Unsupervised Disentanglement of Timbral, Pitch, and Variation Features From Musical Instrument Sounds With Random Perturbation

Keitaro Tanaka*, Yoshiaki Bando†, Kazuyoshi Yoshii‡, and Shigeo Morishima§
* Waseda University, Tokyo, Japan
E-mail: phys.keitaro1227@ruri.waseda.jp Tel/Fax: +81-3-5286-3510
† National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
‡ Kyoto University, Kyoto, Japan
§ Waseda Research Institute for Science and Engineering, Tokyo, Japan

***Abstract*—This paper describes an unsupervised disentangled representation learning method for musical instrument sounds with pitched and unpitched spectra. Since conventional methods have commonly attempted to disentangle timbral features (e.g., instruments) and pitches (e.g., MIDI note numbers and F0s), they can be applied to only pitched sounds. Global timbres unique to instruments and local variations (e.g., expressions and playstyles) are also treated without distinction. Instead, we represent the spectrogram of a musical instrument sound with a variational autoencoder (VAE) that has timbral, pitch, and variation features as latent variables. The pitch clarity or percussiveness, brightness, and F0s (if existing) are considered to be represented in the abstract pitch features. The unsupervised disentanglement is achieved by extracting time-invariant and time-varying features as global timbres and local variations from randomly pitch-shifted input sounds and time-varying features as local pitch features from randomly timbre-distorted input sounds. To enhance the disentanglement of timbral and variation features from pitch features, input sounds are separated into spectral envelopes and fine structures with cepstrum analysis. The experiments showed that the proposed method can provide effective timbral and pitch features for better musical instrument classification and pitch estimation.**

I. Introduction

Musical instrument sounds are typically considered to be characterized in terms of the three major elements of sound, i.e., timbre, pitch, and volume. These elements should not only be analyzed from musical instrument sounds but also be manipulated in an interpretable manner in various tasks such as automatic music transcription [1]–[3] and composition [4], musical instrument classification [5]–[8], timbre modification [9], and style transfer [10]. Music is generally performed with pitched instruments (e.g., strings and wind instruments), unpitched instruments (e.g., drums), and singing voices, all of which produce both pitched and unpitched sounds in reality. We thus seek a disentangled representation learning method that can deal with any kind of pitched and unpitched sounds in a unified manner.

The standard approach to representation learning of musical instrument sounds is to deal with only pitched sounds and use a variational autoencoder (VAE) [11] for disentangling the pitch and timbral features from those sounds [12]–[15]. Since the timbral characteristics of musical instruments and their temporal variations depending on expressions, playstyles, and volumes are thus represented as the timbral features in a lump, musical instrument sounds with different local variations are treated as different instruments even though those sounds are generated from the same instrument. In general, the semitone-level pitches (MIDI note numbers) or fundamental frequencies (F0s) of pitched sounds are directly used as the pitch features for interpretability. This prevents this approach from being applied to unpitched sounds.

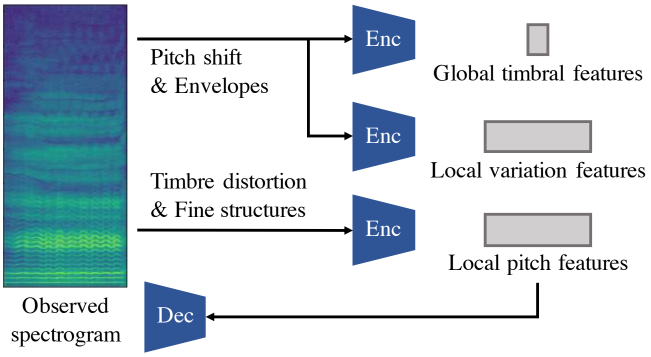

Fig. 1. Proposed VAE with pitch shift, timbre distortion, and cepstrum analysis. It learns, from musical instrument sounds, global timbral, local pitch, and local variation features in unsupervised manner.

