UNSUPERVISED DISENTANGLEMENT OF PITCH AND TIMBRE FOR ISOLATED MUSICAL INSTRUMENT SOUNDS

Yin-Jyun Luo[1,2] **Kin Wai Cheuk**[1,2] **Tomoyasu Nakano**[3]
Masataka Goto[3] **Dorien Herremans**[1,2]
[1] Singapore University of Technology and Design (SUTD), Singapore
[2] Institute of High Performance Computing, A*STAR, Singapore
[3] National Institute of Advanced Industrial Science and Technology (AIST), Japan
{yinjyun_luo, kinwai_cheuk}@mymail.sutd.edu.sg

ABSTRACT

Disentangling factors of variation aims to uncover latent variables that underlie the process of data generation. In this paper, we propose a framework that achieves unsupervised pitch and timbre disentanglement for isolated musical instrument sounds without relying on data annotations or pre-trained neural networks. Our framework, based on variational auto-encoders, takes as input a spectral frame, and encodes pitch and timbre as categorical and continuous variables, respectively. The input is then reconstructed by combining those variables. Under an unsupervised training setting, a major challenge is that encoders are tasked to capture factors of interest with distinct latent representations, without access to the corresponding ground-truth labels. We therefore introduce auxiliary tasks and objectives which leverage pitch shifting as a strategy to create surrogate labels, thereby encouraging the disentanglement of pitch and timbre. Through an ablation study we analyze the impact of the proposed objectives. The evaluation shows the efficacy of the proposed framework for learning disentangled representations, and verifies its applicability to unsupervised pitch classification and conditional spectral synthesis.

1. INTRODUCTION

The generative process from observed data can be described as having multiple latent factors of variation to explain the observations. For example, we may consider that a musical instrument sound consists of its pitch and timbre characteristic as the major underlying factors of variation. The concern of representation learning is to learn a model that captures such explanatory factors which are expected to be transferable to downstream tasks [1].

Disentanglement is said to be crucial for a good representation [1]. A disentangled representation allocates distinct factors of variation into separate dimensions of the representation, which facilitates an interpretable structure. Interventions along certain dimensions thereby only affect the corresponding latent factors, leading to a sparse change to the observation. In this paper, we propose a framework for learning disentangled representations of pitch and timbre, the two dominant factors of an isolated musical instrument sound. Unlike the supervised frameworks that address similar tasks [2, 3], we do not rely on data annotations or networks pre-trained on any form of supervision.

Many of the recent endeavors to achieve disentangled representation learning in an unsupervised setting are based on variational auto-encoders (VAEs) [4]. VAEs depict a data-generating process $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, where a multivariate latent variable $\mathbf{z}$ is first sampled from a prior distribution $p(\mathbf{z})$, and the observation $\mathbf{x}$ is sampled from the conditional distribution $p(\mathbf{x}|\mathbf{z})$ parameterized by a neural network; a variational distribution $q(\mathbf{z}|\mathbf{x})$, also parameterized using a neural network, is introduced to approximate the true posterior $p(\mathbf{z}|\mathbf{x})$.

In order to achieve disentanglement without access to data annotations, recent studies have proposed to impose regularizations on the latent space to promote a factorized aggregated posterior distribution $q(\mathbf{z})$ [5–7]. These works, however, demand further probes (e.g., traversal of latent space) to identify the semantics of the learned representations. One can also leverage prior knowledge of data structure and inject specific constraints [8–11]. For example, factors of variation of speech or video data are categorized as sequence-level (e.g., speakers) and segment-level (e.g., phonetic contents) latent variables [9, 10]. The mentioned prior knowledge, however, is not trivially applicable to factors of interest lacking of structural hierarchy (e.g., an isolated musical instrument sound with a constant pitch has both timbre and pitch as the sequence-level variable).

Given the challenge of disentangled representation learning in the unsupervised setting, literature has also assumed the accessibility to implicit or weak supervision in the form of grouped or paired data [12–14]. Our proposed framework, in contrast, does not require such a form of supervision; instead, we leverage pitch-shifting to create paired data, thereby introducing auxiliary objective functions to enhance feature disentanglement.

The underlying assumption of the proposed framework is that a moderate shift of pitch does not alter the timbre of the original musical instrument sound; we can thereby consider the original and its pitch-shifted version as a pair, and introduce several constraints to promote disentanglement of pitch and timbre. In particular, we adapt the contrastive learning method [15] to our framework, and maximize the similarity measure of the paired data. We also employ cycle-consistency loss [16, 17] to further improve the disentanglement. Moreover, we propose an objective function that explicitly accommodates the information of pitch difference that arises from pitch-shifting [18], which plays a key role for performance improvement. An ablation study is conducted to evaluate the efficacy of the introduced objective functions.

