ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414059

PITCH-TIMBRE DISENTANGLEMENT OF MUSICAL INSTRUMENT SOUNDS BASED ON VAE-BASED METRIC LEARNING

Keitaro Tanaka[1] *Ryo Nishikimi*[2] *Yoshiaki Bando*[3] *Kazuyoshi Yoshii*[2] *Shigeo Morishima*[4]

[1] Waseda University, Japan [2] Kyoto University, Japan
[3] National Institute of Advanced Industrial Science and Technology (AIST), Japan
[4] Waseda Research Institute for Science and Engineering, Japan

ABSTRACT

This paper describes a representation learning method for disentangling an arbitrary musical instrument sound into latent pitch and timbre representations. Although such pitch-timbre disentanglement has been achieved with a variational autoencoder (VAE), especially for a predefined set of musical instruments, the latent pitch and timbre representations are outspread, making them hard to interpret. To mitigate this problem, we introduce a metric learning technique into a VAE with latent pitch and timbre spaces so that similar (different) pitches or timbres are mapped close to (far from) each other. Specifically, our VAE is trained with additional contrastive losses so that the latent distances between two arbitrary sounds of the same pitch or timbre are minimized, and those of different pitches or timbres are maximized. This training is performed under weak supervision that uses only whether the pitches and timbres of two sounds are the same or not, instead of their actual values. This improves the generalization capability for unseen musical instruments. Experimental results show that the proposed method can find better-structured disentangled representations with pitch and timbre clusters even for unseen musical instruments.

***Index Terms*—** Disentangled representation learning, variational autoencoder, metric learning, pitch and timbre modeling.

1. INTRODUCTION

Disentangled representation learning aims to describe complex data as a combination of independent factors so that each factor affects a particular aspect of the data. This makes latent representations interpretable and enables us to intuitively control each factor in data generation. A popular approach is to train a deep latent variable model with a generative adversarial network (GAN) (*e.g.,* InfoGAN [1]) or a variational autoencoder (VAE) (*e.g.,* TVAE [2], β-VAE [3], FactorVAE [4], and HFVAE [5]). In addition to image data (*e.g.,* CelebA [6], MNIST [7], and 3D Chairs [8]), text data [9] and audio data [10] have also been dealt with.

In the field of music information retrieval (MIR), the disentanglement of the three major elements of sound, *i.e.,* volume, pitch, and timbre, from music audio signals has been considered to form the basis of music style transfer [11], automatic music generation [12], music analysis [13], and music recommendation [14]. Since the volume can be computed easily, the disentanglement of the pitch and timbre has mainly been focused on. Mor *et al.* [15], for example, proposed an autoencoder-based music translation method that can change only the timbral characteristics of music audio signals without affecting their pitch content, where multiple decoders corresponding to different timbre are used. Bitton *et al.* [16] proposed an extension of the β-VAE for many-to-many timbre transfer with a pair of the encoder and decoder. The learned latent representations are effective for generating musical instrument sounds, but do not match human perception. To solve this problem, Esling *et al.* [17] proposed a VAE-based method that learns a latent timbre space coherent with human perception by using the multi-dimensional scaling.

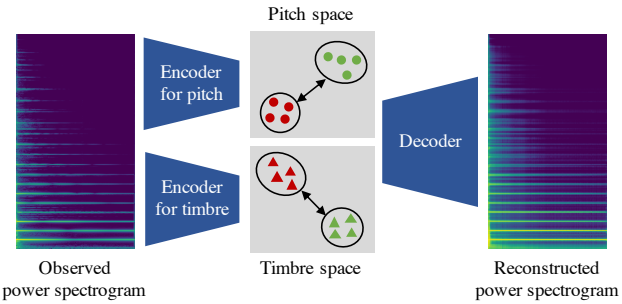

Fig. 1. The proposed VAE with contrastive losses for learning disentangled pitch and timbre representations of music instrument sounds.

