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]

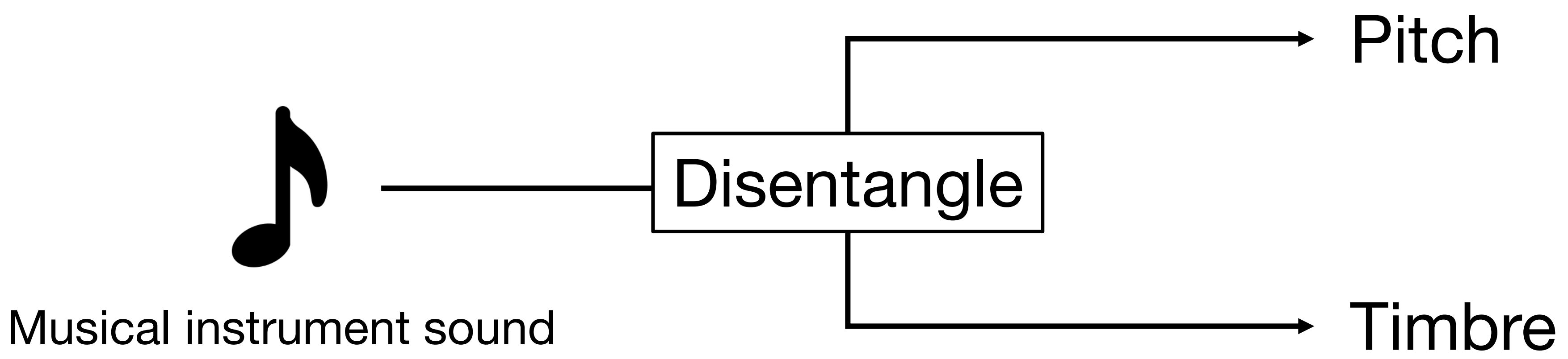
1 Waseda University 2 Kyoto University 3 National Institute of Advanced Industrial Science and Technology (AIST)

4 Waseda Research Institute for Science and Engineering, Japan

𝄞 Our Goal: Pitch-Timbre Disentanglement

Disentangle an **arbitrary** musical instrument sound into latent pitch and timbre representations

- Deal with music sounds played by any harmonic instruments
- Make the latent pitch and timbre spaces **human-interpretable**
- Introduce a metric learning technique into a VAE



What is Disentanglement?

To describe data as a combination of independent factors

- Make latent representations interpretable
- Enable us to intuitively control each factor in data generation

𝄞 What is Disentanglement?

To describe data as a combination of independent factors

- Make latent representations interpretable
- Enable us to intuitively control each factor in data generation

In the field of music information retrieval (MIR),
a sound is disentangled into the three major elements:

Volume Timbre Pitch

What is Disentanglement?

To describe data as a combination of independent factors

- Make latent representations interpretable
- Enable us to intuitively control each factor in data generation

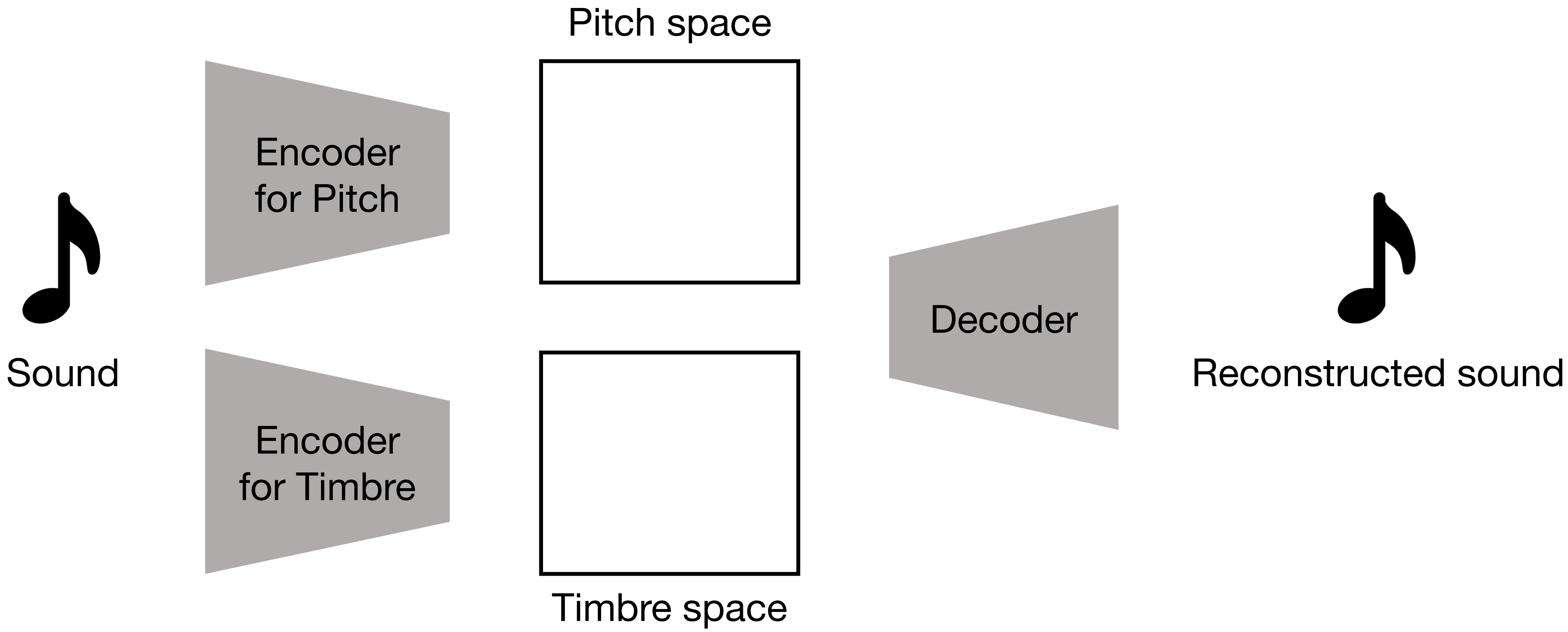
In the field of music information retrieval (MIR),
a sound is disentangled into the three major elements:

Volume — Can be computed easily

Timbre **Pitch** — **Mainly focused**

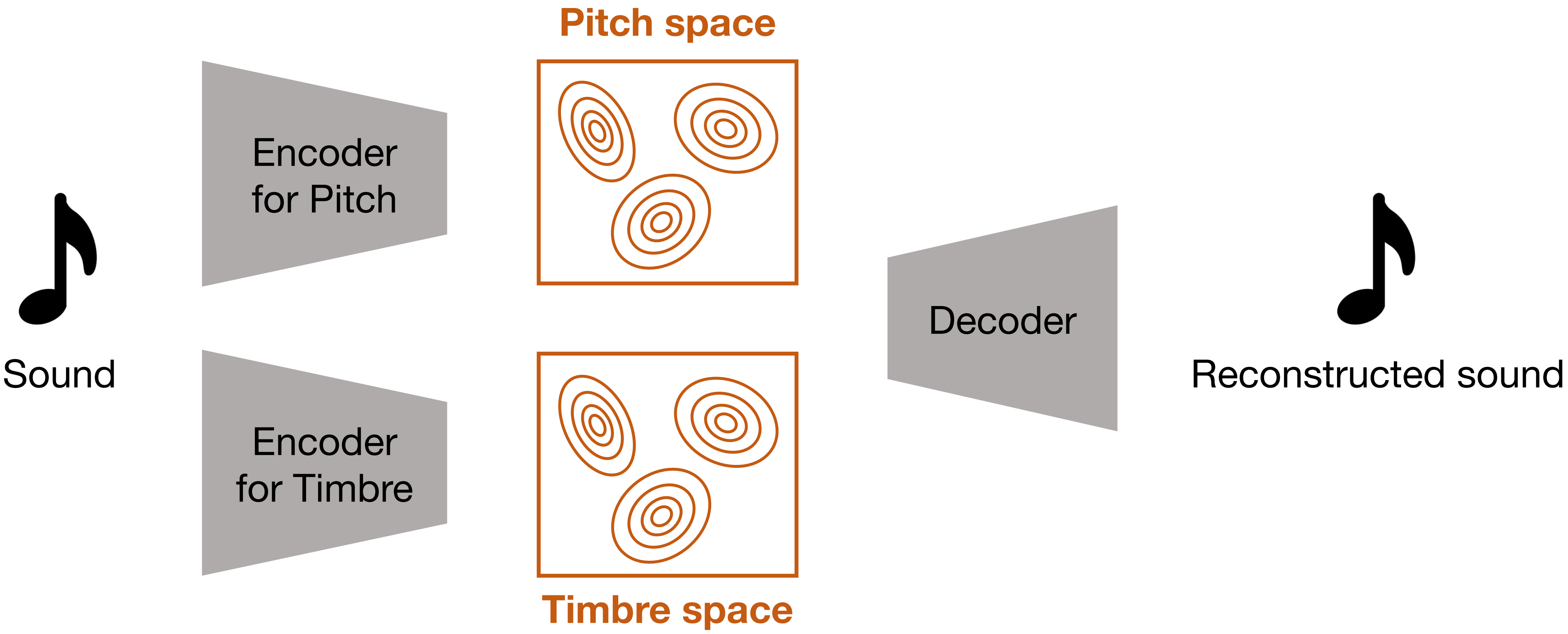
VAE for Disentanglement

A popular approach is to train a variational autoencoder (VAE), as a deep latent variable model