To overcome these limitations, in this paper, we propose a VAE-based representation learning method for timbre-pitch-variation disentanglement of pitched and unpitched sounds (Fig. 1). In music performance, we decide *what instrument* to play and then dynamically control *what* and *how* to play. While the instrument (global timbre) never changes, the sound spectra vary over time according to the pitches (if existing) and timbral variations. The VAE should thus be capable of inferring three kinds of abstract latent features, i.e., global (time-invariant) timbral features, local (time-varying) pitch features, and local variation features from musical instrument sounds, where the timbral features represent the information of

978-616-590-477-3 ©2022 APSIPA

instruments or sources and the variation features represent expressions, playstyles, and volumes. The continuous local pitch features abstractly represent not only fundamental frequencies but also pitch-related characteristics including pitch clarity or percussiveness and brightness. This makes it possible to deal with both pitched and unpitched sounds.

Our VAE consists of three cooperative encoders and a decoder. The encoders infer from a given sound spectrogram the global timbral features, the local variations, and the local pitch features in this order. The decoder aims to reconstruct the original spectrogram from the latent features. The key to disentangled representation learning is to introduce random perturbation with pitch shift and timbre distortion. We further enhance the disentanglement by cepstrum-domain spectrum separation. Specifically, the timbral features are inferred first, and the variations are then inferred framewise in a timbre-conditioned manner. To avoid extracting pitch-related information, both features are inferred from the spectral envelopes extracted from a randomly pitch-shifted version of the original sound spectrogram. In contrast, the pitch features are inferred framewise from the fine (harmonic) structures extracted from a randomly timbre-distorted version of the sound spectrogram.

The main contribution of this study is to represent the timbre and volume of sound in terms of global timbral and local variation features. This representation mitigates the ill-definition in conventional pitch-timbre disentanglement. The latent timbral, pitch, and variation features can be learned from any musical instrument sounds, not limited to pitched sounds in an unsupervised manner. This is achieved by introducing random perturbation and classical signal processing techniques into analysis-and-synthesis formalism based on a VAE. Experimental results show that this approach promotes three-factor disentanglement and provides effective timbral and pitch features for better musical instrument classification and pitch estimation.

II. Related Work

This section reviews recent progress in disentangled representation learning of musical instrument sound. We also glance at disentangled representation in differentiable digital signal processing (DDSP) and speech processing.

A. Disentanglement of Musical Instrument Sounds

Generative models are major tools for disentangled representation learning. In music information retrieval, the disentanglement of the pitch and timbre of music signals has been tackled using autoencoders (AEs) and VAEs. Such disentanglement was first investigated by Mor et al. [16]. They proposed an AE-based music translation method that can change the timbral characteristics of music signals without changing their pitch characteristics. Bitton et al. [17] and Esling et al. [18] then proposed a β-VAE that can handle many sounds in one model and a VAE where the latent timbre space has a human perception-like metric in it, respectively. Those studies dealt with pitch-annotated music signals, while Hung et al. [19] first explored the disentangled representations of pitch and timbre for music style transfer using encoder-decoder structures.

TABLE I
Comparison of existing methods and our method.

Method	Input	Timbre	Pitch	Variation
Luo [12]	Spectrogram	Global	Global	—
Luo [13]	Spectrum	Global	Global	—
Tanaka [15]	Spectrogram	Local	Local	—
Luo [14]	Spectrogram	Global	Local	—
Ours	Spectrogram	Global	Local	Local

The works most related to our study are the series of studies by Luo et al. [12]–[14] and Tanaka et al. [15]. They aimed to disentangle the pitch and timbre of a musical instrument sound. Note that the timbre is assumed to represent an instrument here. We summarize the series in Table I, focusing on the differences in format among their input and latent features. They have commonly attempted to represent an instrument sound using two kinds of disentangled features. Such disentanglement methods, however, treat different temporal characteristics (e.g., expressions and playstyles) of the same instrument as different instruments. In contrast, we consider conventional timbres consisting of global and local components and provide new features, i.e., variations. The variations should include the volumes because the volumes are closely related to expressions. Thus, our method can also cover all three elements of sound.

