UNSUPERVISED PITCH-TIMBRE DISENTANGLEMENT OF MUSICAL INSTRUMENTS USING A JACOBIAN DISENTANGLED SEQUENTIAL AUTOENCODER

Yin-Jyun Luo[1]*, Sebastian Ewert*[2]*, Simon Dixon*[1]

[1]Centre for Digital Music, Queen Mary University of London [2]Spotify

ABSTRACT

Disentangled representation learning seeks to align individual dimensions or separate groups of coordinates of latent factors with attributes of observed data such that perturbing certain latent factors uniquely changes particular attributes. A main challenge in unsupervised disentanglement using autoencoders is that strong regularisation, while necessary for consistent disentanglement, comes at the expense of accurate data reconstruction. To address this, we introduce a teacher-student framework that incorporates a variational sequential autoencoder and a Jacobian constraint that regularises the variation of observations relative to latent factors. In real-world audio recordings of musical instruments, our approach outperforms a state-of-the-art method in both sampling quality and unsupervised pitch-timbre disentanglement.

Index Terms— Disentangled representation, unsupervised learning, variational autoencoder, music instrument

1. INTRODUCTION

Disentangled representation learning captures factorised latent variables corresponding to distinct factors of variation (FoV) [1] and is applied in various modalities [2–8]. To expose semantically meaningful features of sequential data, such as speaker identity and linguistic content from speech, disentangled sequential autoencoders (DSAEs) [9–17] admit a model $p_\theta(\mathbf{x}_{1:\tau}|\mathbf{z}_{1:\tau}, \mathbf{v})p_\theta(\mathbf{z}_{1:\tau})p_\theta(\mathbf{v})$, where $\mathbf{x}_{1:\tau} \in \mathbb{R}^{\tau D_x}$ is conditionally sampled from a time-variant ("local") latent $\mathbf{z}_{1:\tau} \in \mathbb{R}^{\tau D_z}$ and a time-invariant ("global") latent $\mathbf{v} \in \mathbb{R}^{D_v}$. The model is optimised using an evidence lower bound (ELBO) similar to the variational autoencoder (VAE) [18]. DSAE [9] disentangles the global FoV (e.g. speaker identity) from the local ones (e.g. linguistic content) using model assumptions that encourage an inductive bias, which was shown to be required for disentanglement in unsupervised settings [19].

In this paper, we tackle unsupervised learning of disentangled pitch and timbre representations of monophonic musical instruments, represented by $\mathbf{z}_{1:\tau}$ and $\mathbf{v}$, respectively. In particular, our evaluation is performed using real violin and trumpet audio recordings from the URMP dataset [20].

The first author is a research student at the UKRI Centre for Doctoral Training in AI and Music, supported by a scholarship from Spotify.

In practice, DSAE tends to capture information using $\mathbf{z}_{1:\tau}$ and leaves $\mathbf{v}$ unused [14, 16], due to a capacity gap between the two latents caused by the difference in temporal resolution. Further, DSAE showed a considerable sensitivity to hyperparameters such as D_z and D_v [17]. TS-DSAE addresses the problem using a two-stage framework [17]. The first stage promotes learning $\mathbf{v}$, the timbre information, with a strong bottleneck that excludes $\mathbf{z}_{1:\tau}$, the pitch information. The informative $\mathbf{v}$ regularises the training of the full model during the second stage, along with other auxiliary constraints, to ensure unsupervised disentanglement of pitch and timbre. Despite its robustness to the hyperparameters, the auxiliary objectives can potentially over-regularise and harm reconstruction.

A similar teacher-student framework is proposed in [21] for images of human faces. First, a teacher model learns facial attributes supervised from annotated images such that perturbing a latent variable uniquely leads to variation of a certain facial attribute in the generated images. A Jacobian constraint quantifies the variation of an image w.r.t. the perturbation in the latent space and regularises a student network of larger capacity meant for refining reconstruction. It shows a better reconstruction quality compared to baselines, despite the fact that the teacher is supervised by labels.