While the above-mentioned studies focus on timbre representation learning under an condition that pitch representation is given. Hung *et al.* [18] first attempted to find disentangled pitch and timbre representations for music style transfer. Recently, Luo *et al.* [19, 20] proposed a Gaussian mixture VAE to disentangle a musical instrument sound into the two representations. Since this model assumes that the Gaussian distributions correspond to individual pitches and timbres (instruments), it cannot handle unseen pitches and timbres that are not included in the training data.

The need for dealing with *arbitrary* musical instrument sounds with any pitches and timbres turns our attention to metric learning used for representing the dissimilarities of samples as the distances in a latent space [21–25]. The basic approach is to train a deep neural network (DNN) so that similar samples are mapped close to each other and dissimilar samples are mapped far away from each other in a latent space. The key advantage of this approach is that samples of unseen categories (*e.g.,* pitches and timbres) that are not included in the training data can be dealt with (*a.k.a.* zero-shot learning [26,27]), because the DNN is trained by using only the information about the category match or mismatch of any two samples instead of using concrete category labels.

In this paper, we propose a VAE-based method for learning disentangled pitch and timbre representations from arbitrary music instrument sounds (Fig. 1). Our VAE consists of an encoder inferring the latent pitch and timbre representations from a given spectrogram

This paper was partially supported by JST ACCEL No. JPMJAC1602, and JSPS KAKENHI Nos. 16H01744, 19H04137, and 20K21813.

and a decoder representing the generation of the spectrogram from those representations. To make the latent pitch and timbre spaces cluster-structured and interpretable, we introduce contrastive losses so that musical instrument sounds of the same pitch or timbre are mapped close to each other, and those of different pitches or timbres are mapped far from each other. The VAE is trained in a weakly supervised manner, where only the information about the match or mismatch of the pitches and timbres of two spectrograms is used. Since the training of the VAE does not depend on the explicit pitch and timbre labels, it can successfully yield the disentangled representations of musical instrument sounds with unseen pitches and timbres.

The main contribution of this study is to train a VAE based on the contrastive losses for disentangled pitch and timbre representation learning for *arbitrary* musical instrument sounds. The pitch and timbre representations can be obtained without defining the vocabularies of pitches and timbres in advance. We show that the metric learning technique is effective in making the distances of the obtained representations close to the perceptual similarities of pitches and timbres. The latent representations obtained by the VAE potentially contain the other useful information (*e.g.,* tremolo and vibrato) in addition to the pitch and timbre information, which might not be obtained by metric learning only.

2. PROPOSED METHOD

The proposed method is based on weakly supervised metric learning with a VAE for the disentanglement of pitch and timbre. Our goal is to train a VAE that takes as input an observed power spectrogram $\mathbf{X} = \mathbf{x}_{1:T} \in \mathbb{R}_+^{F\times T}$ of an isolated musical instrument sound and outputs a reconstructed power spectrogram $\mathbf{Y} = \mathbf{y}_{1:T} \in \mathbb{R}_+^{F\times T}$ via two latent pitch and timbre representations $\mathbf{Z}^{\mathrm{p}} = \mathbf{z}^{\mathrm{p}}_{1:T} \in \mathbb{R}^{H\times T}$ and $\mathbf{Z}^{\mathrm{t}} = \mathbf{z}^{\mathrm{t}}_{1:T} \in \mathbb{R}^{H\times T}$ ($\mathbf{Z} = \{\mathbf{Z}^{\mathrm{p}}, \mathbf{Z}^{\mathrm{t}}\}$), where T, F, and H represent the number of time frames, that of frequency bins, and the dimension of each latent space, respectively. $\mathbf{x}_t \in \mathbb{R}_+^F$ and $\mathbf{y}_t \in \mathbb{R}_+^F$ are the observed and reconstructed power spectra at frame t, respectively. $\mathbf{z}^{\mathrm{p}}_t \in \mathbb{R}^H$ and $\mathbf{z}^{\mathrm{t}}_t \in \mathbb{R}^H$ indicate latent variables in the pitch and timbre spaces at frame t, respectively.