B. Disentangled Representation in DDSP

Differentiable digital signal processing (DDSP) [20]–[22] has recently been gathering attention because of its high-fidelity sound synthesis. It synthesizes musical instrument sounds from the fundamental frequencies, timbral features, and loudnesses. It can also infer these elements inversely from the observed sound. At this point, we may regard it as disentangled representations. However, DDSP limits its scope to pitched harmonic sounds with concrete fundamental frequencies because of its architecture and cannot handle unpitched percussive sounds. It does not have global timbral features either because it considers all three representations as time-varying features.

C. Disentangled Representation in Speech Processing

From a technical point of view, our work is closely related to the disentanglement of speech into multiple factors [23]–[25]. Qian et al. [23] decomposed speech signals into rhythms, pitches, timbres, and contents via three information bottlenecks. Choi et al. [24] did so into linguistic, pitch, speaker, and energy information by input waveform perturbation. Most recently, Du et al. [25] disentangled the signals into emotional styles and speaker identities. However, our research distinguishes itself from those in that the target disentangled features and the perturbation methods are derived from the generation process and properties specific to music performances and instrument sounds.

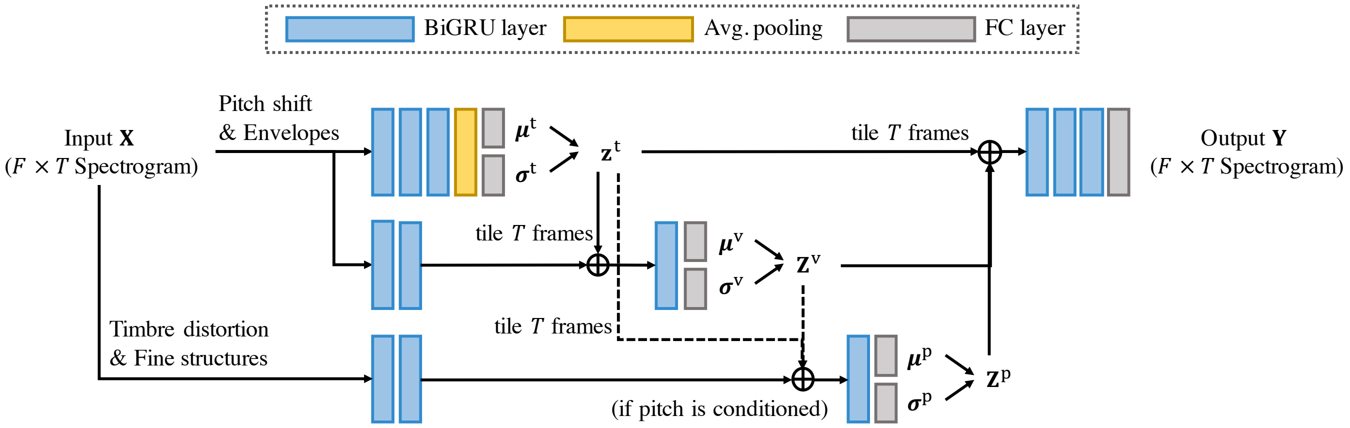
BiGRU layer
Avg. pooling
FC layer
Input **X**
($F \times T$ Spectrogram)
Pitch shift
& Envelopes
$\boldsymbol{\mu}^{\mathrm{t}}$
$\boldsymbol{\sigma}^{\mathrm{t}}$
$\mathbf{z}^{\mathrm{t}}$
tile T frames
Output **Y**
($F \times T$ Spectrogram)
tile T frames
$\boldsymbol{\mu}^{\mathrm{v}}$
$\boldsymbol{\sigma}^{\mathrm{v}}$
$\mathbf{Z}^{\mathrm{v}}$
Timbre distortion
& Fine structures
tile T frames
$\boldsymbol{\mu}^{\mathrm{p}}$
$\boldsymbol{\sigma}^{\mathrm{p}}$
$\mathbf{Z}^{\mathrm{p}}$
(if pitch is conditioned)

Fig. 2. Implementation of proposed VAE, consisting of three encoders to infer latent variables and decoder to generate sounds. $\oplus$ represents concatenation of multiple tensors. $\mathbf{z}^{\mathrm{t}}$, $\mathbf{Z}^{\mathrm{v}}$, and $\mathbf{Z}^{\mathrm{p}}$ are sampled probabilistically with reparameterization trick during training and given deterministically in inference. Pitch shift and timbre distortion are applied only during training.