We consider a generative process that samples a categorical and a continuous latent variable, referred to as pitch and timbre, respectively, and samples the data conditioned on both the variables. This manifests the discrete nature of pitch and introduces a strong inductive bias crucial to the success of unsupervised disentanglement [19], which is made feasible as each sample in this study corresponds to a pitch class in the equal tempered scale.

For evaluation, classifiers are built to predict ground-truth pitch and instrument labels, which take as input the learned timbre representation. The low accuracy for pitch, and the high accuracy for instrument indicate a disentangled timbre representation. We also evaluate the pitch latent variable in terms of the metrics used for clustering tasks, which demonstrates the model's capability of unsupervised pitch classification. Attributed to the interpretability of the disentangled representation, we can achieve pitch-conditioning spectral synthesis whereby disentanglement is evaluated through the lens of conditional generation. We also propose a metric that accounts for consistency and diversity of pitch of the generated data. Our main contributions can be summarized as follows:

- Propose a framework based on VAEs to tackle unsupervised disentanglement of pitch and timbre.
- Leverage pitch-shifting which enables the auxiliary objectives that further introduce inductive biases to improve disentanglement.
- Design a metric that accounts for pitch consistency and diversity which quantifies the performance of disentanglement.

We present the proposed framework and the auxiliary objective functions in Section 2, and detail the implementation along with the experimental setup in Section 3. The evaluation methods and the proposed metric are elaborated in Section 4, followed by experimental results and discussions in Section 5. The paper is concluded in Section 6.

2. METHOD

In this section, we describe the proposed framework, and present the auxiliary objective functions that are introduced to further enhance the model.

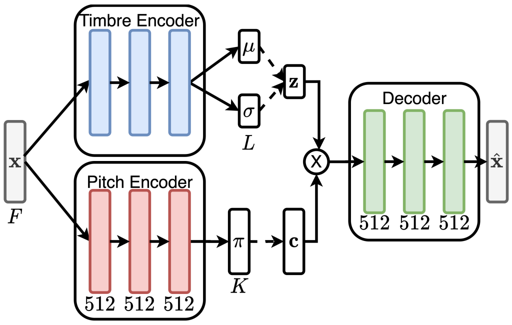

Figure 1: The proposed framework. The dashed lines denote sampling, and the cross denotes concatenation.

2.1 Overview

Figure 1 illustrates the proposed framework, which depicts a data-generating process of $\mathbf{x} \in \mathbb{R}^F$ being sampled from a conditional distribution $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c})$, referred to as a decoder, where $\mathbf{c} \in \mathbb{R}^K$ is a categorical latent variable for pitch, and $\mathbf{z} \in \mathbb{R}^L$ is a continuous latent variable for timbre. θ denotes the parameters of the decoder. Variational distributions $q_\phi(\mathbf{c}|\mathbf{x})$ and $q_\phi(\mathbf{z}|\mathbf{x})$, referred to as the pitch and timbre encoder, are introduced to approximate the true posterior distributions. The parameters of the two encoders are collectively denoted as ϕ. Under the framework of variational inference, the generative model is optimized through the evidence lower bound (ELBO) of $p_\theta(\mathbf{x})$:

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

For the continuous latent variable $\mathbf{z}$, we follow the literature [4] assuming $p(\mathbf{z}) = \mathcal{N}(0, I)$ and $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_\phi(\mathbf{x}), diag(\sigma^2_\phi(\mathbf{x})))$. For the categorical latent variable $\mathbf{c}$, we let $p(\mathbf{c}) = U(0, 1)$, a standard uniform distribution over number of categories K, and $q_\phi(\mathbf{c}|\mathbf{x}) = Cat(\mathbf{c}|\pi_\phi(\mathbf{x}))$. We can treat the pitch encoder as a pitch classifier that can be trained altogether with the entire network without pitch labels.

A major challenge for the unsupervised disentanglement is that the pitch encoder and timbre encoder are tasked to capture pitch and timbre features, respectively, without access to the corresponding labels. The presented model manifests the discrete nature of pitch with the categorical variable, thereby encouraging the pitch encoder to leave timbral information to the timbre encoder.

