with the main difference being that we leverage the strong bottleneck during the constrained training to naturally promote a global information-rich posterior which can be directly used as the sequence-level prior for the complete model training stage. On the other hand, FHVAE initialises and learns the prior from scratch, which lacks a stronger inductive bias and a discriminative objective function is reported to be helpful. Also, learning of the sequence-level priors is amortised by the global encoder in our model, whereby memory consumption does not scale with the number of training data as in FHVAE.

The vanilla DSAE [Li and Mandt, 2018] is proposed as an elegant minimalistic model to achieve disentanglement, as shown in Fig. 2. However, we demonstrate its tendency to collapse the global latent space in Section 6, which is likely due to the over-simplified standard Gaussian prior. R-WAE [Han *et al.*, 2021] minimises the Wasserstein distance between the aggregated posterior and the prior instead, estimated by maximum mean discrepancy or generative adversarial networks, either of which is not trivial in terms of parameter tuning and optimisation. S3-VAE [Zhu *et al.*, 2020] and C-DSVAE [Bai *et al.*, 2021] exploit self-supervised learning and employ either domain-specific ad-hoc loss functions or data augmentation. The proposed TS-DSAE is free from any form of supervision, adversarial training, or domain-dependent data augmentation.

VDSM adopts a pre-training stage as well as a scheme of KL-annealing to promote usage of the global latent space [Vowels *et al.*, 2021], which is similar to our constrained training. The main differences, however, are that we train only the global variable during "pre-trainig", and avoid KL-annealing to save the tuning efforts. Further, VDSM employs n decoders, each of which is responsible for a unique identity of a video object, where n is set manually depending on the dataset. This makes it less general, requires rather heavy computation, and might complicate the optimisation process. Lezama [2019] proposes a progressive autoencoder-based framework to tackle the "shortcut problem" for static data. The framework first trains a network with a low capacity latent space in order to learn the factors of interest, and subsequently increases the latent space capacity to improve data reconstruction. The final model utilises supervision from human annotations to learn the factors of interest. Our two-stage training shares a similar idea, but differs in that TS-DSAE operates without any supervision and models sequential data.

Our constrained training stage is also reminiscent of multi-view representation learning. For example, VCCA [Wang *et al.*, 2016] formulates a model that samples different views of a common object from distributions conditioned on a shared latent variable. NestedVAE [Vowels *et al.*, 2020] learns the common factors using staged information bottlenecks by training a low-level VAE given the latent space derived from a high-level VAE. In our model, given an input sequence, we treat multiple time frames as the different "views" of a common underlying factor which is the global factor.

There has been a lack of exploration in unsupervised disentangled representation for music audio. Both Luo *et al.* [2020] and Cífka *et al.* [2021] exploit self-supervised learning to decorrelate instrument pitch and timbre. Similar to our work, the latter models monophonic melodies. Yet, it employs pitch-shifting which is domain-dependent, and constrains the local capacity by learning discrete latent variables which might pose optimisation challenges. We maintain the simplicity of DSAE and improve the robustness in a simple yet effective way, which is not limited to any certain modality.

5 Experimental Setup

5.1 Datasets

We consider both an artificial and a real-world music audio dataset. The former facilitates the control over the underlying factors of variation, while the latter demonstrates applicability of the proposed model to realistic data.

dMelodies. The artificial dataset is compiled by synthesising audio from monophonic symbolic music gathered from dMelodies [Pati *et al.*, 2020]. Each melody is a two-bar sequence with 16 eighth notes, subject to several global factors, i.e., tonic, scale, and octave, and local factors, i.e., direction of arpeggiation, and rhythm. In order to facilitate analysis, we normalise the global factors by considering only the melodies of C Major in the fourth octave. We also discard melodies starting or ending with the rest note to avoid spurious amplitude values and boundaries during audio synthesis with FluidSynth.[1] We randomly pick 3k samples from the remaining melodies which are then split into 80% training and 20% validation sets, and synthesise audio of sampling rate 16kHz using sound fonts of violin and trumpet from MuseScore_General.sf3.[2] The amplitude of each audio sample is normalised with respect to its maximum value. The number of samples rendered with the two instruments is uniformly distributed.