To overcome the trade-off between reconstruction and unsupervised disentanglement, we combine the two frameworks and propose a Jacobian DSAE (J-DSAE), which significantly improves TS-DSAE in *both* reconstruction and timbre similarity measured in the Fréchet Audio Distance (FAD) [22], *and* disentanglement in terms of raw pitch accuracy (RPA) [23] of extracted pitch contours. [1]

2. BACKGROUND

2.1. Two-stage disentangled sequential autoencoder

The DSAE [9] models sequential data and learns both a global latent $\mathbf{v}$ and a local latent $\mathbf{z}_{1:\tau}$ by optimising the ELBO:

$$\begin{aligned}\mathcal{L}(\theta, \phi; \mathbf{x}_{1:\tau}) = \ & \mathbb{E}_{q_\phi(\mathbf{z}_{1:\tau}, \mathbf{v}|\cdot)}\big[\log p_\theta(\mathbf{x}_{1:\tau}|\mathbf{z}_{1:\tau}, \mathbf{v})\big] \\ -\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{1:\tau}|\cdot)\|p_\theta(\mathbf{z}_{1:\tau})\big) & - \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{v}|\cdot)\|p(\mathbf{v})\big),\end{aligned} \tag{1}$$

where the conditional variable $\mathbf{x}_{1:\tau}$ is omitted in q_ϕ for brevity. The posterior $q_\phi(\mathbf{z}_{1:\tau}, \mathbf{v}|\cdot)$ is a product of two diagonal Gaus-

[1]Audio samples are available at `http://www.jaco-dsae.xyz/`.

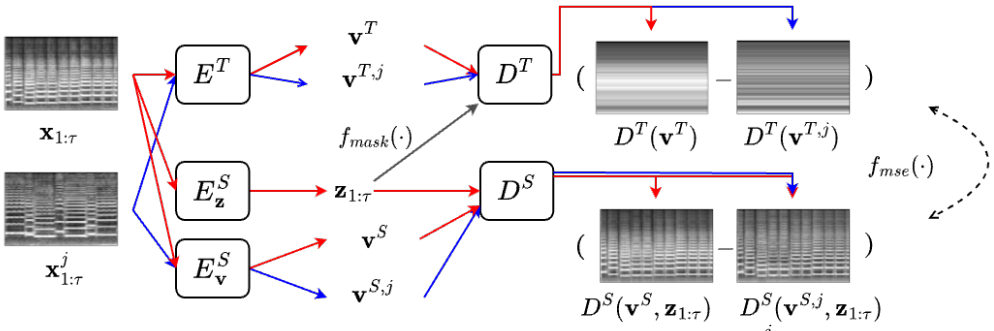

Fig. 1. Illustration of the Jacobian constraint (Eq. 8). In practice, the paired sequence $\mathbf{x}^j_{1:\tau}$ is randomly sampled and does not necessarily share or differ in particular attributes, thus **the framework is fully unsupervised**. E and D denote the encoder and decoder, and the superscripts T and S refer to the teacher and student modules. $\mathbf{v}$ and $\mathbf{z}_{1:\tau}$ are global and local latents.

sians parameterised by separate neural network encoders to disentangle $\mathbf{v}$ from $\mathbf{z}_{1:\tau}$. The dynamic prior $p_\theta(\mathbf{z}_{1:\tau})$ is also a product of diagonal Gaussians that can be implemented by an LSTM. We simply set $p(\mathbf{v}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Although the temporal resolution gap introduces a strong inductive bias for disentanglement, DSAE has been reported to collapse $\mathbf{v}$ and capture all information using $\mathbf{z}_{1:\tau}$ [14, 16, 17].

TS-DSAE [17] addresses the issue by not updating the local encoder $q_\phi(\mathbf{z}_{1:\tau}|\cdot)$ in the first stage, leading the model to use $\mathbf{v}$. Setting $\mathbf{z}_{1:\tau}$ of $p_\theta(\mathbf{x}_{1:\tau}|\mathbf{z}_{1:\tau}, \mathbf{v})$ to zero, as in our experiments, also achieves a similar result. During the second stage, the training recovers Eq. 1 with the non-informative prior $p(\mathbf{v})$ replaced by $q_{\phi_C}(\mathbf{v}|\cdot)$, where ϕ_C denotes the encoder at the end of the first stage, which helps preserve the learnt information.