2.1. Generative model

We formulate a probabilistic model of the observed spectrogram $\mathbf{X}$ with latent representations $\mathbf{Z} = \{\mathbf{Z}^{\mathrm{p}}, \mathbf{Z}^{\mathrm{t}}\}$ as follows:

$$p_\theta(\mathbf{X}, \mathbf{Z}) = p_\theta(\mathbf{X}|\mathbf{Z})p(\mathbf{Z}), \tag{1}$$

where $p_\theta(\mathbf{X}|\mathbf{Z})$ is a likelihood function of $\mathbf{Z}$ for $\mathbf{X}$, $p(\mathbf{Z})$ is the prior of $\mathbf{Z}$, and θ is a set of model parameters. We formulate a deep generative model $p_\theta(\mathbf{X}|\mathbf{Z})$ as

$$p_\theta(\mathbf{X}|\mathbf{Z}) = \prod_{f=1}^{F}\prod_{t=1}^{T} \mathrm{Exponential}(x_{ft}|[\boldsymbol{\kappa}_\theta(\mathbf{Z})]_{ft}) \tag{2}$$

$$= \prod_{f=1}^{F}\prod_{t=1}^{T} \frac{1}{[\boldsymbol{\kappa}_\theta(\mathbf{Z})]_{ft}} \exp(-x_{ft}/[\boldsymbol{\kappa}_\theta(\mathbf{Z})]_{ft}), \tag{3}$$

where $\boldsymbol{\kappa}_\theta(\mathbf{Z})$ is the FT-dimensional output of a DNN with parameters θ that takes $\mathbf{Z}$ as input and the notation $[\mathbf{A}]_{ij}$ indicates the ij-th element of $\mathbf{A}$. $p(\mathbf{Z})$ is set to a standard Gaussian distribution as

$$p(\mathbf{Z}) = p(\mathbf{Z}^{\mathrm{p}})p(\mathbf{Z}^{\mathrm{t}}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^{\mathrm{p}}_t|\mathbf{0}_H, \mathbf{I}_H)\mathcal{N}(\mathbf{z}^{\mathrm{t}}_t|\mathbf{0}_H, \mathbf{I}_H), \tag{4}$$

where $\mathbf{0}_H$ is the all-zero vector of size H and $\mathbf{I}_H$ is the identity matrix of size $H \times H$.

2.2. VAE-based training

Given an observed spectrogram $\mathbf{X}$, we aim to infer the latent representations $\mathbf{Z}$ and estimate the model parameter θ in a maximum-likelihood sense. Because the DNN-based formulation of our generative model makes the posterior distribution $p_\theta(\mathbf{Z}|\mathbf{X})$ intractable, we compute it approximately with a VAE. More specifically, we introduce a variational distribution $q_\phi(\mathbf{Z}|\mathbf{X}) = q_\phi(\mathbf{Z}^{\mathrm{p}}|\mathbf{X})q_\phi(\mathbf{Z}^{\mathrm{t}}|\mathbf{X})$ parameterized by ϕ and optimize it such that the Kullback-Leibler (KL) divergence from $q_\phi(\mathbf{Z}|\mathbf{X})$ to $p_\theta(\mathbf{Z}|\mathbf{X})$ is minimized. In this paper, $q_\phi(\mathbf{Z}|\mathbf{X})$ is implemented with a DNN parameterized by ϕ as follows:

$$q_{\phi^*}(\mathbf{Z}^*|\mathbf{X}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^*_t|[\boldsymbol{\mu}^*_{\phi^*}(\mathbf{X})]_t, [\boldsymbol{\sigma}^{*2}_{\phi^*}(\mathbf{X})]_t), \tag{5}$$

where $*$ represents "p" or "t", $\boldsymbol{\mu}^*_{\phi^*}(\mathbf{X})$ and $\boldsymbol{\sigma}^{*2}_{\phi^*}(\mathbf{X})$ are the FT-dimensional outputs of the DNN with parameters ϕ^*. Similar to the deep generative model, the outputs of the DNN represent the parameters of probabilistic distributions.

