Paper ID: 5290

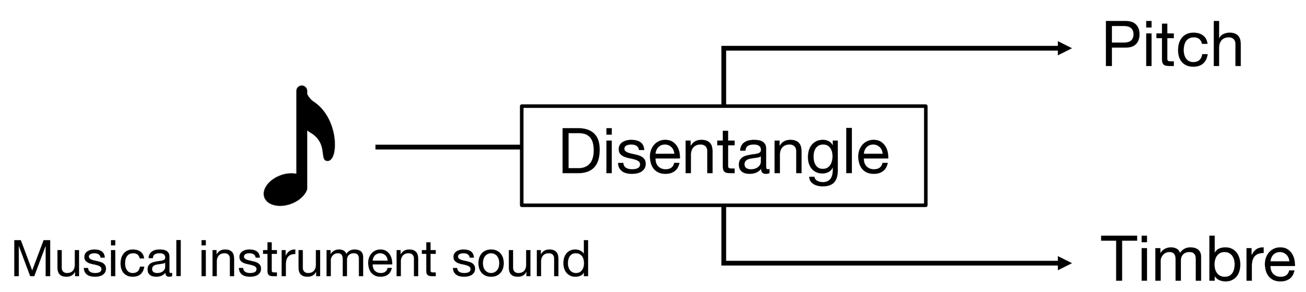
Pitch-Timbre Disentanglement of Musical Instrument Sounds Based on VAE-Based Metric Learning

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

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

Backgrounds

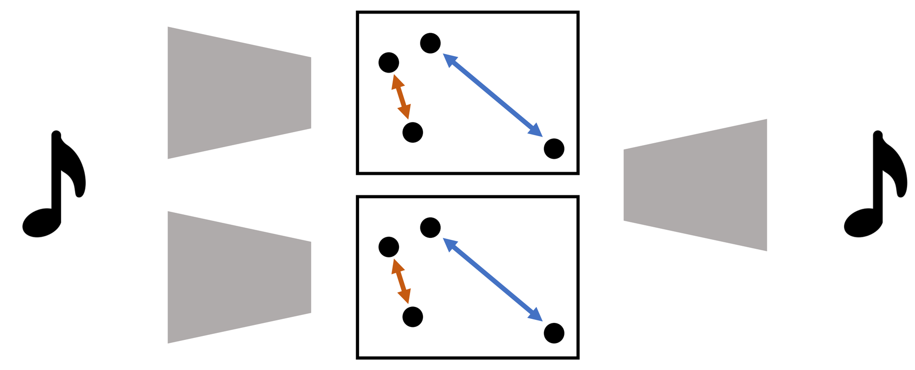
What is **Disentanglement**?



- To describe data as a combination of independent factors
- A **sound** can be disentangled into **pitch and timbre**
- Conventional approach cannot treat an arbitrary musical instrument sound **because of using the concrete category labels**