Disentanglement can be further promoted by minimising:

$$\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{v}|\mathbf{x}^{\mathbf{v}\to\mathbf{v}^j}_{1:\tau})\|q_\phi(\mathbf{v}|\mathbf{x}^j_{1:\tau})\big), \tag{2}$$

$$\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{1:\tau}|\mathbf{x}^{\mathbf{v}\to\mathbf{v}^j}_{1:\tau})\|q_\phi(\mathbf{z}_{1:\tau}|\mathbf{x}_{1:\tau})\big), \tag{3}$$

$$\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{v}|\mathbf{x}^{\mathbf{z}_{1:\tau}\to\mathbf{z}^j_{1:\tau}}_{1:\tau})\|q_\phi(\mathbf{v}|\mathbf{x}_{1:\tau})\big), \text{and} \tag{4}$$

$$\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{1:\tau}|\mathbf{x}^{\mathbf{z}_{1:\tau}\to\mathbf{z}^j_{1:\tau}}_{1:\tau})\|q_\phi(\mathbf{z}_{1:\tau}|\mathbf{x}^j_{1:\tau})\big), \tag{5}$$

where $\mathbf{x}^{\mathbf{z}_{1:\tau}\to\mathbf{z}^j_{1:\tau}}_{1:\tau} \sim p_\theta(\mathbf{x}_{1:\tau}|\mathbf{z}^j_{1:\tau}, \mathbf{v})$, $\mathbf{z}^j_{1:\tau} \sim q_\phi(\mathbf{z}_{1:\tau}|\mathbf{x}^j_{1:\tau})$ and $\mathbf{v} \sim q_\phi(\mathbf{v}|\mathbf{x}_{1:\tau})$, similar for $\mathbf{v} \to \mathbf{v}^j$. That is, $\mathbf{x}^j_{1:\tau}$ is randomly sampled to pair with $\mathbf{x}_{1:\tau}$, and their latents are swapped and combined to condition the sampling of a synthesised sequence. The posterior over the swapped latent should match that of the paired sequence (Eqs. 2 and 5), while the posterior over the other latent should remain unchanged (Eqs. 3 and 4).

Although Eqs. 2 - 5 further encourage unsupervised disentanglement, they introduce unnecessarily strong constraints, which adversely affect sampling quality. This, in turn, has repercussions on reconstruction quality and timbre similarity to targets in attribute translation.

2.2. Jacobian constraint

In [21], a teacher-student framework is proposed to first capture the FoV of interest with a small latent space, which is later enlarged to recover the nuisance factors. Given images of human faces, a low-capacity deterministic AE, or the teacher, learns representations $\mathbf{y} \in \mathbb{R}^k$ specified by annotations of facial attributes. The student network is then built by adding $\mathbf{z} \in \mathbb{R}^d$, $d > k$, to the teacher, which helps to improve reconstruction. Importantly, the model is regularised so that:

$$\frac{\partial \hat{\mathbf{x}}^T}{\partial \mathbf{y}^T} = \frac{\partial \hat{\mathbf{x}}^S}{\partial \mathbf{y}^S}, \tag{6}$$

where $\hat{\mathbf{x}}$ and $\mathbf{y}$ are reconstructed images and latent variables of the teacher (superscript T) and the student (superscript S) network, respectively. That is, the Jacobian, or change in observation w.r.t. latent perturbation, is supposed to be consistent between the teacher and the student network. Eq. 6 can be translated into $\mathcal{L}_{Jac}$ that regularises the student[2]:

$$\begin{aligned}\mathcal{L}_{Jac}(E^S, D^S; \mathbf{x}) = \|\mathbf{y}^S - \mathbf{y}^T\|_2^2 \\ + \|(D^T(\mathbf{y}^{T,j}) - D^T(\mathbf{y}^T)) \\ - (D^S(\mathbf{y}^{S,j}, \mathbf{z}) - D^S(\mathbf{y}^S, \mathbf{z}))\|_2^2,\end{aligned} \tag{7}$$