2.2 Gumbel-Softmax Distribution

In particular, we let $\mathbf{c}_k$ be a one-hot encoding of pitch, indexed at k, that is sampled from the $q_\phi(\mathbf{c}|\mathbf{x})$. To enable back-propagation through sampling of the discrete node, a common technique is to approximate $\arg\max$ with the Gumbel-Softmax distribution [20]. We specifically employ the straight-through estimator, which forward-passes the one-hot vector $\mathbf{c}_k$, and approximates its gradient with that of the Gumbel-Softmax distribution.

2.3 Auxiliary Objective Functions

Based on the underlying assumption that a moderate shift of pitch does not change the timbre of a musical instrument sound, we exploit pitch-shifting, thereby enabling the following auxiliary objective functions to enhance the model. We refer to $\mathbf{x}$ and $\mathbf{x}'$ respectively for the original sample and the pitch-shifted version throughout, and $(\mathbf{z}, \mathbf{c})$ and $(\mathbf{z}', \mathbf{c}')$ are the corresponding latent variables.

2.3.1 Latent Regression

One obvious auxiliary loss function enabled by pitch-shifting would be $\mathcal{L}_{\text{regression}} = \|\mathbf{z} - \mathbf{z}'\|_2^2$, which we include in the ablation study for comparison.

2.3.2 Contrastive Learning

We adapt SimCLR [15], a discriminative approach for representation learning [15, 21, 22], to our generative framework. Particularly, each sample in a minibatch of size N is pitch-shifted randomly upward or downward to a number of semitones, resulting in an augmented minibatch of size $2N$. A positive pair of data is defined as $(\mathbf{x}, \mathbf{x}')$, and the other $2(N-1)$ pairs are treated as negative ones, instead of being explicitly defined. The loss function for a positive pair, indexed as (i, j), is then defined as

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{m=1}^{2N} \mathbb{1}_{[m \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_m)/\tau)}, \quad (2)$$

where $\mathbb{1}_{[m \neq i]} \in \{0, 1\}$ is an indicator function evaluated as 1 if and only if $m \neq i$, and τ is a temperature parameter. The final loss $\mathcal{L}_{\text{contrast}}$ is obtained by aggregating $\mathcal{L}_{i,j}$ across all positive pairs. Following SimCLR [15], the cosine similarity is used as the similarity measure $\text{sim}(\cdot, \cdot)$.

Intuitively, the loss function attracts $\mathbf{z}$ and $\mathbf{z}'$ and repels the other negative pairs that are possibly derived from different instruments, which is expected to encourage the timbre encoder to extract pitch-invariant latent variables whereby the disentanglement is improved.

2.3.3 Cycle-consistency Loss

Cycle-consistency has been proposed to address unpaired image-to-image translation between different domains [16], which has been incorporated with VAEs to learn disentangled representations for images [17, 23].

We adopt the approach to further constrain the model and encourage disentanglement. Specifically, let E_p, E_t, and D denote the pitch encoder, the timbre encoder, and the decoder, respectively; whereby the cycle-consistency loss is defined as

$$\begin{aligned}\mathcal{L}_{cycle} =& \|E_t(D(\mathbf{z}, \mathbf{c}'_k)) - \mathbf{z}\|_2^2 + \|E_t(D(\mathbf{z}', \mathbf{c}_k)) - \mathbf{z}'\|_2^2 \\ &+ \text{CE}(E_p(D(\mathbf{z}, \mathbf{c}'_k)), k') + \text{CE}(E_p(D(\mathbf{z}', \mathbf{c}_k)), k),\end{aligned} \quad (3)$$

where $k = \arg\max_k q_\phi(\mathbf{c}|\mathbf{x})$ (similar for k'), and $\text{CE}(\cdot, \cdot)$ refers to cross-entropy loss.

Intuitively, given $(\mathbf{z}, \mathbf{c}'_k)$, D should generate a sample embodying timbre $\mathbf{z}$ and pitch category k', such that E_t and E_p can correctly predict $\mathbf{z}$ and k' in order to minimize the loss (similar for $(\mathbf{z}', \mathbf{c}_k)$). The objective function is expected to enforce D to faithfully render the given conditioning signals, and to further encourage E_t and E_p to encode the respective factors. Empirically, we freeze the weights of E_t and E_p when back-propagating $\mathcal{L}_{cycle}$ as suggested in the literature [24].

2.3.4 Surrogate Label Loss

We also propose to exploit the information of the shifted amount of semitones. Specifically, we minimize $\mathcal{L}_{\text{surrogate}} = \text{CE}(E_p(\mathbf{x}'), y')$. The surrogate label for $\mathbf{x}'$ is $y' = k + \delta$, where $k = \arg\max_k q_\phi(\mathbf{c}|\mathbf{x})$, $\delta \in [-S, S]$ denotes the shifted number of semitones, and S is the maximum amount of shift. This enforces E_p to acknowledge the pitch difference. As shown in Section 5, this loss term plays a key role in reaching the best-performing model.