III. Proposed Method

Our method trains a VAE to disentangle a musical instrument sound into global (time-invariant) timbral features and local (time-varying) pitch and variation features in an unsupervised manner (Fig. 2). Let $\mathbf{x}_{1:N}$ be a matrix consisting of N vectors $\mathbf{x}_n$ $(n = 1, \dots, N)$. This VAE is trained to represent a log-amplitude spectrogram $\mathbf{X} \triangleq \mathbf{x}_{1:T} \in \mathbb{R}^{F\times T}$ of an isolated musical instrument sound, where F and T are frequency and time-frame indices, respectively. The proposed training forces the network to estimate latent representations $\mathbf{Z} \triangleq \{\mathbf{z}^{\mathrm{t}}, \mathbf{z}^{\mathrm{v}}_{1:T}, \mathbf{z}^{\mathrm{p}}_{1:T}\}$ consisting of global timbres $\mathbf{z}^{\mathrm{t}} \in \mathbb{R}^{D^{\mathrm{t}}}$, local variations $\mathbf{Z}^{\mathrm{v}} \triangleq \mathbf{z}^{\mathrm{v}}_{1:T} \in \mathbb{R}^{D^{\mathrm{v}}\times T}$, and local pitch features $\mathbf{Z}^{\mathrm{p}} \triangleq \mathbf{z}^{\mathrm{p}}_{1:T} \in \mathbb{R}^{D^{\mathrm{p}}\times T}$ from the observation, where D^* ($*$ represents "t," "v," or "p") is the dimension of the latent space.

A. Generative Model

Following [12]–[14], we first formulate a generative process of an observed log-amplitude spectrogram $\mathbf{X}$. Our generative model represents each time-frequency bin $x_{ft} \in \mathbb{R}$ of $\mathbf{X}$ as a Gaussian distribution with latent variables $\mathbf{Z}$:

$$x_{ft} \sim \mathcal{N}(x_{ft} | \mu_{\theta,ft}(\mathbf{z}^{\mathrm{t}}, \mathbf{z}^{\mathrm{v}}_{1:T}, \mathbf{z}^{\mathrm{p}}_{1:T}), \sigma^2), \tag{1}$$

where $\mu_{\theta,ft}(\mathbf{z}^{\mathrm{t}}, \mathbf{z}^{\mathrm{v}}_{1:T}, \mathbf{z}^{\mathrm{p}}_{1:T}) \in \mathbb{R}$ is the output of a DNN (decoder) with parameters θ, and $\sigma^2 \in \mathbb{R}_+$ is a hyperparameter representing the variance of the spectrogram. We assume that each of the latent features follows a standard Gaussian distribution:

$$\mathbf{z}^{\mathrm{t}} \sim \mathcal{N}(\mathbf{z}^{\mathrm{t}} \mid \mathbf{0}_{D^{\mathrm{t}}}, \mathbf{I}_{D^{\mathrm{t}}}), \tag{2}$$

$$\mathbf{z}^{\mathrm{v}}_{1:T} \sim \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^{\mathrm{v}}_t \mid \mathbf{0}_{D^{\mathrm{v}}}, \mathbf{I}_{D^{\mathrm{v}}}), \tag{3}$$

$$\mathbf{z}^{\mathrm{p}}_{1:T} \sim \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^{\mathrm{p}}_t \mid \mathbf{0}_{D^{\mathrm{p}}}, \mathbf{I}_{D^{\mathrm{p}}}), \tag{4}$$

where $\mathbf{0}_{D^*}$ is an all-zero vector of size D^*, and $\mathbf{I}_{D^*}$ is an identity matrix of size $D^* \times D^*$. These priors make each dimension of latent variables have independent features [26] with good interpretability. This generative model itself, however, does not guarantee that each of the three latent features corresponds to the timbre, variation, and pitch features. We thus train the model with random perturbation in an unsupervised manner.