Instead of directly maximizing $\log p_\theta(\mathbf{X})$ with respect to the model parameters θ, we maximize its variational lower bound $\mathcal{L}^{\mathrm{vae}}$ derived by introducing $q_\phi(\mathbf{Z}|\mathbf{X})$ as follows:

$$\mathcal{L}^{\mathrm{vae}} = \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})}[\log p_\theta(\mathbf{X}|\mathbf{Z})] - \mathcal{D}_{\mathrm{KL}}(q_\phi(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z})), \tag{6}$$

where $\phi = \{\phi^{\mathrm{p}}, \phi^{\mathrm{t}}\}$, and the equality holds, *i.e.,* $\mathcal{L}^{\mathrm{vae}}$ is maximized, if and only if $q_\phi(\mathbf{Z}|\mathbf{X}) = p_\theta(\mathbf{Z}|\mathbf{X})$. Note that this condition cannot be satisfied because $p_\phi(\mathbf{Z}|\mathbf{X})$ is hard to compute. Because the gap between $\log p_\theta(\mathbf{X})$ and $\mathcal{L}^{\mathrm{vae}}$ in Eq. (6) is equal to the KL divergence from $q_\phi(\mathbf{Z}|\mathbf{X})$ to $p_\theta(\mathbf{Z}|\mathbf{X})$, the minimization of the KL divergence is equivalent to the maximization of $\mathcal{L}^{\mathrm{vae}}$. The expectation term of $\mathcal{L}^{\mathrm{vae}}$ is calculated with the reparameterization trick [28], and the KL divergence can be calculated analytically so that the networks can be optimized with a gradient method.

The VAE-based training can disentangle the observed spectrogram $\mathbf{X}$ into the two latent representations of $\mathbf{Z}^{\mathrm{p}}$ and $\mathbf{Z}^{\mathrm{t}}$ such that they are statistically independent. The obtained representations, however, are not necessarily useful because it does not reflect the perceptual similarity of the latent representations. We deal with this limitation by introducing metric learning to encourage the perceptual disentanglement of the latent space.

2.3. Pairwise metric learning for disentanglement

The encoders transform an observed spectrum $\mathbf{x}_t$ into the latent variables $\mathbf{z}^{\mathrm{p}}_t$ and $\mathbf{z}^{\mathrm{t}}_t$ in the two latent spaces. Ideally, latent representations of the same pitch (timbre, *i.e.,* instrument) should be close to each other, and those of different pitches (timbres) should be far away from each other in each latent space.

To acquire such latent representations, we use a pairwise metric. We consider the mini-batch training with N spectrograms $\{\mathbf{X}_n\}_{n=1}^N$ of musical instrumental sounds and randomly select two spectrograms $\mathbf{X}_i$ and $\mathbf{X}_j$ ($i \neq j$) from the spectrograms, where N is the even number of the batch size, and $i, j \in \{1, \ldots, N\}$ are the indices of the training samples. The latent variables $\mathbf{Z}^*_i$ and $\mathbf{Z}^*_j$ for each pair of the spectrograms are then obtained by using the encoders. The metric learning is conducted with contrastive loss functions $\mathcal{L}^{\mathrm{p}}_c$ and $\mathcal{L}^{\mathrm{t}}_c$. The loss function for pitch $\mathcal{L}^{\mathrm{p}}_c$ is calculated as

$$\mathcal{L}^{\mathrm{p}}_c = \begin{cases} \displaystyle\sum_{i,j\in N} (\mathcal{D}^{\mathrm{p}}_{ii} + \mathcal{D}^{\mathrm{p}}_{jj} + \mathcal{D}^{\mathrm{p}}_{ij}) \\ \qquad (\text{if } i \text{ and } j \text{ are from the same pitch}) \\ \displaystyle\sum_{i,j\in N} (\mathcal{D}^{\mathrm{p}}_{ii} + \mathcal{D}^{\mathrm{p}}_{jj} - \mathcal{D}^{\mathrm{p}}_{ij}) \qquad (\text{otherwise}), \end{cases} \tag{7}$$

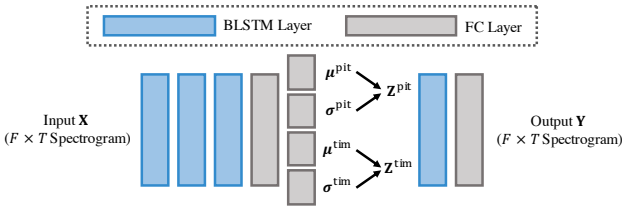

Fig. 2. The implementation of the VAE.