𝄞 Conventional Approach

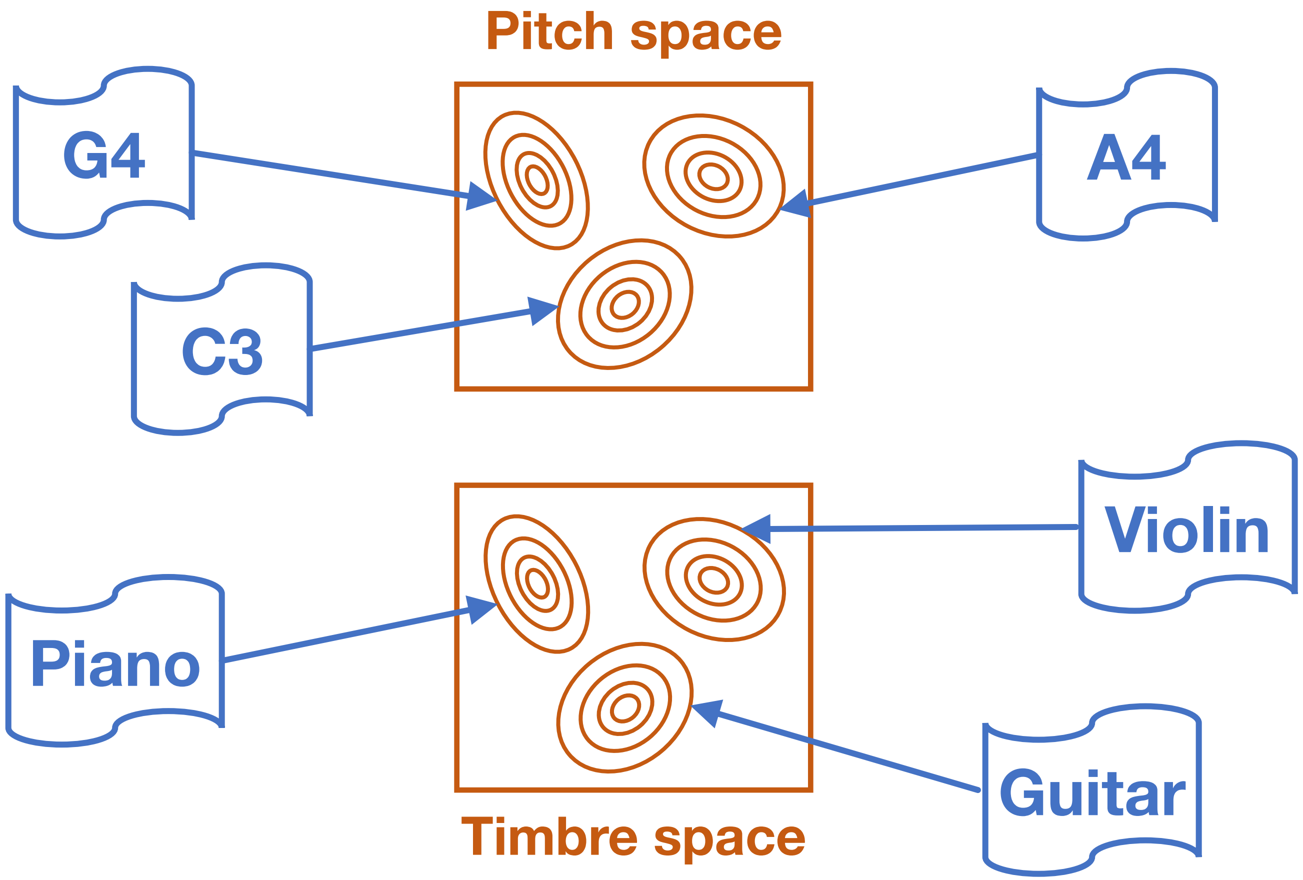
Assume the Gaussian distributions correspond to individual pitches and timbres (instruments)



Luo et al. “Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders”, ISMIR 2019.

𝄞 Conventional Approach

Use concrete category labels



Luo et al. “Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders”, ISMIR 2019.

Motivation

The conventional approach **did not aim to treat an arbitrary musical instrument sound**

- Set a **finite number** of Gaussian distributions mixtures
- Must prepare all the target labels and data beforehand
- **Cannot handle unseen pitches and timbres** that are not included in the training data

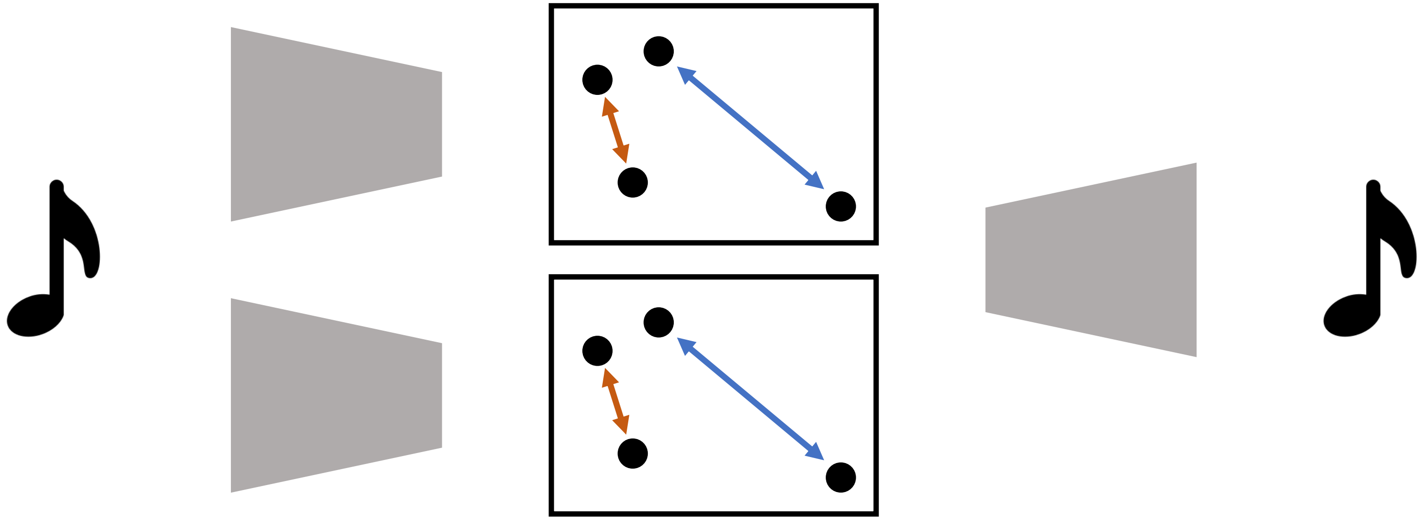
Disentanglement **without using the concrete category labels** enables to treat an **arbitrary** musical instrument sound