URMP. For the real-world audio recordings, we select the violin and trumpet tracks from the URMP dataset [Li *et al.*, 2019]. We follow the preprocessing by Hayes *et al.* [2021], where the amplitude of each audio recording, resampled to 16kHz, is normalised in a corpus-wide fashion for each instrument subset. The audio samples are then divided into four-second segments, and segments with mean pitch confidence lower than 0.85 are discarded, as assessed by the full CREPE model [Kim *et al.*, 2018], a state-of-the-art pitch extractor. The process results in 1,545 violin and 534 trumpet samples in the training set, and 193 violin and 67 trumpet samples for validation.

Note that for both datasets, we expect the underlying local and global factors to be melody and instrument identity, respectively. We transform the audio samples and represent the data as log-amplitude mel-spectrogram with 80 mel filter banks, derived from a short-time Fourier transform with a 128ms Hann window and 16ms hop, leading to $\mathbf{x}_{1:T} \in \mathbb{R}^{80\times 251}$.

5.2 Implementation

Architecture. We study the two models proposed in the original DSAE [Li and Mandt, 2018], "factorised q" and "full q" as shown in Fig. 2. We use `net-[layers]` to denote architectures of modules, where `net` indicates types of the network, and `[layers]` is a list specifying the numbers of neurons at each layer. `Tanh` is used as the non-linear activation between layers of FCNs, and we use long short-term memory (LSTM) for RNNs. If a Gaussian parameterisation layer follows, we append the notation `Gau-L` which encompasses two linear layers with parameters $\mathbf{w}_1$ and $\mathbf{w}_2$ projecting the output hidden states $\mathbf{h}$ to $\mu_{\mathbf{w}_1}(\mathbf{h}) \in \mathbb{R}^L$ and $\log \sigma^2_{\mathbf{w}_2}(\mathbf{h}) \in \mathbb{R}^L$, respectively, where the Gaussian variable living in an L-dimensional space is then sampled from $\mathcal{N}\big(\mu_{\mathbf{w}_1}(\mathbf{h}), \mathrm{diag}(\sigma^2_{\mathbf{w}_2}(\mathbf{h}))\big)$.

For factorised q, we implement the global encoder $q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$ as `FCN-[64,64]-Avg-Gau-16`, where `Avg` denotes average pooling across the time-axis, and we keep the size of $\mathbf{v}$ fixed as 16 across our main experiments; and the local encoder $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ as `FCN-[64,64]-Gau-{8,16,32}`, where we investigate different sizes of $\mathbf{z}_{1:T}$. For the transition network $p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})$, we use `RNN-[32,32]-Gau-{8,16,32}`. The decoder $p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{v})$ is `FCN-[64,64]-Gau-80` taking as input the concatenation of $\mathbf{z}_{1:T}$ and time-axis broadcast $\mathbf{v}$. Note that, following the convention of VAEs, the Gaussian layer of the decoder parameterises $\mathcal{N}\big(\mu_{\mathbf{w}_1}(\mathbf{h}), \mathbf{1}\big)$ which evaluates the likelihood $p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{v})$ as the squared L2-norm between the output of the decoder and $\mathbf{x}_t$.

For full q, $q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$ follows that of factorised q; and $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}, \mathbf{v})$ corresponds to `biRNN-[64,64]-Gau-{8,16,32}` which takes as input the concatenation of $\mathbf{x}_{1:T}$ and time-axis broadcast $\mathbf{v}$ inferred from $q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$. `biRNN` denotes a bi-LSTM, where the outputs of the forward and backward LSTM are averaged along the time-axis before the Gaussian layer. Both the transition network and decoder follow those of factorised q.