Approach

Introduce a **metric learning technique**

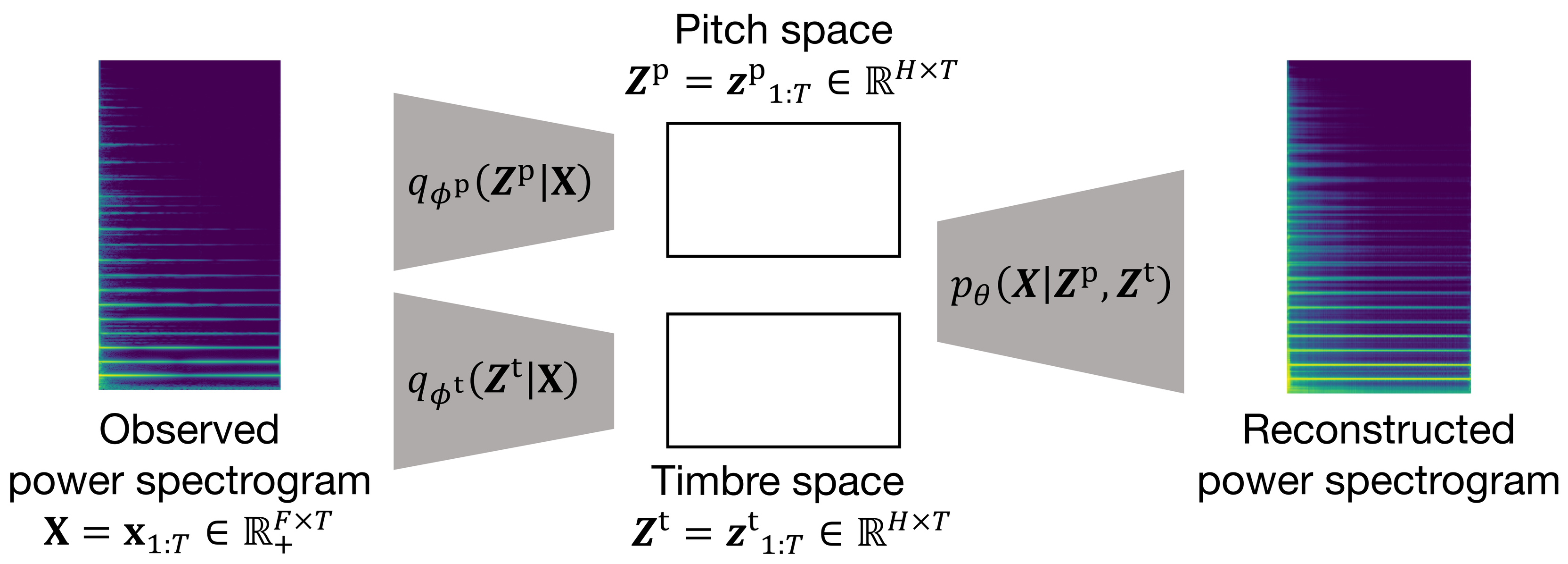


- A technique used for **representing the dissimilarities** of samples as the distances in a latent space
- **Similar** samples are mapped **close** to each other
- **Dissimilar** samples are mapped **far away** from each other

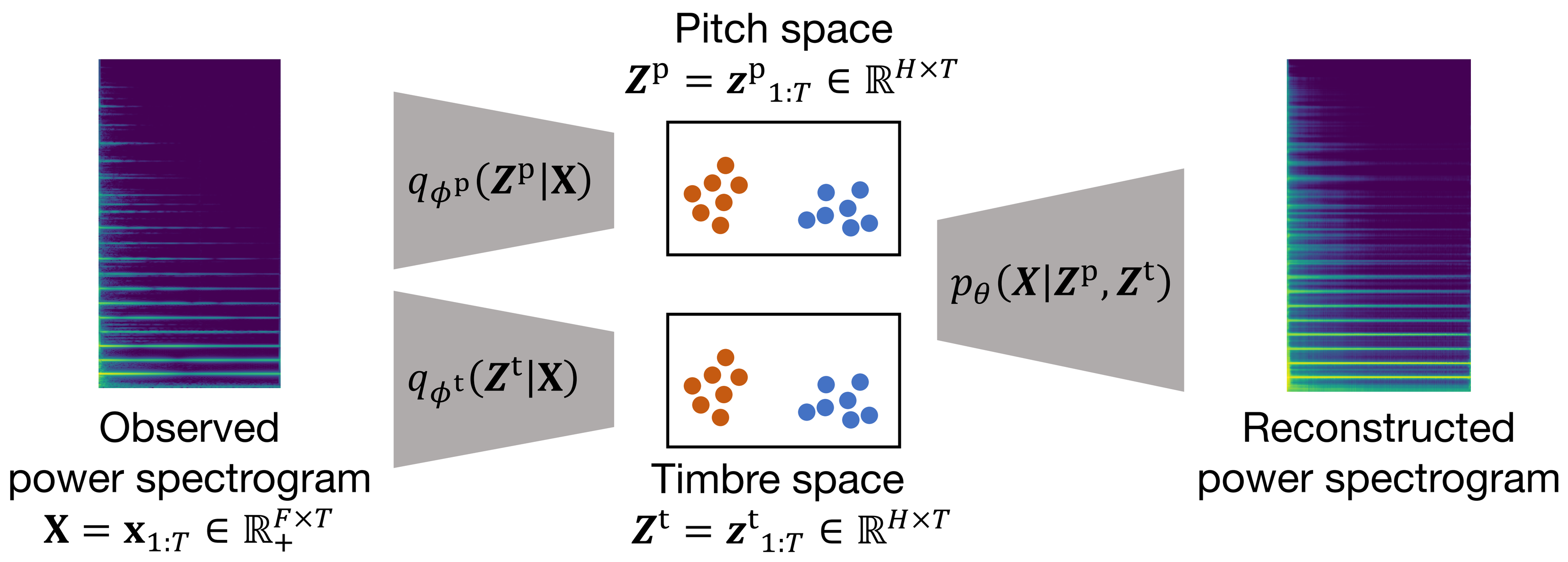
Method

Generative Model of the Observed Spectrogram

Formulate a probabilistic model of the observed spectrogram $\mathbf{X}$ with latent representations $\boldsymbol{Z}^{\mathrm{p}}$ and $\boldsymbol{Z}^{\mathrm{t}}$