where $\mathcal{D}^*_{ii}, \mathcal{D}^*_{jj}, \mathcal{D}^*_{ij}$ are the sums of distances between the two latent variables $\mathbf{Z}^{\mathrm{p}}_i$ and $\mathbf{Z}^{\mathrm{p}}_j$. They are defined as follows:

$$\mathcal{D}^{\mathrm{p}}_{i,i} = \sum_{t_1=1}^{T-1} \sum_{t_2=t_1+1}^{T} \|\mathbf{z}^{\mathrm{p}}_{it_1} - \mathbf{z}^{\mathrm{p}}_{it_2}\|, \tag{8}$$

$$\mathcal{D}^{\mathrm{p}}_{j,j} = \sum_{t_1=1}^{T-1} \sum_{t_2=t_1+1}^{T} \|\mathbf{z}^{\mathrm{p}}_{jt_1} - \mathbf{z}^{\mathrm{p}}_{jt_2}\|, \tag{9}$$

$$\mathcal{D}^{\mathrm{p}}_{i,j} = \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} \|\mathbf{z}^{\mathrm{p}}_{it_1} - \mathbf{z}^{\mathrm{p}}_{jt_2}\|, \tag{10}$$

where $\|\cdot\|$ is the euclidean norm of a vector. The loss function for timbre $\mathcal{L}^{\mathrm{t}}_c$ is calculated in the same manner. The values of these contrastive losses take large values when latent representations of the same pitch (timbre) are far away from each other or those of different pitches (timbres) are close to each other. Thus, we can encourage the latent spaces of the proposed generative model to be more perceptually interpretable.

We train the proposed networks in a weakly supervised manner, where only information on whether pitches and timbres of a pair of observed spectrograms fed into the proposed VAE are identical or not is required, and their actual labels are not necessary. The training is conducted with the following total loss function $\mathcal{L}^{\mathrm{total}}$ combining the VAE loss in Eq. (6) and the contrasive loss functions in Eq. (7):

$$\mathcal{L}^{\mathrm{total}} = -\mathcal{L}^{\mathrm{vae}} + \alpha\mathcal{L}^{\mathrm{p}}_c + \beta\mathcal{L}^{\mathrm{t}}_c, \tag{11}$$

where α and β are hyperparameters to control the weights of the contrastive loss functions.

3. EVALUATION

This section describes experiments conducted to evaluate the performance of the proposed method for the disentanglement.

3.1. Data

To evaluate the proposed method, we used musical instrument sounds from the RWC Music Database [29] except for Shakuhachi, Soprano, and Alto. Each file in the database is annotated with an instrument name and includes the sounds of the whole pitches that the instrument can play. We split the sound signals in each file into individual pitch sounds automatically with mute detection and removed the silence regions at the start of the split sounds with onset detection. We selected the sounds of pitches from C3 to B5. We split the obtained sounds (40914 files, fifty instruments) into three sets: a training set (29957 sounds, forty instruments) and an evaluation set (10957 sounds, test instruments). For determining the training termination with 2-fold cross-validation, the evaluation set was split into two subsets, each of which has five instruments. The three sets shared pitches but did not share instruments.

Table 1. The denseness and the divergence in each space.

Methods	Pitch representations		Timbre representations	
	Denseness	Divergence	Denseness	Divergence
Vanilla VAE	3.334	2.279	3.640	1.541
Proposed VAE	**2.891**	**3.551**	**3.420**	**2.654**

All sounds were resampled at 22050 Hz, and we used only the first two seconds of each sound. We used a short-time Fourier transform (STFT) with a Hann window of 1024 samples and a shifting interval of 256 samples to obtain the spectrogram of $F = 513$ and $T = 173$. We normalized each spectrogram such that the average power of each spectrogram was one.