Optimisation. Our implementation is based on `PyTorch v1.9.0` and we use ADAM [Kingma and Ba, 2014] with default parameters $lr = 0.001$, and $[\beta_1, \beta_2] = [0.9, 0.999]$ without weight decay. We use a batch size of 128, and train the models for 4k epochs at most; we employ early stopping if Eq. (2) obtained from the validation set stops improving for 300 epochs. For the models adopting the proposed two-stage training frameworks presented in Section 3, we set the number of epochs for the first stage $C = 300$ for all cases, to which we find the performance insensitive.

[1] https://www.fluidsynth.org/

[2] https://musescore.org/en/handbook

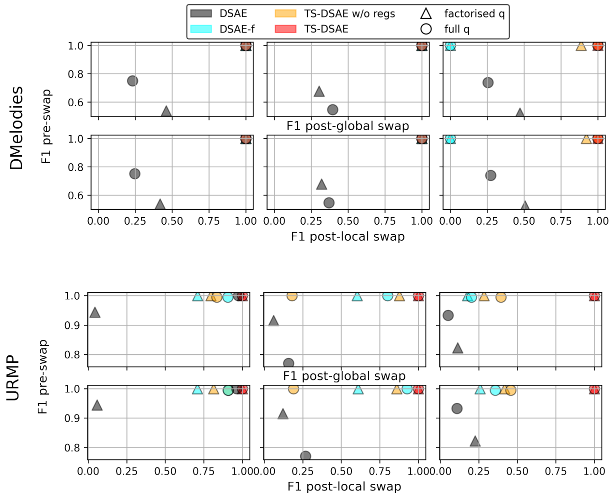

Figure 3: Macro F1 score of instrument classification derived from applying LDA to the global latent space. Size of the local latent space increases from left to right columns, 8, 16, and 32, respectively. See Section 6.1 for details.

6 Experiments and Results

We consider three baseline methods: 1) *DSAE*; 2) *DSAE-f*, where we employ the constrained training and freeze the global encoder after C epochs; and 3) *TS-DSAE w/o regs*, where we adopt the two-stage training framework without introducing the four terms from Section 3.2. We do not include the models mentioned in Section 4 [Zhu *et al.*, 2020; Bai *et al.*, 2021; Han *et al.*, 2021] which is left for future work, because the main focus is to improve upon DSAE with minimum modifications, and thus provide a superior backbone model which can be complementary with the existing methods.

6.1 Instrument Classification

We first evaluate disentanglement through the lens of instrument classification. In particular, we train a linear discriminant analysis (LDA) classifier taking as inputs $\mathbf{v} \sim q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$, the global latent variables sampled from a learnt model, derived from the training set, and evaluate its classification accuracy for instrument identity in terms of the macro F1-score on the validation set. We pair each sequence i from the validation set with another sequence j recorded with the other instrument, and perform the scheme of inference, replacement, decoding, and inference. Following the notation in Section 3.2, $\mathbf{v}^{\mathbf{v}^i \to \mathbf{v}^j}$ should be predictive of the instrument of sample j; while $\mathbf{v}^{\mathbf{z}^i_{1:T} \to \mathbf{z}^j_{1:T}}$ should reflect the original instrument of sample i. We report three metrics including accuracy before the replacement (pre-swap), after replacing $\mathbf{v}$ (post-global swap), and after replacing $\mathbf{z}_{1:T}$ (post-local swap). Note that we use the mean parameters of the Gaussian posterior $q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$ to train the LDA.

The results are summarised in Fig. 3. The proposed TS-DSAE (red), with either factorised or full q, is consistently located at the top right corner of the plot, across all the sizes of the local latent space. This indicates its robust disentanglement as well as a linearly separable global latent space. From the left to right column, the competing methods DSAE-f (cyan) and TS-DSAE without the additional regularisations (orange) move from top right to left-hand side of the plot, showing the inclination for a collapsed global latent space with the increased local latent capacity. Being located at lower left of the plot, DSAE (gray) attains the worst performance in most configurations. This highlights the issue of positing the standard Gaussian prior in the global latent space.

The overall high pre-swap and low post-swap F1 especially towards high-dimensional $\mathbf{z}_t$ implies that the decoder tends to ignore $\mathbf{v}$, even though the mean parameter of $q_\phi(\mathbf{v}|\mathbf{x}_{1:T})$ is discriminative w.r.t. the instrument identity. The competing models appear to suffer the most from the size of $\mathbf{z}_t$ as a large local latent space can easily capture all the necessary information for reconstruction.