Transform two observed spectrograms $\mathbf{X}_1$ and $\mathbf{X}_2$ into the latent variables $\boldsymbol{Z}^{\mathrm{p}}$ and $\boldsymbol{Z}^{\mathrm{t}}$ independently



- Conduct pairwise metric learning for these $2T$ samples

Pairwise Metric Learning for Disentanglement

Contrastive loss functions $\mathcal{L}_{\mathrm{c}}^{\mathrm{p}}$ (for pitch) and $\mathcal{L}_{\mathrm{c}}^{\mathrm{t}}$ (for timbre)

$$\mathcal{L}_{\mathrm{c}}^{\mathrm{p}} = \mathcal{D}_{11}^{\mathrm{p}} + \mathcal{D}_{22}^{\mathrm{p}} \pm \mathcal{D}_{12}^{\mathrm{p}}$$ ($\mathcal{L}_{\mathrm{c}}^{\mathrm{t}}$ is calculated in a similar way)

The sum of distances between all latent variable pairs of the same spectrogram ($\mathbf{X}_1$ and $\mathbf{X}_2$)

The sum of distances between all latent variable pairs from different spectrograms
+: if $\mathbf{X}_1$ and $\mathbf{X}_2$ have the same pitch
−: otherwise

- $\mathcal{L}_{\mathrm{c}}^{\mathrm{p}}$ and $\mathcal{L}_{\mathrm{c}}^{\mathrm{t}}$ pull similar samples close to each other and keep dissimilar samples far from each other

Training with the Weakly Supervised Learning

The training is conducted with the total loss function $\mathcal{L}^{\mathrm{total}}$

$$\mathcal{L}^{\mathrm{total}} = -\mathcal{L}^{\mathrm{vae}} + \alpha\mathcal{L}_{\mathrm{c}}^{\mathrm{p}} + \beta\mathcal{L}_{\mathrm{c}}^{\mathrm{t}}$$

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

Hyperparameters to control the weights

- Only information on **whether pitches and timbres of a pair of observed spectrograms are identical or not is required**
- Their actual labels are not necessary
- **An arbitrary musical instrument sound can be treated**

Evaluation

Evaluate **denseness** and **divergence** for unseen musical instruments

- **Denseness** shows how **close** the latent variables with **the same pitch or timbre** label are
- Calculated as:

$$\frac{1}{M}\sum_{m=1}^{M}\frac{1}{9N_m}\sum_{n=1}^{N_m}\sum_{t=1}^{9}\left\|\mathbf{z}_{mnt}^{\mathrm{p}} - \boldsymbol{\eta}_m^{\mathrm{p}}\right\|$$ (Timbre in a similar way)

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

- **Divergence** shows how **far** the latent variables with **different pitch or timbre** labels are
- Calculated as:

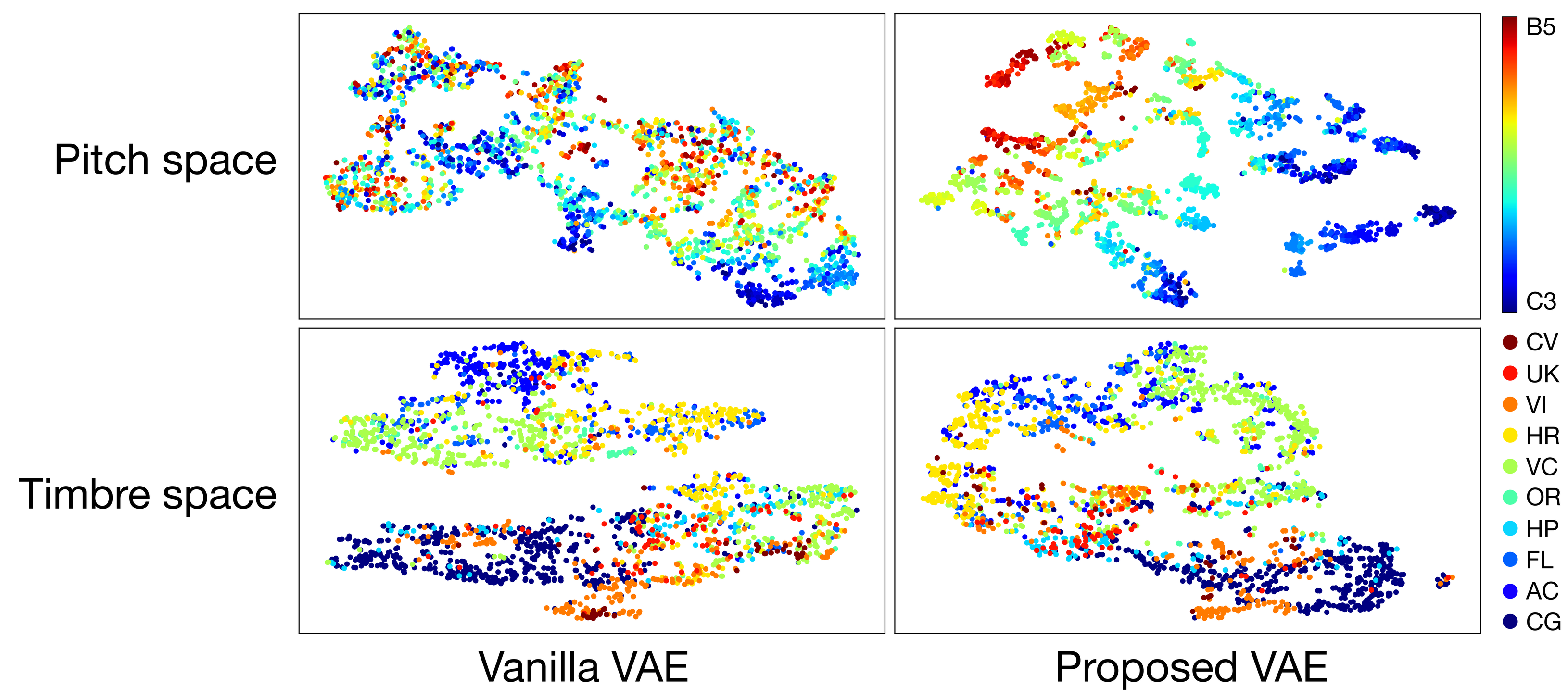
$$\frac{2}{M(M-1)}\sum_{m_1=1}^{M-1}\sum_{m_2=1}^{M}\left\|\boldsymbol{\eta}_{m_1}^{\mathrm{p}} - \boldsymbol{\eta}_{m_2}^{\mathrm{p}}\right\|$$ (Timbre in a similar way)

Result

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**

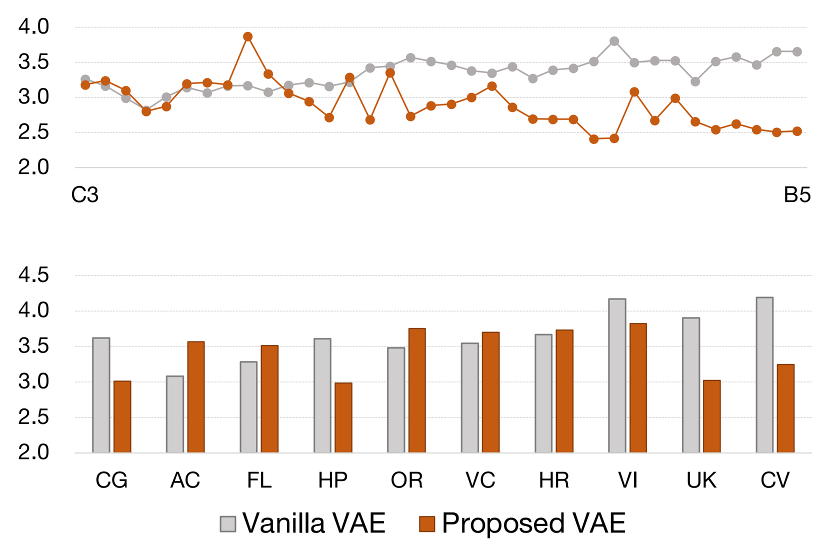
- **The denseness got smaller**, and **the divergence got larger** in both latent spaces by introducing the metric learning

Visualizations of the pitch and timbre spaces



Denseness and divergence in details

- Denseness



- Divergence