To sum up, based on the underlying assumption that moderate shift of pitch does not alter timbre characteristics, all the objective functions are made possible thanks to the pitch-shifting strategy. The aggregated objective function to be maximized thereby becomes

$$\begin{aligned}\mathcal{L} = \mathcal{L}(\theta, \phi; \mathbf{x}) - (&\lambda_1 \mathcal{L}_{\text{regression}} + \lambda_2 \mathcal{L}_{\text{contrast}} \\ &+ \lambda_3 \mathcal{L}_{\text{cycle}} + \lambda_4 \mathcal{L}_{\text{surrogate}}),\end{aligned} \quad (4)$$

where λ_1, λ_2, λ_3, and λ_4 denote the weights of each loss term. We conduct an ablation study to investigate the efficacy of each auxiliary objective in terms of the metrics elaborated in Section 4.

3. EXPERIMENTAL SETUP

In this section, we describe the dataset used to evaluate our framework along with the implementation details.

3.1 Dataset

We train the framework using a subset of Studio-On-Line (SOL) [25], which includes 1,885 samples of 12 musical instruments and 82 possible pitches. We resample all the recordings to 22,050Hz, after which they are converted to short-time Fourier transform (STFT) with a 92ms of Hann window and 11ms of hop size. Mel-spectrograms with 256 filterbanks are then derived from power magnitude spectrum of the STFT. The dataset is split into a training set (90%) and validation set (10%), both of which have a same distribution of instruments. The magnitude of the Mel-spectrogram is logarithmically scaled, and min-max normalized within $[-1, 1]$ using the minimum and maximum values in the training set. The normalization is performed corpus-wide to preserve the variety of dynamics.

As a preliminary study, we extract the spectral frame at 200ms from the processed spectrograms, a time instant that usually displays the sustained part of a musical note; a datum is therefore referred to as a spectrum $\mathbf{x} \in \mathbb{R}^{256}$ of a Mel-spectrogram. To facilitate the timbre encoder to extract pitch-invariant features, we further derive 30-dimensional Mel-frequency cepstral coefficients (MFCCs) from the Mel-spectrograms. Therefore, the inputs to the timbre and pitch encoder are $\mathbf{x}_{\text{MFCC}} \in \mathbb{R}^{30}$ and

$\mathbf{x}_{\text{Mel-spec}} \in \mathbb{R}^{256}$, respectively. For convenience, we refer $\mathbf{x}$ to input data and do not distinguish $\mathbf{x}_{\text{MFCC}}$ and $\mathbf{x}_{\text{Mel-spec}}$ in the text and Figure 1. Note that the reconstruction target for evaluating $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c})$ remains as the Mel-spectrum.

As mentioned in Section 2, pitch-shifting is employed to augment the model by enabling the auxiliary objective functions. This is performed by stretching or shrinking an audio waveform with linear interpolation, which results in pitch-shifting in the frequency domain.

3.2 Implementation Details

Both the pitch and timbre encoders are comprised of three 512-unit fully-connected (FC) layers. They differ in the parametric layer; the pitch encoder outputs a categorical distribution $q_\phi(\mathbf{c}|\mathbf{x})$ through a FC layer with number of units equal to $K = 82$, i.e., the number of possible pitches. We henceforth refer $\mathbf{c}$ to *pitch category*, which differentiates from the pitch labels $\mathbf{y}$. The timbre encoder, on the other hand, contains a Gaussian parametric layer, which outputs two L-dimensional vectors, $L = 8$, representing mean $\mu_\phi(\mathbf{x})$ and variance $\sigma^2_\phi(\mathbf{x})$.

$\mathbf{c} \sim q_\phi(\mathbf{c}|\mathbf{x})$ and $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ are concatenated as the input to the decoder for reconstructing $\mathbf{x}$. The straight-through Gumbel-Softmax estimator [20] and the reparameterization trick [4] enable gradients to back-propagate through the parametric layers with stochastic gradient descent. The decoder is also composed of three 512-unit FC layers, which finally outputs $\hat{\mathbf{x}} \in \mathbb{R}^{256}$. Except for the two parametric layers, we use `tanh` as the activation function, and batch normalization follows thereafter.