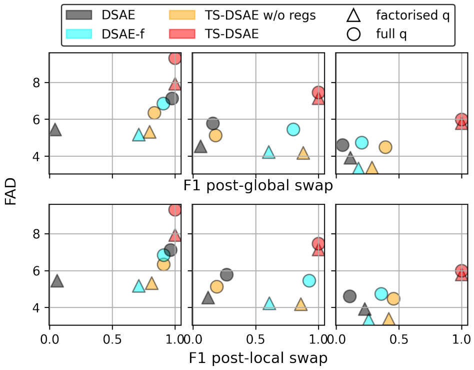

Figure 4: FAD (the lower the better) of reconstruction versus macro F1 score for instrument classification, evaluated using URMP. See Section 6.2 for details.

6.2 Reconstruction Quality

We examine the trade-off between disentanglement and reconstruction in terms of Fréchet Audio Distance (FAD) [Kilgour *et al.*, 2019] which is reported to correlate with auditory perception. We only report the results for URMP in Fig. 4 as both datasets reach a similar summary. As expected, FAD is improved with increasing $\mathbf{z}_t$ dimension. However, TS-DSAE is the only model that overcomes the trade-off, in the sense that competing models lose their ability to disentangle (move from right to left of the plot) with the improved FAD.

6.3 Raw Pitch Accuracy

In this section, we evaluate $\mathbf{z}_{1:T}$ by applying the full CREPE model [Kim *et al.*, 2018] to audio re-synthesised from the mel-spectrogram. The conversion is done by `InverseMelScale` and `GriffinLim` accessible from `torchaudio v0.9.0`. Using the notation from Section 3.2, we extract pitch contours from reconstructed samples (pre-swap), $\mathbf{x}_{1:T}^{\mathbf{v}^i \to \mathbf{v}^j}$ (post-global swap) which is supposed to mirror the pitch contour of $\mathbf{x}^i_{1:T}$, and $\mathbf{x}_{1:T}^{\mathbf{z}^i_{1:T} \to \mathbf{z}^j_{1:T}}$ (post-local swap) which is supposed to follow the pitch contour of $\mathbf{x}^j_{1:T}$. Note that for models with trivial $\mathbf{v}$, the accuracy of post-global swap will remain high as the decoder is independent of $\mathbf{v}$. We extract pitch contours from the input data as the ground-truth and report the raw pitch accuracy (RPA) with a 50-cent threshold [Salamon *et al.*, 2014].

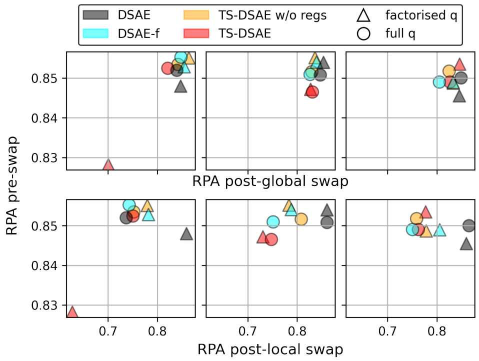

Figure 5: RPA assessed using CREPE on URMP.

We report the results with URMP in Fig. 5. TS-DSAE consistently improves with the increasing size of $\mathbf{z}_t$ in terms of RPA. Except for post-local swap, TS-DSAE performs comparably with the competing models towards the larger $\mathbf{z}_t$, and achieves disentanglement at once.

6.4 Richer Decoders

To mitigate the trade-off, we further construct and evaluate a richer decoder where the reconstruction of $\mathbf{x}_t$ is conditioned on $\mathbf{z}_{1:T}$, i.e., the entire sequence of local latent variables, instead of $\mathbf{z}_t$. We set the size of $\mathbf{z}_t$ to 16, and the inference network to factorised q, and compare DSAE, TS-DSAE, and the TS-DSAE augmented with the enriched decoder.