3.2. Model configuration

Our VAE-based method utilized the bidirectional long short-term memory (BLSTM) architecture for its encoder and decoder to capture temporal characteristics of sounds (Fig. 2). The encoder consisted of three-layers of BLSTMs with 2×300 cells and fully connected (FC) layers. We set the dropout rate to 0.3 for each BLSTM layer of the encoder. A shared FC layer transformed 600 dimensions into 256 dimensions. The outputs of the BLSTMs and the FC layer were passed through the leaky ReLU. Four independent FC layers independently transformed 256 dimensions into $H = 16$ dimensions to represent means and variances of the latent variables. The decoder consisted of a one-layer BLSTM with 2×300 cells followed by an FC layer. The output of the BLSTM was also passed through the leaky ReLU. Since each time-frequency bin of a spectrogram took a non-negative value, we applied softplus for the outputs of the FC layer. The batch size N was sixteen. We used an Adam [30] optimizer with a learning rate of 0.001. The denseness and divergence in the latent spaces were able to be controlled by changing the weights α and β for the contrastive losses. We experimentally confirmed that the proposed method failed in reconstruction when the weights were greater than 0.5, and we set them to 0.2.

3.3. Evaluation criteria

We evaluate the following denseness and divergence that indicate how close the latent variables with the same pitch or timbre label are, and how far the latent variables with different pitch or timbre labels are, respectively. We calculate these criteria of pitch space as

$$\text{Denseness} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{9N_m}\sum_{n=1}^{N_m}\sum_{t=1}^{9}\|\mathbf{z}^{\mathrm{p}}_{mnt} - \boldsymbol{\eta}^{\mathrm{p}}_m\|, \tag{12}$$

$$\text{Divergence} = \frac{2}{M(M-1)}\sum_{m_1=1}^{M-1}\sum_{m_2=m_1+1}^{M}\|\boldsymbol{\eta}^{\mathrm{p}}_{m_1} - \boldsymbol{\eta}^{\mathrm{p}}_{m_2}\|, \tag{13}$$

$$\boldsymbol{\eta}^{\mathrm{p}}_m = \frac{1}{9N_m}\sum_{n=1}^{N_m}\sum_{t=1}^{9}\mathbf{z}^{\mathrm{p}}_{mnt}, \tag{14}$$

where M, N_m, and $\boldsymbol{\eta}_m$ indicate the number of the pitches, that of sounds with pitch m, and the mean of all latent variables with pitch m, respectively. We calculate those of timbre space in the same manner. We selected only the first nine frames (about 100 ms) because some sounds include silent regions. We normalized each latent space to remove the impact of its scale.