- Instead, use **similarities and dissimilarities** of samples

𝄞 Key Idea

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



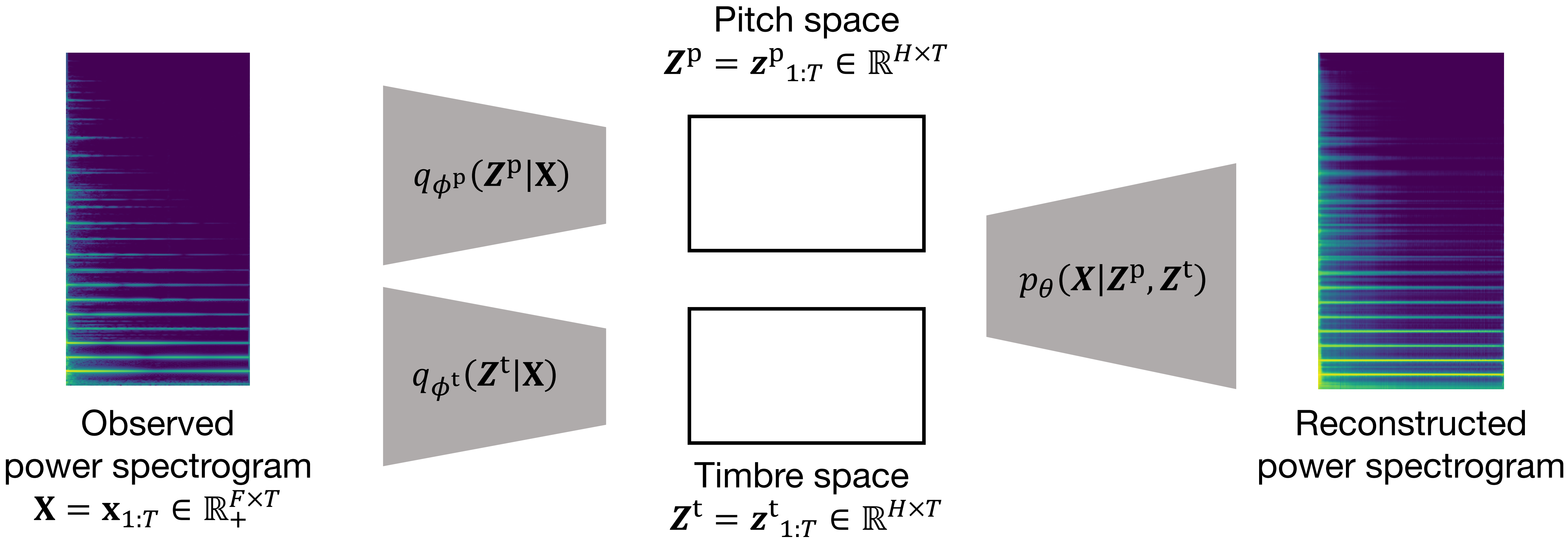
𝄞 Key Idea

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

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)