where $\mathbf{y}^U = E^U(\mathbf{x})$ and (E^U, D^U) denotes the encoder-decoder pair. The superscript $U \in \{T, S\}$ denotes the teacher or student network, and j is the index of a randomly paired $\mathbf{x}^j$. Intuitively speaking, given a teacher network that has already learnt the factors of interest $\mathbf{y}^T$, the first term anchors $\mathbf{y}^S$ to $\mathbf{y}^T$ in order to retain the information learnt by the teacher, and the second term promotes the consistency of the variation of the outputs relative to that of $\mathbf{y}^U$ between the teacher and student networks. The framework outperforms an adversarial baseline in terms of reconstruction quality [21].

There are considerable differences between TS-DSAE and [21]: TS-DSAE is unsupervised, models sequential data and is based on a probabilistic model, in contrast to [21], which employs deterministic autoencoders and models annotated static images. However, we highlight their connections in Section 3 and introduce the Jacobian constraint into TS-DSAE, in order to combat the trade-off between reconstruction and unsupervised disentanglement.

[2]The derivation is detailed in Eq. 2 - 7 in [21].

3. METHOD

3.1. TS-DSAE as a teacher-student framework

Let the global encoder $E^T(\mathbf{x}_{1:\tau}) \coloneqq q_{\phi_C}(\mathbf{v}|\mathbf{x}_{1:\tau})$ and the decoder $D^T(\mathbf{v}) \coloneqq p_{\theta_C}(\mathbf{x}_{1:\tau}|\mathbf{0}, \mathbf{v})$, where ϕ_C and θ_C are the parameters in the last epoch C of the first stage. Similarly, $E^S_{\mathbf{v}}(\mathbf{x}_{1:\tau}) \coloneqq q_\phi(\mathbf{v}|\mathbf{x}_{1:\tau})$, $E^S_{\mathbf{z}}(\mathbf{x}_{1:\tau}) \coloneqq q_\phi(\mathbf{z}_{1:\tau}|\mathbf{x}_{1:\tau})$, and $D^S(\mathbf{v}, \mathbf{z}_{1:\tau}) \coloneqq p_\theta(\mathbf{x}_{1:\tau}|\mathbf{z}_{1:\tau}, \mathbf{v})$, are the two encoders and the decoder that continue training during the second stage.

The connection between TS-DSAE and a teacher-student framework becomes clearer with the above definitions: The first stage produces a teacher VAE (with $\mathbf{z}_{1:\tau}$ masked), which regularises the second stage training of a student DSAE through $\mathcal{D}_{\mathrm{KL}}\big(E^S_{\mathbf{v}}(\cdot)\|E^T(\cdot)\big)$.

Revealing the connection makes it straightforward to apply Eq. 7 to regularise the second stage training of TS-DSAE:

$$\begin{aligned}\mathcal{L}_{Jac}(E^S_{\mathbf{v}}, E^S_{\mathbf{z}}, D^S; \mathbf{x}_{1:\tau}) =& \mathcal{D}_{\mathrm{KL}}\big(E^S_{\mathbf{v}}(\mathbf{x}_{1:\tau})\|E^T(\mathbf{x}_{1:\tau})\big) \\ &+\|(D^T(\mathbf{v}^{T,j}) - D^T(\mathbf{v}^T)) \\ &-(D^S(\mathbf{v}^{S,j}, \mathbf{z}_{1:\tau}) - D^S(\mathbf{v}^S, \mathbf{z}_{1:\tau}))\|_2^2,\end{aligned} \tag{8}$$