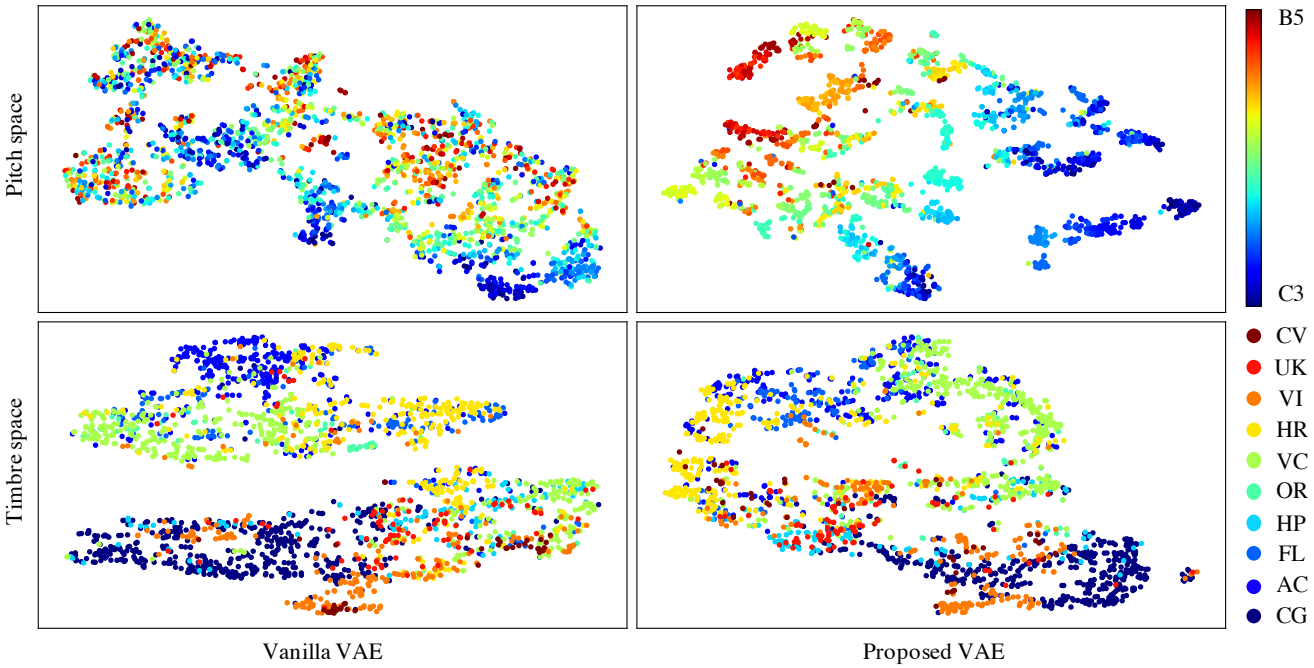

Fig. 3. Visualizations of the pitch and timbre spaces.

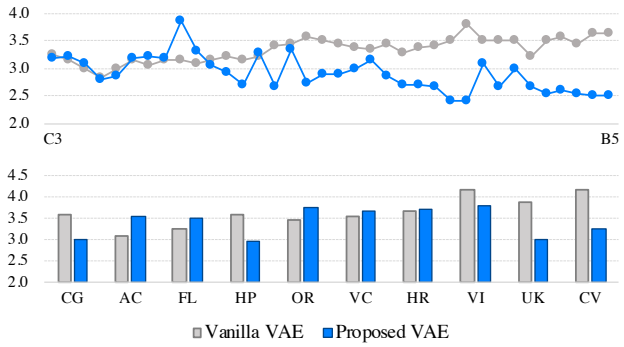

Fig. 4. The disentanglement performances in the denseness. The upper and lower figures are the denseness in the pitch and timbre spaces, respectively.

3.4. Experimental results

Experimental results are shown in Table. 1. The denseness (smaller is better) got smaller, and the divergence (larger is better) got larger in both latent pitch and timbre spaces by introducing the metric learning. Fig. 3 visualizes the latent pitch and timbre spaces using t-distributed stochastic neighbor embedding (t-SNE) [31]. The proposed method found better-structured disentangled representations with pitch and timbre clusters for unseen musical instruments. In Fig. 4, the proposed method achieved better denseness for most pitches and timbres by using the contrastive losses. In comparing the upper left and upper right matrices in Fig. 5, the values of the elements around the diagonal got smaller, and those distant from the diagonal got larger. This indicates that our method succeeded in making latent variables of similar pitches close to each other and those of different pitches far away from each other in the latent pitch space. In the lower matrices in Fig. 5, the off-diagonal elements of the right matrix have larger values than those of the left matrix. These results demonstrated that the proposed method succeeded in mapping the different timbres to be distant from each other.

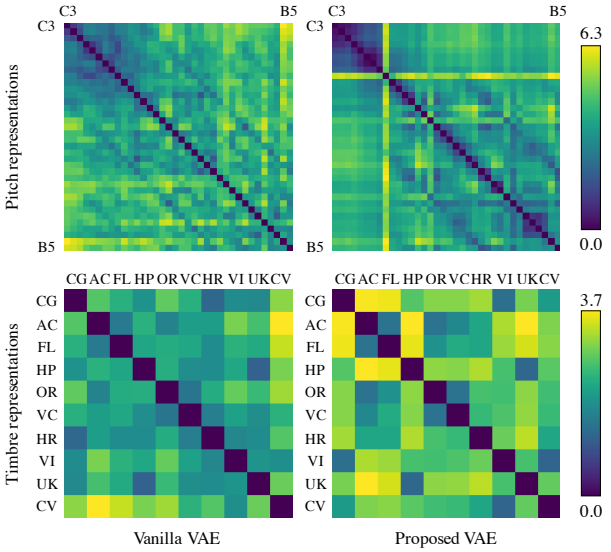

Fig. 5. The disentanglement performances in the divergence.