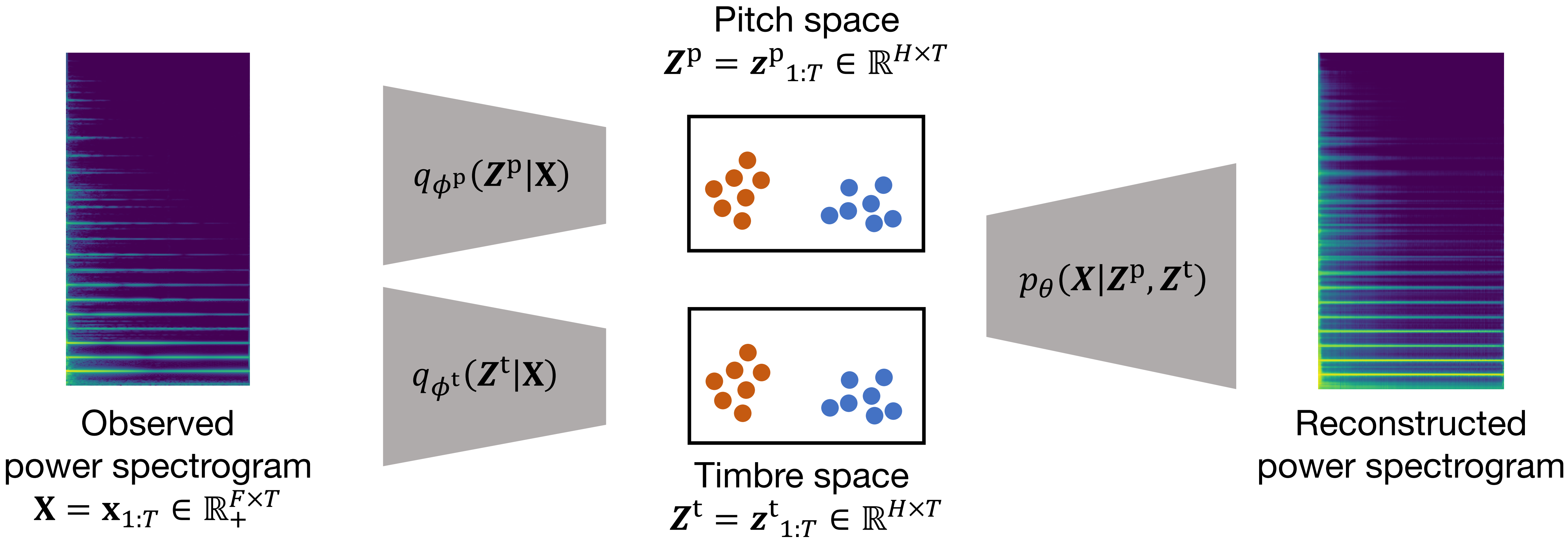
Method

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



𝄞 Method

Second, 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



Method

Third, conduct pairwise metric learning with 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)

𝄞 Method

Third, conduct pairwise metric learning with 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 $\mathbf{X}_2$)

Method

Third, conduct pairwise metric learning with 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

𝄞 Method

Third, conduct pairwise metric learning with 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

Method

Train the networks in a weakly supervised manner

- 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

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

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

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

Hyperparameters
to control the weights

Experimental Evaluation

Data

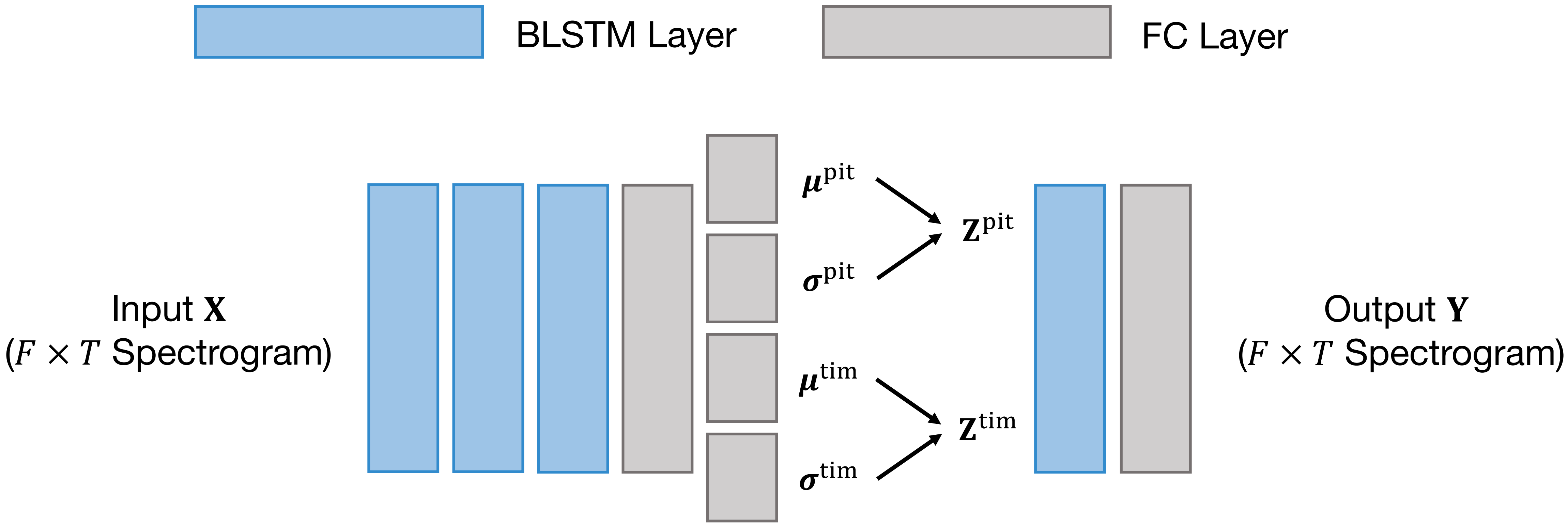
- Used instrument sounds from the RWC Music Database

Goto et al., "RWC music database: Music genre database and musical instrument sound database", ISMIR 2003.

- Excepted for Shakuhachi, Soprano, and Alto
- Selected the sounds of pitches from C3 to B5
- Split into three sets:
  - Training set (29957 sounds, 40 instruments)
  - Evaluation sets (10957 sounds, 10 instruments, 2-fold cross-validation)
- The three sets shared pitches but **did not share instruments**

Experimental Evaluation

Model Configuration



Experimental Evaluation

Evaluation Criteria

- **Denseness** (Smaller is better)
  - How **close** the latent variables with **the same pitch or timbre** label are
    - Calculated as: $\frac{1}{M}\sum_{m=1}^{M}\frac{1}{9N_m}\sum_{n=1}^{N_m}\sum_{t=1}^{9}\left\|\boldsymbol{z}_{mnt}^{\mathrm{p}}-\boldsymbol{\eta}_{m}^{\mathrm{p}}\right\|$ (Timbre in a similar way)
    - $\boldsymbol{\eta}_{m}^{\mathrm{p}}=\frac{1}{9N_m}\sum_{n=1}^{N_m}\sum_{t=1}^{9}\boldsymbol{z}_{mnt}^{\mathrm{p}}$