where $\mathbf{v}^{T,j} \sim E^T(\mathbf{x}^j_{1:\tau})$, $\mathbf{v}^T \sim E^T(\mathbf{x}_{1:\tau})$, $\mathbf{v}^{S,j} \sim E^S_{\mathbf{v}}(\mathbf{x}^j_{1:\tau})$, $\mathbf{v}^S \sim E^S_{\mathbf{v}}(\mathbf{x}_{1:\tau})$, and $\mathbf{z}_{1:\tau} \sim E^S_{\mathbf{z}}(\mathbf{x}_{1:\tau})$. The teacher network (E^T, D^T) is frozen and is used to regularise the student network $(E^S_{\mathbf{v}}, E^S_{\mathbf{z}}, D^S)$. The probabilistic formulation replaces the mean squared error, or the first term of Eq. 7 with the distributional difference in Eq. 8. The second term is illustrated in Fig. 1: the Jacobian is calculated given a *random* pair of samples from a mini-batch and no explicit supervision is used. The variation in the observation domain caused by $\mathbf{v}^U$ and $\mathbf{v}^{U,j}$, where $U \in \{T, S\}$, is constrained to be consistent between the teacher and student networks.

3.2. Optimising J-DSAE

To summarise, J-DSAE first maximises $\mathcal{L}_T$ for C epochs:

$$\begin{aligned}\mathcal{L}_T(\theta, \phi; \mathbf{x}_{1:\tau}) = \mathbb{E}_{q_\phi(\mathbf{z}_{1:\tau}, \mathbf{v}|\cdot)}\big[\log p_\theta(\mathbf{x}_{1:\tau}|\mathbf{0}, \mathbf{v})\big] \\ -\mathcal{D}_{\mathrm{KL}}(\mathbf{z}_{1:\tau}) - \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{v}|\cdot)\|p(\mathbf{v})\big),\end{aligned} \tag{9}$$

and continues to maximise an alternative objective:

$$\begin{aligned}\mathcal{L}_S(\theta, \phi; \mathbf{x}_{1:\tau}) = \mathbb{E}_{q_\phi(\mathbf{z}_{1:\tau}, \mathbf{v}|\cdot)}\big[\log p_\theta(\mathbf{x}_{1:\tau}|\mathbf{z}_{1:\tau}, \mathbf{v})\big] \\ -\mathcal{D}_{\mathrm{KL}}(\mathbf{z}_{1:\tau}) - \mathcal{L}_{Jac}(E^S_{\mathbf{v}}, E^S_{\mathbf{z}}, D^S; \mathbf{x}_{1:\tau}),\end{aligned} \tag{10}$$

where $\mathcal{D}_{\mathrm{KL}}(\mathbf{z}_{1:\tau}) \coloneqq \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{1:\tau}|\cdot)\|p_\theta(\mathbf{z}_{1:\tau})\big)$.

Crucially, apart from $\mathcal{L}_{Jac}$, another difference between J-DSAE and TS-DSAE is that the former removes the loss terms of Eqs. 2 - 5, leading to superior sampling quality and disentanglement, as we show in Section 5.

Unlike in Section 2.2, J-DSAE adopts the Jacobian constraint for sequential data instead of images, leverages the inductive bias and training mechanism of TS-DSAE to be free of explicit forms of supervision, and optimises a VAE with prior distributions rather than a deterministic AE.

4. EXPERIMENT

4.1. Dataset

Following the TS-DSAE benchmark [17], we use a subset of the URMP dataset [20] that is composed of 1,545 (193) violin and 534 (67) trumpet training (validation) samples. Audio recordings are resampled to 16 kHz and divided into four-second chunks that are transformed into mel spectrograms of 80 filter banks, resulting in $\mathbf{x}_{1:\tau} \in \mathbb{R}^{80\times 251}$. We expect to capture the overall timbre of the instrument using $\mathbf{v}$ and the time-varying pitch using $\mathbf{z}_{1:\tau}$.

4.2. Implementation

We closely follow the architectural implementation of TS-DSAE [17] [3] and set (D_v, D_z) as $(16, 32)$. We implement both a factorised and an autoregressive (AR) decoder. $D^U_{\text{Fac}} \coloneqq \prod_{t=1}^{\tau} p_\theta(\mathbf{x}_t|\mathbf{v}, \mathbf{z}_t)$ and $D^U_{\text{AR}} \coloneqq \prod_{t=1}^{\tau} p_\theta(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{v}, \mathbf{z}_{<t})$. We use ADAM as the optimiser with a learning rate 10^{-3} and a batch size of 128, and set $C = 300$ epochs as in the original implementation.