4. CONCLUSION

This paper presented the VAE-based method for disentangling a musical instrument sound into latent pitch and timbre representations. We utilized the metric learning technique to control each disentangled space based on the similarity of sounds. We also proposed a weakly supervised learning method to achieve the metric learning. We experimentally confirmed that our method more successfully disentangled the latent pitch and timbre representations compared to the vanilla VAE. Our future work includes extending the proposed method to map one instrument sound into one pitch representation and one timbre representation to embed the temporal information of input sounds in the latent spaces. We also plan to incorporate other metric learning techniques into the proposed method to obtain better latent spaces for improving the quality of the reconstructions.

5. REFERENCES

[1] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 2172–2180.

[2] H. Ishfaq, A. Hoogi, and D. Rubin, "TVAE: Triplet-based variational autoencoder using metric learning," in *arXiv:1802.04403*, 2018, pp. 1–4.

[3] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations (ICLR)*, 2017.

[4] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning (ICML)*, 2018.

[5] B. Esmaeili, H. Wu, S. Jain, A. Bozkurt, N. Siddharth, B. Paige, D. H. Brooks, J. Dy, and J. Meent, "Structured disentangled representations," in *Proceedings of Machine Learning Research (PMLR)*, 2019, pp. 2525–2534.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[8] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic, "Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] S. Ravfogel, Y. Elazar, J. Goldberger, and Y. Goldberg, "Unsupervised distillation of syntactic information from contextualized word representations," in *arXiv:2010.05265*, 2020.

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

[11] S. Dai, Z. Zhang, and G. G. Xia, "Music style transfer: A position paper," in *International Workshop on Music Metacreation (MUME)*, 2018.

[12] J. Briot, G. Hadjeres, and F. Pachet, "Deep learning techniques for music generation – A survey," *Computational Synthesis and Creative Systems*, pp. 1–249, 2020.

[13] R. Yang, D. Wang, Z. Wang, T. Chen, J. Jiang, and G. Xia, "Deep music analogy via latent representation disentanglement," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 596–603.

[14] D. Paul and S. Kundu, "A survey of music recommendation systems with a proposed music recommendation system," *Emerging Technology in Modelling and Graphics*, pp. 279–285, 2020.

[15] N. Mor, L. Wolf, A. Polyak, and Y. Taigman, "A universal music translation network," in *arXiv:1805.07848*, 2018.

[16] A. Bitton, P. Esling, and A. Chemla-Romeu-Santos, "Modulated variational auto-encoders for many-to-many musical timbre transfer," in *arXiv:1810.00222*, 2018.

[17] P. Esling, A. Chemla–Romeu-Santos, and A. Bitton, "Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 175–181.

[18] Y. Hung, I. Chiang, Y. Chen, and Y. Yang, "Musical composition style transfer via disentangled timbre representations," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 4697–4703.

[19] Y. Luo, K. Agres, and D. Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 746–753.

[20] Y. Luo, K. W. Cheuk, T. Nakano, M. Goto, and D. Herremans, "Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 700–707.

[21] R. Lu, K. Wu, Z. Duan, and C. Zhang, "Deep ranking: Triplet matchnet for music metric learning," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 121–125.

[22] J. Royo-Letelier, R. Hennequin, V. Tran, and M. Moussallam, "Disambiguating music artists at scale with audio metric learning," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 622–629.

[23] F. Karsdorp, P. Kranenburg, and E. Manjavacas, "Learning similarity metrics for melody retrieval," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 478–485.

[24] M. C. McCallum, "Unsupervised learning of deep features for music segmentation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 346–350.

[25] A. Schindler and P. Knees, "Multi-task music representation learning from multi-label embeddings," in *Content-Based Multimedia Indexing (CBMI)*, 2019, pp. 1–6.

[26] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *National Conference on Artificial Intelligence*, 2008, pp. 646–651.

[27] M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[29] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2003, pp. 229–230.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv:1412.6980*, 2014.

[31] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, pp. 2579–2605, 2008.