- **Divergence** (Larger is better)
  - 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)

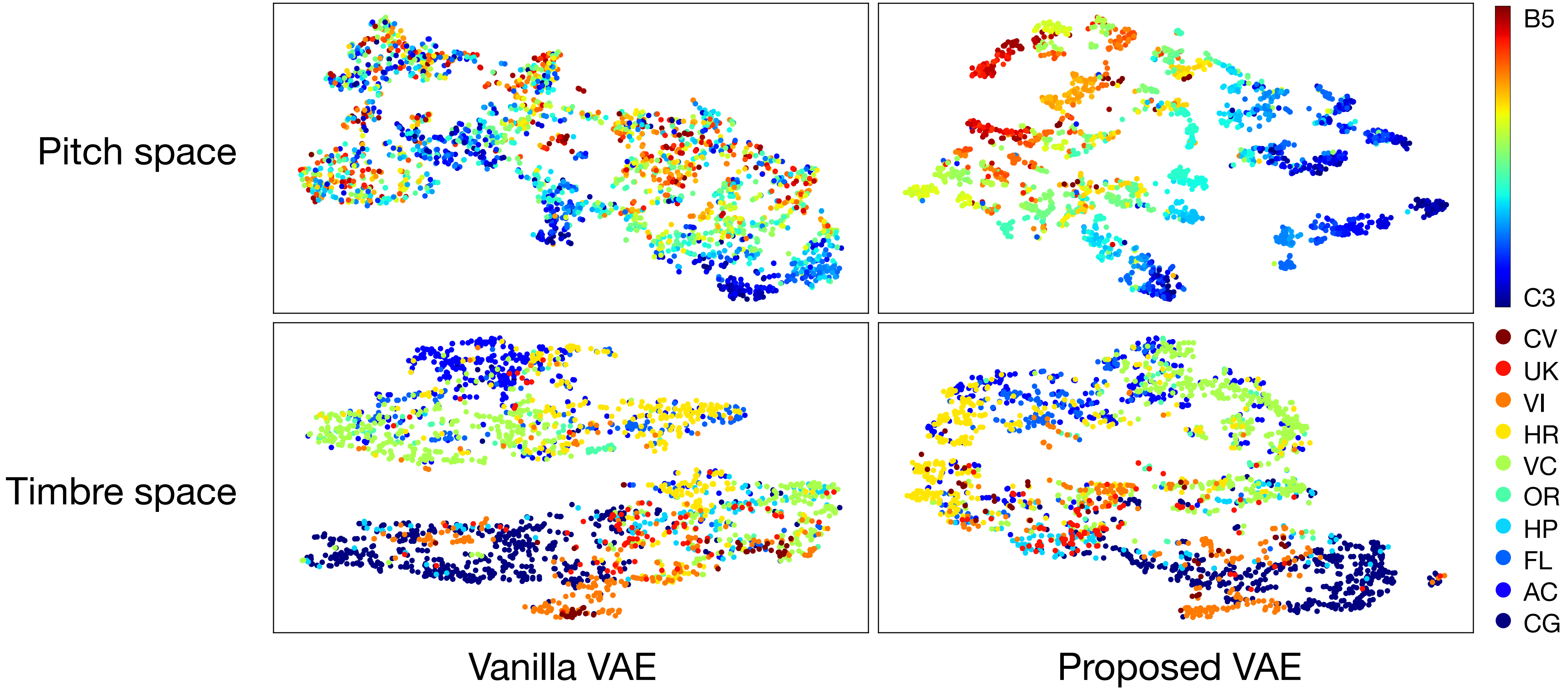
Experimental Results

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

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**

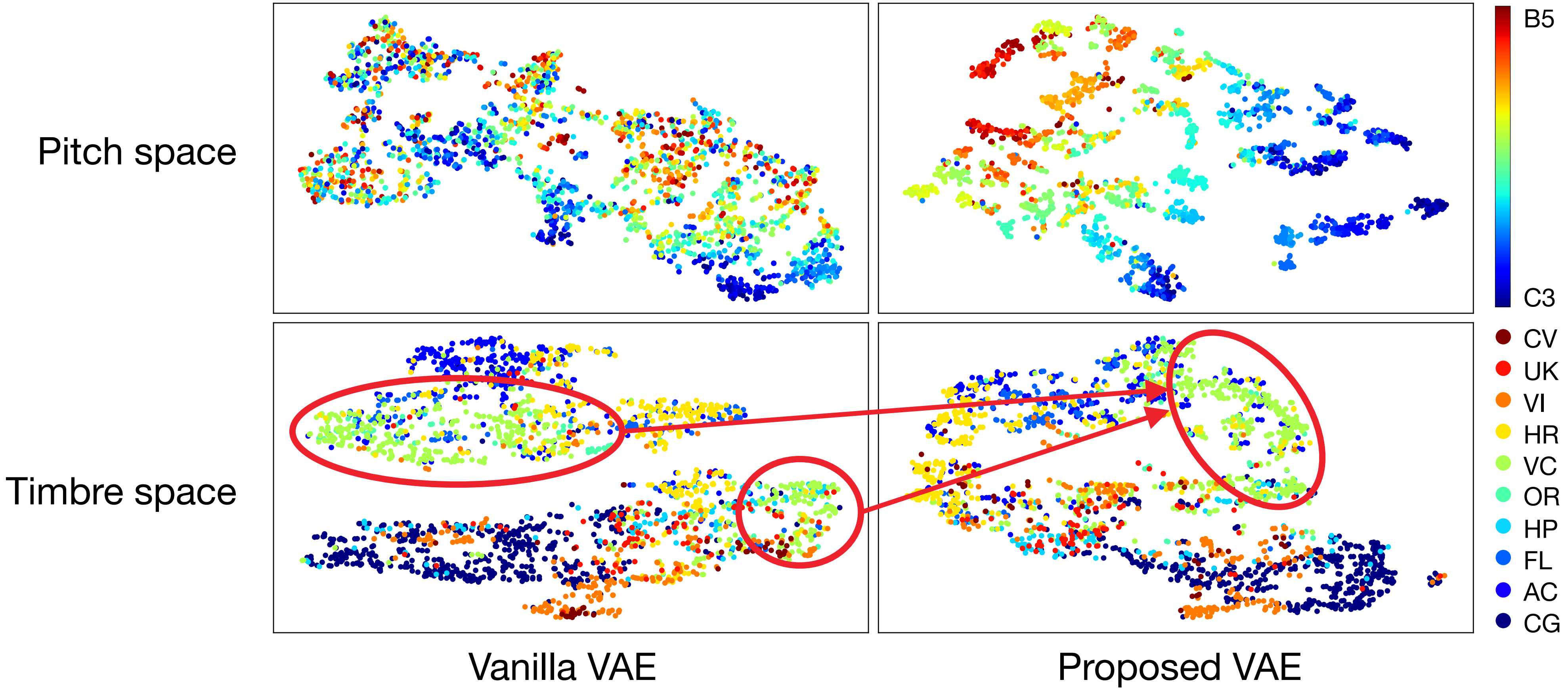
Experimental Results

Found better-structured disentangled representations with pitch and timbre clusters for unseen musical instruments



Experimental Results

Found better-structured disentangled representations with pitch and timbre clusters for unseen musical instruments