4.3. Evaluation protocol

We train an instrument classifier (IC) using the training set. Then, given the validation set, the classifier predicts the instrument given $\mathbf{x}^{\mathbf{z}_{1:\tau}\to\mathbf{z}^j_{1:\tau}}_{1:\tau}$ and $\mathbf{x}^{\mathbf{v}\to\mathbf{v}^j}_{1:\tau}$ defined in Section 2.1. We denote the former by $\mathbf{z}$-*swap* and the latter by $\mathbf{v}$-*swap*. The ground truth instrument of $\mathbf{z}$-swap remains that of the input sample, while the ground truth as a result of $\mathbf{v}$-swap becomes the instrument label of the sample indexed j.

We use the full CREPE model [24] to extract pitch contours and evaluate them against target pitch contours in terms of RPA with a pitch tolerance of 50 cents [23]. The target of $\mathbf{z}$-swap is the pitch contour of the sample indexed j, while the target of $\mathbf{v}$-swap is dictated by the input sample.

For sampling quality, we measure FAD, which is reported to correlate with auditory perception [22], by comparing embeddings of actual and generated samples that we extract using the pre-trained IC, which captures timbre information. *Recon.* in Table 1 reports FAD by comparing the reconstruction and the actual data. FAD of $\mathbf{v}$-*swap* ($\mathbf{z}$-*swap*) compares $\mathbf{x}^{\mathbf{v}\to\mathbf{v}^j}_{1:\tau}$ ($\mathbf{x}^{\mathbf{z}_{1:\tau}\to\mathbf{z}^j_{1:\tau}}_{1:\tau}$) and the subset of true recordings played with the instrument of $\mathbf{x}^j_{1:\tau}$ ($\mathbf{x}_{1:\tau}$), thus measuring how well the timbre is translated (preserved) by its similarity to the true data. *Rand.* first samples $\mathbf{x}^{\text{rand}}_{1:\tau} \sim D^S(\mathbf{v}^{\text{rand}}, \mathbf{z}_{1:\tau})$, where $\mathbf{v}^{\text{rand}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{z}_{1:\tau} \sim E^S_{\mathbf{z}}(\mathbf{x}_{1:\tau})$, then predicts $\hat{y}_{\text{rand}}$ given $\mathbf{x}^{\text{rand}}_{1:\tau}$ using the pre-trained IC, and calculates FAD given $\mathbf{x}^{\text{rand}}_{1:\tau}$ and true samples whose ground truths are $\hat{y}_{\text{rand}}$. The rationale is that if $\mathbf{x}^{\text{rand}}_{1:\tau}$ carries a timbre of $\hat{y}_{\text{rand}}$ according to the IC, it is supposed to achieve a low FAD to true samples annotated with $\hat{y}_{\text{rand}}$.

[3] https://github.com/yjlolo/dSEQ-VAE

	Global (F1) ↑		Local (RPA) ↑		$\text{FAD}_{\text{Violin}}$ ↓				$\text{FAD}_{\text{Trumpet}}$ ↓			
	v-swap	**z**-swap	**v**-swap	**z**-swap	Recon.	**v**-swap	**z**-swap	Rand.	Recon.	**v**-swap	**z**-swap	Rand.
TS-DSAE	1.00	1.00	0.87	0.86	3.49	4.26 (0.15)	4.53 (0.09)	4.85 (0.09)	2.08	4.00 (0.48)	3.46 (0.36)	2.88 (0.02)
J-DSAE	1.00	1.00	**0.93**	**0.92**	**2.17**	**3.22** (0.06)	**3.37** (0.07)	**3.37** (0.02)	**1.20**	**2.73** (0.49)	**2.25** (0.27)	**2.12** (0.14)
TS-DSAE (AR)	1.00	1.00	0.82	0.83	2.18	3.19 (0.11)	3.47 (0.07)	3.56 (0.03)	1.41	3.61 (0.48)	2.92 (0.38)	2.59 (0.24)
J-DSAE (AR)	1.00	1.00	**0.90**	**0.90**	**1.49**	**2.35** (0.05)	**2.57** (0.03)	**2.56** (0.07)	**0.89**	**1.05** (0.12)	**0.99** (0.04)	**1.88** (0.09)

Table 1. Disentanglement in terms of macro F1 score and RPA, alongside audio quality measured in FAD. The numbers in parentheses are standard deviation of three random paired sequences (for **v**- and **z**-swap) or samples of **v** (for Rand.).