All the experiments are conducted with a batch size of 256, and the model parameters are optimized using Adam [26] with a learning rate of 10^{-4}. The model stops training if the objective function (Equation (4)) does not improve over 300 epochs, i.e., we do not use any of the metrics presented in Section 4 as the stopping criteria, which assures absence of leakage of label information. We conduct an ablation study of the loss terms in Equation (4); however, we do not perform an exhaustive search for the corresponding weights, and instead evaluate $\lambda_i \in \{0, 1\}$ to investigate their effects. Fine-tuning the weights is left for future work.

4. EVALUATION METRICS

Our evaluation protocol relies on the properties of disentangled representations. From the synthesis point of view, the pitch of the synthesized spectrum should be invariant to perturbations in the timbre space as much as possible; from the perspective of analysis, the timbre space (pitch space) should mostly accommodate timbre information (pitch information) while minimizing clues for pitch (timbre). Accordingly, we consider the following metrics. Note that ground-truth annotations and pre-trained classifiers are employed only for evaluation purpose.

4.1 Classification Accuracy

We train logistic regression models which take as input the learned timbre latent variable $\mathbf{z}$ and predict labels of instrument and pitch. A well disentangled timbre representation should yield high accuracy for instrument, and low accuracy for pitch.

4.2 Clustering Accuracy (ACC)

During testing, the pitch encoder outputs a categorical distribution $q_\phi(\mathbf{c}|\mathbf{x})$ from which a pitch category of $\mathbf{x}$ can be assigned as $k = \arg\max_k q_\phi(\mathbf{c}|\mathbf{x})$. We can thereby consider it as a clustering task and calculate ACC [27] using pitch labels. Furthermore, since we do not train our model with pitch labels, the mapping from the inferred pitch categories to the pitch labels is unknown. For each category, we thus assign a pitch label that occurs the most within that category, and pitch classification accuracy can be calculated accordingly. This approximated pitch mapping is termed PM. Both ACC and PM are served to evaluate the unsupervised pitch classification.

4.3 Fréchet Inception Distance (FID)

We exploit FID [28] to quantitatively measure the quality of the synthesized spectrum. The metric measures the distributional difference between two multivariate Gaussians, which are fit to features derived from the real and generated samples, respectively. In our case, the features are extracted from a pre-trained instrument classifier, using the identical training data, which shares the same architecture with the encoder.

4.4 Consistency-Diversity Score (CDS)

In order to assess the model's capability of pitch-conditional generation, we propose a metric, termed CDS, to account for consistency of $p_k(\mathbf{y}|\hat{\mathbf{x}})$ and diversity of $\mathbb{E}_k[p_k(\mathbf{y}|\hat{\mathbf{x}})]$, where $p_k(\mathbf{y}|\hat{\mathbf{x}}) = p(\mathbf{y}|D(\mathbf{z}, \mathbf{c}_k))$ is the posterior distribution of a pre-trained pitch classifier given the generated samples $\hat{\mathbf{x}}$; and $\mathbb{E}_k[\cdot]$ denotes marginalization over k, where $k \in \{1, 2, \ldots, K\}$. Note that we can not simply measure pitch classification accuracy given the generated samples, as the true mapping from categories to pitch labels is unknown under the unsupervised setting.

Intuitively, the pre-trained pitch classifier should consistently output similar posterior distribution $p_k(\mathbf{y}|\hat{\mathbf{x}})$, if the generated samples $\hat{\mathbf{x}}$ are synthesized conditioned on a fixed $\mathbf{c}_k$ regardless of $\mathbf{z}$; and the aggregated distribution $\mathbb{E}_k[p_k(\mathbf{y}|\hat{\mathbf{x}})]$ should be uniformly distributed over $\mathbf{y}$, which indicates that the generated samples $\hat{\mathbf{x}}$, when conditioned on different $\mathbf{c}_k$'s, are predicted as having different pitches. Formally, CDS combines the two indicators as follows:

$$\begin{aligned} \text{CDS} &= -\mathbb{E}_k[H(p_k(\mathbf{y}|\hat{\mathbf{x}}))] + H(\mathbb{E}_k[p_k(\mathbf{y}|\hat{\mathbf{x}}])) \\ &= \mathbb{E}_k[\mathcal{D}_{\text{KL}}(p_k(\mathbf{y}|\hat{\mathbf{x}}) \| \mathbb{E}_k[p_k(\mathbf{y}|\hat{\mathbf{x}})])], \end{aligned} \tag{5}$$

namely, the marginal KL-divergence of the per-category and the aggregated posterior. A higher CDS thus hinds toward better consistency and diversity of pitch manifested by the generated pitch-conditioning spectrum.