As shown in Fig. 6, the enriched model maintains the perfect accuracy for instrument classification for both datasets, while improving FAD over its counterpart with the factorised decoder. Note that using dMelodies, the model outperforms DSAE equipped with the factorised decoder in terms of FAD.

We leave the evaluation for the full range of configurations for future work, including autoregressive decoders that could cause posterior collapse even for vanilla VAEs.

6.5 Multiple Global Factors

We now consider both the fourth and fifth octaves when synthesising the dMelodies dataset, introducing octave number as the other global factor of variation in addition to instrument identity. We train the decoder-enriched TS-DSAE described in Section 6.4, and show the results in Fig. 7. In particular, we replace $\mathbf{v}$ inferred from the source at the lower left, with that derived from one of the three targets displayed in the top row, and generate novel samples shown from the second to last columns of the bottom row.

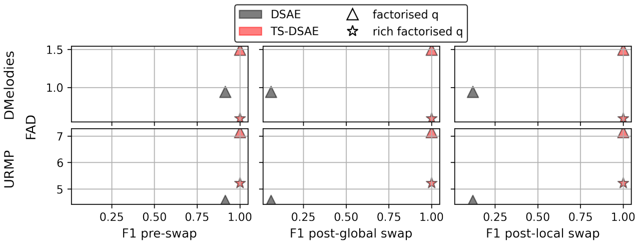

Figure 6: FAD (the lower the better) against disentanglement in terms of instrument classification.

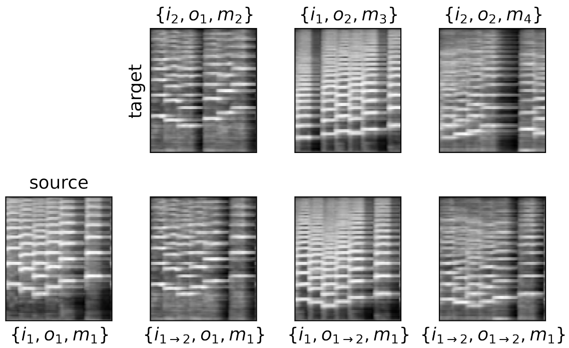

Figure 7: Global latent replacement using the top three samples as the targets and the sample at the bottom left as the source.

We use $\{i, o, m\}$ to denote the instrument, octave, and melody of each sample, respectively. For example, the source $\{i_1, o_1, m_1\}$ and the first target $\{i_2, o_1, m_2\}$ share the same octave but differ in the instrument, characterised by the spectral distribution along the frequency axis. As a result of replacing $\mathbf{v}$, the target instrument i_2 is manifested in the outcome $\{i_{1\to2}, o_1, m_1\}$, while the octave remains unchanged. Similarly, the second target $\{i_1, o_2, m_3\}$ differs from the source with the octave, characterised by the level of pitch contour; therefore, swapping $\mathbf{v}$ only transforms the octave for the output $\{i_1, o_{1\to2}, m_1\}$. Finally, the sample $\{i_{1\to2}, o_{1\to2}, m_1\}$ results from using the target $\{i_2, o_2, m_4\}$ that does not share any of the attributes with the source, where both the instrument and octave are converted. Importantly, the source melody m_1 remains intact in the three transformed samples, suggesting the global-local disentanglement.

7 Conclusion and Future Work

We have proposed TS-DSAE, a robust framework for unsupervised sequential data disentanglement, which has been shown to consistently work over a wide range of settings. [3] Our evaluation focuses on the ability to robustly achieve disentanglement, and we leave evaluations on multi-modal data generation from unconditional prior sampling for future work. We would also like to verify the applicability of TS-DSAE to modalities beyond the music audio datasets.