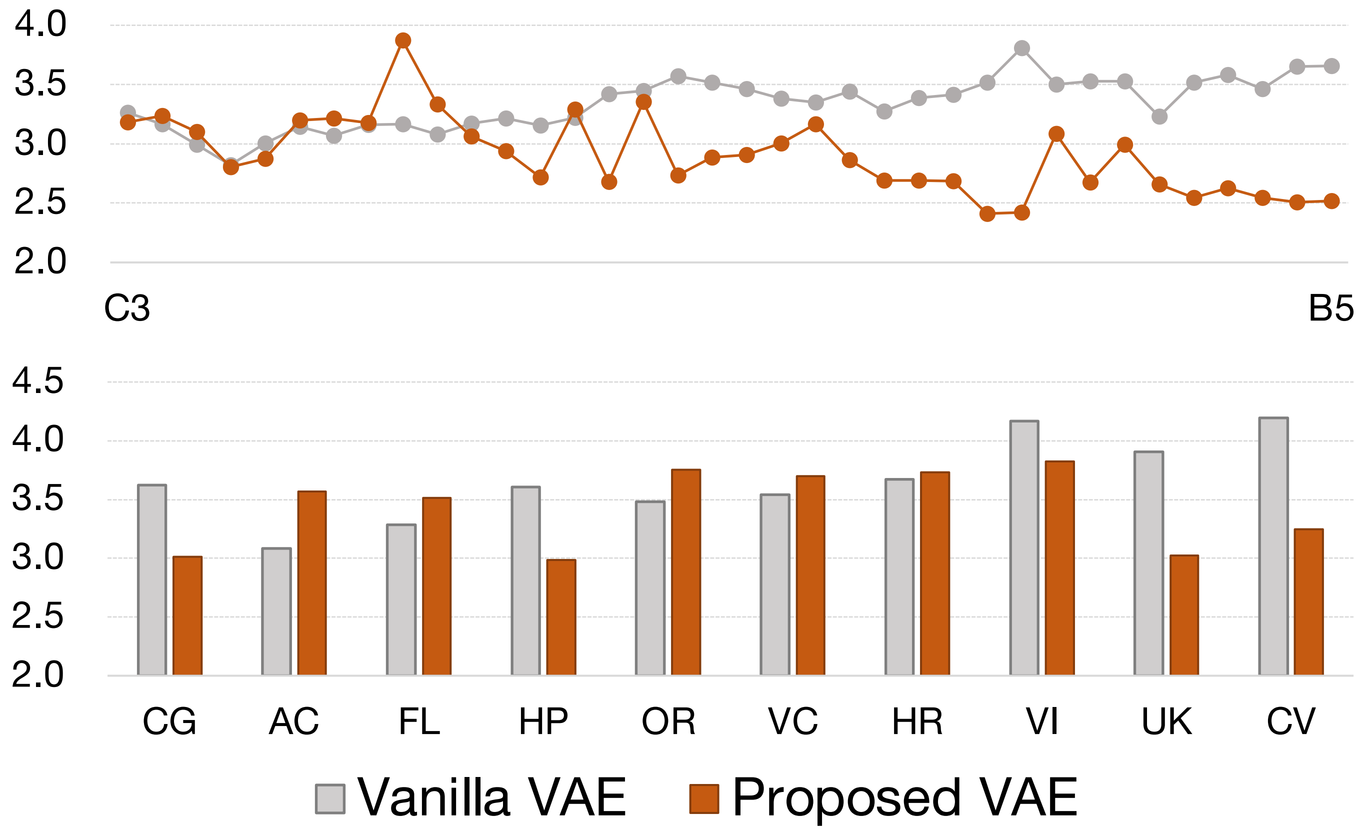


Experimental Results

Denseness: Achieved better denseness for most pitches and timbres by using the contrastive losses

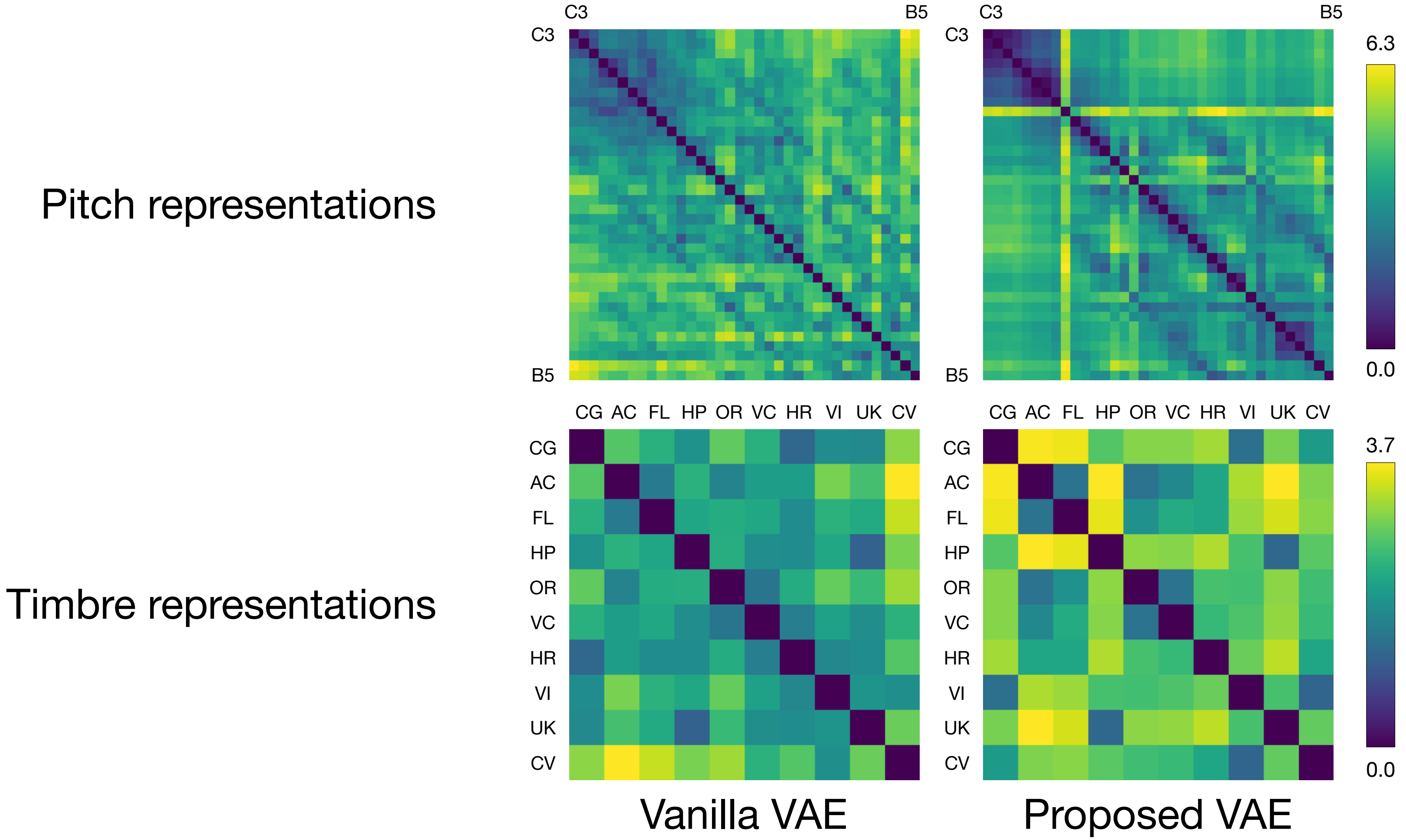
Pitch representations

Timbre representations



Experimental Results

Divergence: Succeeded in mapping the different timbres to be distant from each other



Summary

We proposed the VAE-based method for disentangling a musical instrument sound into latent pitch and timbre representations

- Deal with music sounds played by any harmonic instruments
- Make the latent pitch and timbre spaces human-interpretable
- Introduce a metric learning technique into a VAE
- Successfully disentangled the latent pitch and timbre representations compared to the vanilla VAE

Thank you for watching!