λ_1	λ_2	λ_3	λ_4		Pitch	Instrument	Combine	ACC	PM	FID$_{\text{recon}}$	FID$_{\text{rand}}$	CDS
0	0	0	0	♭	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
				♯	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**

Table 1: The ablation study. For simplicity, we focus more on examining individual effects and do not exhaust all combinations. Each model (per row) is evaluated over all the evaluation metrics. For *Pitch* and *FID*, lower numbers indicate better performance, while the rest suggest the otherwise. The best-performing *unsupervised* models (♯, M0-M6) are highlighted.

Note that CDS bears resemblance to Inception Score (IS) [29]; the latter, however, was originally proposed to evaluate visual quality of synthetic images, whereas CDS evaluates the extent to which the model faithfully renders the conditional signal and enables correct classification.

5. EXPERIMENTS AND RESULTS

We train the framework with different configurations of the objective function $\mathcal{L}$ (Equation (4)), and quantify the performance of disentanglement with the metrics detailed in Section 4. For each model configuration, we initialize the model parameters with five random seeds, and report an averaged score along with standard deviation for each metric.

The results are summarized in Table 1, where we highlight the best-performing *unsupervised* models for each metric. Each row represents a model configuration; the symbol ♭ denotes the pitch-supervised model, which is trained to minimize an additional cross-entropy loss between the categorical distribution $q_\phi(\mathbf{c}|\mathbf{x})$ and the pitch labels, and is treated as a reference. The symbol ♯ denotes the unsupervised model trained without pitch-shifting. The rest M0-M6 are all unsupervised, utilizing pitch-shifting with maximum two semitones upward or downward.

For convenience, E_p, E_t, and D denote the pitch encoder, the timbre encoder, and the decoder, throughout.

5.1 Timbre Space Classification

Using the learned timbre representation $\mathbf{z}$, which we replace with $\mu_\phi(\mathbf{x})$ (μ in Figure 1) as the input feature to the logistic regression models, we obtain a relatively low accuracy for pitch classification, and a high accuracy for instrument classification, as shown in columns *Pitch* and *Instrument* in Table 1. As mentioned previously, low and high accuracy of pitch and instrument, respectively, indicate disentanglement of the timbre representation; we thereby aggregate the two metrics by $\frac{1}{2}(1-a_{\text{pitch}}+a_{\text{instrument}})$ shown in column *Combine*, where a_{pitch} and $a_{\text{instrument}}$ are the classification accuracy.

The pitch-supervised (reference) model attains the best aggregated score, as E_p is explicitly trained to classify pitch, thereby preventing pitch leak to E_t. Among the proposed models, M6 outperforms in terms of the aggregated score contributed by low a_{pitch}, which implies that combining all the auxiliary loss terms helps prevent pitch from leaking into the timbre space. Pitch-shifting alone improves the baseline unsupervised model significantly; this might be due to the rather imbalanced pitch distribution of the data. The high score attained without additional losses implies the efficacy of the proposed architectural design on disentangling pitch and timbre, given the augmented data.

Individually adding the auxiliary objective functions does not contribute much to the performance. For example, while M1-M4 degrade the aggregated score, combining M1-M3 (M5) approaches the best-performing model (M6). Notably, we can see that the proposed surrogate label loss further improves the performance of M5, which similarly applies to other metrics that follow.

5.2 Unsupervised Pitch Classification

As described in Section 4.2, we can consider E_p as a pitch classifier trained without labels as in our proposed models. We thus evaluate the performance with ACC.

We also report the pitch classification accuracy derived by the approximated mapping from pitch categories to pitch labels, which is the PM described in Section 4.2.

The supervised model can therefore be treated as the upper bound of pitch classification accuracy attained by the unsupervised E_p. M3 is the best model in terms of both ACC and PM, which could be attributed to the cycle-consistency that acknowledges the pitch-swapping scheme during training. This however promotes pitch leak to the timbre space as shown in column $Pitch$, which implies that an accurate E_p does not guarantee the absence of pitch leak, and, without supervision, more constraints are necessary to maintain both the pitch accuracy and timbre disentanglement, as demonstrated by M6.