5. RESULTS

5.1. Quantitative metrics

In Table 1, J-DSAE outperforms TS-DSAE in terms of disentanglement according to RPA. Meanwhile, it overcomes the trade-off and excels in terms of FAD of Recon., suggesting superior reconstruction quality; **v**-swap (**z**-swap), indicating better timbre translation (preservation) after attribute swapping; and Rand., showing more realistic unconditional samples of timbre. Moreover, the factorised J-DSAE also outperforms AR TS-DSAE despite the low-capacity decoder D^U_{Fac}.

It suggests that replacing Eq. 2 - 5 with the Jacobian constraint significantly improves TS-DSAE. We hypothesise that the encoders and decoder could collaborate to minimise the auxiliary losses by encoding latent information that does not necessarily correspond to natural variation in the observation domain. On the other hand, the Jacobian constraint directly regularises the observation built upon the success of (E^T, D^T) instead of matching only the latent distributions.

5.2. Sampling from timbre space

We also train a J-DSAE with the size of the global latent space $D_v = 2$ (while D_z remains 32). In the left panel of Figure 2, we show $D^T(\mathbf{v}^{\text{grid}})$, where $\mathbf{v}^{\text{grid}} = (m, n)$ and $m, n \in \{-3, 0, 3\}$ are the coordinates of the timbre space. Note that only global attributes without temporal variation are rendered by the teacher decoder, which verifies the effectiveness of the Jacobian constraint (Eq. 8) in preserving overall timbre information. By representing the timbre as spectral distributions composed of horizontal strips, we can observe, from the top left to the bottom right, the transition from trumpet to violin or a shift from a high to low spectral centroid.

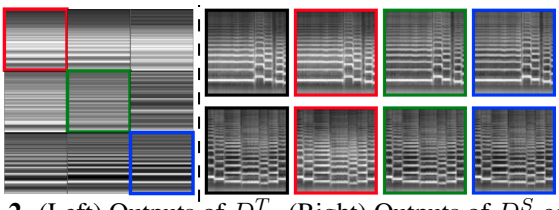

Fig. 2. (Left) Outputs of D^T. (Right) Outputs of D^S conditioned on timbre in the left panel and pitch from the leftmost column. Refer to text for details.

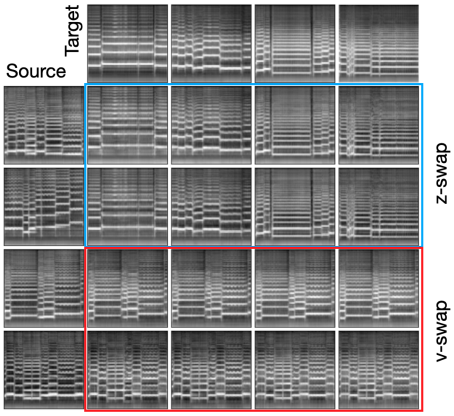

Fig. 3. Attribute translation. See Section 5.3.

Denote the diagonal by $\mathbf{v}^{\text{diag}}$ and the left column in the right panel by $\mathbf{x}^{\text{seed}}_{1:\tau}$. The second to last columns are $D^S(\mathbf{v}^{\text{diag}}, \mathbf{z}^{\text{seed}}_{1:\tau})$, where $\mathbf{z}^{\text{seed}}_{1:\tau} \sim E^S_{\mathbf{z}}(\mathbf{x}^{\text{seed}}_{1:\tau})$. Fixing $\mathbf{z}^{\text{seed}}_{1:\tau}$ while varying $\mathbf{v}^{\text{diag}}$, they share the pitch of $\mathbf{x}^{\text{seed}}_{1:\tau}$ and differ in the timbre dictated by $\mathbf{v}^{\text{diag}}$. Moreover, from each of the second to the last columns, the spectral distribution is consistent over the two samples, as they are conditioned on a common $\mathbf{v}^{\text{diag}}$.