Despite the drastic increase in robustness, the difficulty of balancing disentanglement and reconstruction remains challenging [Lezama, 2019]. Scaling the regularisation terms differently might be helpful as mentioned in Section 3.2. Moreover, DSAEs probabilistic graphical model forces the input sequence to have a single global latent variable fixed over time, which could be too restrictive for more general use cases where sequences do not have stationary factors but ones that evolve slowly. Therefore, adopting a hierarchy of latent variables encoding information at multiple frame rates [Saxena *et al.*, 2021] can be a favorable relaxation of DSAEs. A potential extension of our two-stage training is to have multiple stages of constrained training with progressively larger network capacity, thereby accommodating the said hierarchy, which can also be seen as a temporal extension of Li *et al.* [2020].

[3] The implementation and audio samples are accessible from https://github.com/yjlolo/dSEQ-VAE.

Acknowledgments

The first author is a research student at the UKRI CDT in AI and Music, supported by Spotify.

References

[Bai *et al.*, 2021] Junwen Bai, Weiran Wang, and Carla Gomes. Contrastively disentangled sequential variational autoencoder. *Advances in Neural Information Processing Systems*, 2021.

[Bengio, 2013] Yoshua Bengio. Deep learning of representations: Looking forward. In *Proceedings of the International Conference on Statistical Language and Speech Processing*, 2013.

[Cífka *et al.*, 2021] Ondřej Cífka, Alexey Ozerov, Umut Şimşekli, and Gaël Richard. Self-supervised VQ-VAE for one-shot music style transfer. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2021.

[Han *et al.*, 2021] Jun Han, Martin Renqiang Min, Ligong Han, Li Erran Li, and Xuan Zhang. Disentangled recurrent Wasserstein autoencoder. In *Proceedings of the International Conference on Learning Representations*, 2021.

[Hayes *et al.*, 2021] Ben Hayes, Charalampos Saitis, and György Fazekas. Neural waveshaping synthesis. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2021.

[Hsu *et al.*, 2017] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. *Advances in Neural Information Processing Systems*, 2017.

[Khurana *et al.*, 2019] Sameer Khurana, Shafiq Rayhan Joty, Ahmed Ali, and James Glass. A factorial deep Markov model for unsupervised disentangled representation learning from speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2019.

[Kilgour *et al.*, 2019] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of INTERSPEECH*, 2019.

[Kim *et al.*, 2018] Jong Wook Kim, Justin Salamon, Peter Qi Li, and Juan Pablo Bello. CREPE: A convolutional representation for pitch estimation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2018.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.

[Lezama, 2019] José Lezama. Overcoming the disentanglement vs reconstruction trade-off via Jacobian supervision. In *Proceedings of the International Conference on Learning Representations*, 2019.

[Li and Mandt, 2018] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. In *Proceedings of the International Conference on Machine Learning*, 2018.

[Li *et al.*, 2019] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 2019.

[Li *et al.*, 2020] Zhiyuan Li, Jaideep Vitthal Murkute, Prashnna Kumar Gyawali, and Linwei Wang. Progressive learning and disentanglement of hierarchical representations. In *Proceedings of the International Conference on Learning Representations*, 2020.

[Locatello *et al.*, 2019] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning*, 2019.

[Luo *et al.*, 2020] Yin-Jyun Luo, Kin Wai Cheuk, Tomoyasu Nakano, Masataka Goto, and Dorien Herremans. Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2020.

[Pati *et al.*, 2020] Ashis Pati, Siddharth Gururani, and Alexander Lerch. dMelodies: A music dataset for disentanglement learning. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2020.

[Salamon *et al.*, 2014] Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, and Gaël Richard. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 2014.

[Saxena *et al.*, 2021] Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. *Advances in Neural Information Processing Systems*, 2021.

[Vowels *et al.*, 2020] Matthew James Vowels, Necati Cihan Camgöz, and Richard Bowden. NestedVAE: Isolating common factors via weak supervision. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2020.

[Vowels *et al.*, 2021] Matthew James Vowels, Necati Cihan Camgoz, and Richard Bowden. VDSM: Unsupervised video disentanglement with state-space modeling and deep mixtures of experts. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2021.

[Wang *et al.*, 2016] Weiran Wang, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arxiv:1610.03454*, 2016.

[Zhu *et al.*, 2020] Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3VAE: Self-supervised sequential VAE for representation disentanglement and data generation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2020.