5.3 Spectral Synthesis

We now turn our attention to the evaluation of generative tasks. In particular, we first evaluate the timbre representation by FID between the synthesized spectrum and the real one. To be more specific, the synthesized data are generated by D which takes as input $\mathbf{z}$ and $\mathbf{c}_k$, where $\mathbf{z} \sim p(\mathbf{z})$ and $k = \arg\max_k q_\phi(\mathbf{c}|\mathbf{x})$; that is, we first infer $\mathbf{c}_k$ of the

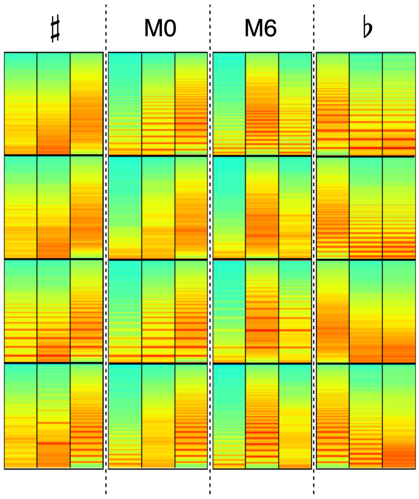

Figure 2: Pitch-conditioning spectrum generation. Each column represents a model, the bottom row refers to seed samples, and the top three rows correspond to generated samples conditioned on different pitch categories.

validation set, which is then combined with the randomly sampled $\mathbf{z}$ for decoding.

$\text{FID}_{\text{recon}}$ measures between the real and the reconstructed data, while FID_{rand} is for the real and the synthesized data. $\text{FID}_{\text{recon}}$ can thus be treated as a lower bound of the metric. From Table 1, it is clear that M6 prevails. As discussed earlier, adding the proposed $\mathcal{L}_{\text{surrogate}}$ to M5 makes the best model in terms of FID.

Interestingly, the supervised model does not achieve satisfying performance, which implies that the discriminability gain of E_p does not correlate well with the generative quality of timbre features. This similarly applies to the model that employs only the contrastive loss (M2).

5.4 Pitch-Conditioning Synthesis

Next we evaluate the disentanglement through the lens of conditional generation. Particularly, we first infer $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ from the validation set, which we directly take the mean $\mu_\phi(\mathbf{x})$ as the representative latent variable. We then enumerate all possible pitch categories $k \in \{1, 2, ..., K\}$, each of which is converted to a one-hot vector and concatenated with the inferred $\mathbf{z}$. D consumes the pitch-conditioned latent vector, and generates samples $\hat{\mathbf{x}}$ which are then classified by a pre-trained pitch classifier. This computes $p(\mathbf{y}|D(\mathbf{z}, \mathbf{c}_k))$, and CDS is derived by Equation (5). We report $\exp(\text{CDS})$ to restrict the value in $\{1, K\}$.

The supervised model performs well, due to the available pitch labels during training. M6 outperforms all the unsupervised models. Notably, $\mathcal{L}_{\text{surrogate}}$ alone (M4) outperforms M0, and, once again, the loss term further improves M5 to reach the best model M6.

The proposed $\mathcal{L}_{\text{surrogate}}$ synergizes with other loss terms, as evidenced by comparing M5 and M6 in terms of most metrics. This is probably attributed to the extra information from the amount of pitch-shift, which enables the model to explicitly account for the pitch difference [18].

Among all metrics, the t-test only yields a significant difference between the bold and M0 in terms of FID. Apart from the relatively high variances obtained by M0, this could be due to the small sample size (five random seeds) and the suboptimal configuration of values of loss weights, which we will investigate in future work.

5.5 Qualitative Study

We conclude our evaluation with a qualitative study on pitch-conditioning synthesis, as demonstrated in Figure 2. For each model (column), the bottom row refers to three reconstructed samples (with corresponding $\mathbf{z}$'s) which are the seed spectrums sampled from the validation set. Each of the rest of the rows corresponds to generated samples conditioned on the same $\mathbf{z}$'s but a different $\mathbf{c}$.

As a result, for each model (column), a good performance is indicated by a matched harmonic pattern across all three frames in a row (consistency), and diverse harmonic patterns across the top three rows (diversity). Note that the seed samples (bottom row) and the three conditioning pitches are not fixed across the four models, thus a direct comparison is not available. Nonetheless, we can have a rough idea that model ♯ does not perform as well as others, as harmonic patterns do not clearly appear aligned except for the one at the third row. This to some extent verifies the proposed CDS, in terms of which model ♯ attains the worst performance, although a study of larger scale is required for a faithful verification.

We can also observe that the overall timbre, characterized by the spectral energy distribution, stays rather consistent along each frame of each column despite the change of the pitch condition, which verifies the disentanglement.