5.3. Attribute translation

Figure 3 shows attribute translation. The blue block (**z**-swap) shows that the source timbre is preserved and combined with the target pitch. The red block (**v**-swap) shows that the source timbre is translated with the pitch unchanged. The rendering of a latent variable is largely independent of the other.

6. CONCLUSION

A strong regularisation is usually found in autoencoder-based models that facilitates learning disentangled representation in an unsupervised manner, which could hamper reconstruction or synthesis quality. By adapting the Jacobian constraint, we have proposed J-DSAE that overcomes the trade-off and improves upon TS-DSAE in terms of both unsupervised disentanglement of pitch and timbre as well as sampling quality.

Similarly to TS-DSAE, the success of J-DSAE depends on the teacher network that captures factors of global attributes. A more comprehensive study of the conditions that satisfy the requirement is left for future work.

7. REFERENCES

[1] Y. Bengio, "Deep learning of representations: Looking forward," *Int. Conference on Statistical Language and Speech Processing*, 2013.

[2] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," *Advances in Neural Information Processing Systems*, 2017.

[3] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," *Advances in Neural Information Processing Systems*, 2017.

[4] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," *Int. Conference on Machine Learning*, 2017.

[5] O. Cífka, A. Ozerov, U. Şimşekli, and G. Richard, "Self-supervised VQ-VAE for one-shot music style transfer," *Int. Conference on Acoustics, Speech and Signal Processing*, 2021.

[6] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, "Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders," *Int. Conference on Acoustics, Speech and Signal Processing*, 2020.

[7] Y.-J. Luo, K. W. Cheuk, T. Nakano, M. Goto, and D. Herremans, "Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds," *Int. Society for Music Information Retrieval Conference*, 2020.

[8] H. Zhang and S. Dixon, "Disentangling the horowitz factor: Learning content and style from expressive piano performance," *Int. Conference on Acoustics, Speech and Signal Processing*, 2023.

[9] Y. Li and S. Mandt, "Disentangled sequential autoencoder," *Int. Conference on Machine Learning*, 2018.

[10] J. Bai, W. Wang, and C. Gomes, "Contrastively disentangled sequential variational autoencoder," *Advances in Neural Information Processing Systems*, 2021.

[11] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3VAE: Self-supervised sequential VAE for representation disentanglement and data generation," *Int. Conference on Computer Vision and Pattern Recognition*, 2020.

[12] J. Han, M. R. Min, L. Han, L. E. Li, and X. Zhang, "Disentangled recurrent Wasserstein autoencoder," *Int. Conference on Learning Representations*, 2021.

[13] M. J. Vowels, N. C. Camgoz, and R. Bowden, "VDSM: Unsupervised video disentanglement with state-space modeling and deep mixtures of experts," *Int. Conference on Computer Vision and Pattern Recognition*, 2021.

[14] J. Lian, C. Zhang, and D. Yu, "Robust Disentangled Variational Speech Representation Learning for Zero-Shot Voice Conversion," *Int. Conference on Acoustics, Speech and Signal Processing*, 2022.

[15] 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.

[16] H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng, "Disentangled Speech Representation Learning for One-Shot Cross-Lingual Voice Conversion Using ß-VAE," *Spoken Language Technology Workshop*, 2023.

[17] Y.-J. Luo, S. Ewert, and S. Dixon, "Towards robust unsupervised disentanglement of sequential data — a case study using music audio," *Int. Joint Conference on Artificial Intelligence*, 2022.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Int. Conference on Learning Representations*, 2014.

[19] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," *Int. Conference on Machine Learning*, 2019.

[20] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, 2019.

[21] J. Lezama, "Overcoming the disentanglement vs reconstruction trade-off via Jacobian supervision," *Int. Conference on Learning Representations*, 2019.

[22] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," *INTERSPEECH*, 2019.

[23] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, 2014.

[24] J. W. Kim, J. Salamon, P. Q. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," *Int. Conference on Acoustics, Speech and Signal Processing*, 2018.