6. CONCLUSION AND FUTURE WORK

We have proposed a VAE-based framework for unsupervised learning of disentangled pitch and timbre representation. The framework accommodates a categorical and a continuous latent variable, with the former embodying the discrete nature of pitch. We exploit pitch-shifting which enables the auxiliary objective functions, that are shown to potentially enhance the performance in terms of the quantitative evaluation.

A major challenge for future research is to infer pitch values from the categorical assignment, without access to ground-truth annotations. Furthermore, the proposed model imposes a strong inductive bias to the pitch encoder, by restricting degree of freedom through a one-hot encoded categorical pitch variable. This might pose a challenge when a tuning difference among instruments is present in the dataset. Increasing the capacity of the pitch representation while maintaining enough constraints for disentanglement is a direction for future work. We also aim to train the framework on a larger and more structured dataset [30], and to evaluate the method on data with larger time scale, for which we aim to learn dynamical latent factors on top of the global variables that we have studied.

7. ACKNOWLEDGMENTS

This work was supported by MOE Tier 2 grant no. MOE2018-T2-2-161 and SRG ISTD 2017 129, and JST ACCEL Grant Number JPMJAC1602. We also appreciate the efforts and insightful comments made by the reviewers.

8. REFERENCES

[1] Y. Bengio, "Deep learning of representations: Looking forward," in *Proc. of the Int. Conf. on Statistical Language and Speech Processing*. Springer, 2013, pp. 1–37.

[2] Y.-J. Luo, K. Agres, and D. Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders," in *Proc. of the Int. Society for Music Information Retrieval Conf.*, 2019, pp. 746–753.

[3] Y.-N. Hung, Y.-A. Chen, and Y.-H. Yang, "Learning disentangled representations for timber and pitch in music audio," *arXiv preprint arXiv:1811.03271*, 2018.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. of the Int. Conf. on Learning Representations*, 2014.

[5] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, M. Shakir, and A. Lerchner, "Beta-vae: Learning basic visual concepts with a constrained variational framework," in *Proc. of the Int. Conf. on Learning Representations*, 2017.

[6] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. of the Int. Conf. on Machine Learning*, 2018.

[7] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," in *Proc. of the Int. Conf. on Learning Representations*, 2018.

[8] M. Fraccaro, S. Kamronn, U. Paquet, and O. Winther, "A disentangled recognition and nonlinear dynamics model for unsupervised learning," in *Proc. of the Int. Conf. on Neural Information Processing Systems*, 2017, pp. 3601–3610.

[9] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in *Proc. of the Int. Conf. on Neural Information Processing Systems*, 2017.

[10] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Proc. of the Int. Conf. on Neural Information Processing Systems*, 2017, pp. 1878–1889.

[11] Y. Li and S. Mandt, "Disentangled sequential autoencoder," in *Proc. of the Int. Conf. on Machine Learning*, 2018.

[12] A. Ruiz, O. Martinez, X. Binefa, and J. Verbeek, "Learning disentangled representations with reference-based variational autoencoders," *arXiv preprint arXiv:1901.08534*, 2019.

[13] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multilevel variational autoencoder: Learning disentangled representations from grouped observations," in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2018.

[14] J. Chen and K. Batmanghelich, "Weakly supervised disentanglement by pairwise similarities," in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2019.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of the Int. Conf. on Machine Learning*, 2020.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2017, pp. 2223–2232.

[17] A. H. Jha, S. Anand, M. Singh, and V. Veeravasarapu, "Disentangling factors of variation with cycle-consistent variational auto-encoders," in *Proc. of the European Conf. on Computer Vision*, 2018, pp. 805–820.

[18] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020.

[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," in *Proc. of the Int. Conf. on Machine Learning*, 2019.

[20] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. of the Int. Conf. on Learning Representations*, 2017.

[21] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2017, pp. 2051–2060.

[22] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. of the Int. Conf. on Learning Representations*, 2018.

[23] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. of the European Conf. on Computer Vision*, 2018, pp. 35–51.

[24] J. Lezama, "Overcoming the disentanglement vs reconstruction trade-off via Jacobian supervision," in *Proc. of the Int. Conf. on Learning Representations*, 2019.

[25] G. Ballet, R. Borghesi, P. Hoffmann, and F. Levy, “Studio Online 3.0: An Internet “killer application” for remote access to IRCAM sounds and processing tools,” *Journee Informatique Musicale*, 1999.

[26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[27] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Proc. of the Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2003, pp. 267–273.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *Proc. of the Int. Conf. on Neural Information Processing Systems*, 2017, pp. 6626–6637.

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” in *Proc. of the Int. Conf. on Neural Information Processing Systems*, 2016, pp. 2234–2242.

[30] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.