B. Variational Inference for Unsupervised Training

Our goal is to train the DNN to maximize the log-marginal likelihood $\log p_\theta(\mathbf{X})$. Because the DNN-based formulation of our generative model makes $\log p_\theta(\mathbf{X})$ intractable, we introduce an encoder network with parameters ϕ representing $q_\phi(\mathbf{Z}|\mathbf{X})$ to approximately calculate the log-marginal likelihood [11]. Specifically, we train the encoder and decoder networks to maximize the following lower bound of the log-marginal likelihood $\mathcal{L}$:

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})}[\log p_\theta(\mathbf{X}|\mathbf{Z})] - \mathcal{D}_{\mathrm{KL}}(q_\phi(\mathbf{Z}|\mathbf{X}) || p(\mathbf{Z})), \tag{5}$$

where $\mathcal{D}_{\mathrm{KL}}(\cdot||\cdot)$ represents the Kullback-Leibler (KL) divergence. The parameters of the decoder θ are inferred in a maximum-likelihood sense and those of the encoders ϕ are optimized to minimize the KL divergence from $q_\phi(\mathbf{Z}|\mathbf{X})$ to $p_\theta(\mathbf{Z}|\mathbf{X}) \propto p_\theta(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})$. The lower bound can be calculated analytically with Monte-Carlo approximation as in the training of VAEs [11] and optimized with gradient ascent. In this paper, we consider two types of encoding architectures: the timbre-variation conditional pitch model and the timbre-variation independent pitch model.

1) Timbre-Variation Conditional Pitch Model: In this model, the pitch features are conditioned by the timbral and variation features:

$$q_\phi(\mathbf{Z}|\mathbf{X}) = q_{\phi^{\mathrm{t}}}(\mathbf{z}^{\mathrm{t}}|\mathbf{X}) q_{\phi^{\mathrm{v}}}(\mathbf{z}^{\mathrm{v}}_{1:T}|\mathbf{X}, \mathbf{z}^{\mathrm{t}}) q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\mathbf{X}, \mathbf{Z}^{\mathrm{t,v}}), \tag{6}$$

where $\mathbf{Z}^{\mathrm{t,v}} \triangleq \{\mathbf{z}^{\mathrm{t}}, \mathbf{z}^{\mathrm{v}}_{1:T}\}$. Each posterior is given by

$$q_{\phi^{\mathrm{t}}}(\mathbf{z}^{\mathrm{t}}|\mathbf{X}) = \mathcal{N}(\mathbf{z}^{\mathrm{t}}|\boldsymbol{\mu}^{\mathrm{t}}_{\phi^{\mathrm{t}}}(\mathbf{X}), \mathrm{diag}(\boldsymbol{\sigma}^{\mathrm{t}2}_{\phi^{\mathrm{t}}}(\mathbf{X}))), \tag{7}$$

$$q_{\phi^{\mathrm{v}}}(\mathbf{z}^{\mathrm{v}}_{1:T}|\mathbf{X}, \mathbf{z}^{\mathrm{t}}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^{\mathrm{v}}_t|[\boldsymbol{\mu}^{\mathrm{v}}_{\phi^{\mathrm{v}}}(\mathbf{X}, \mathbf{z}^{\mathrm{t}})]_t, \mathrm{diag}([\boldsymbol{\sigma}^{\mathrm{v}2}_{\phi^{\mathrm{v}}}(\mathbf{X}, \mathbf{z}^{\mathrm{t}})]_t)), \tag{8}$$

$$q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\mathbf{X}, \mathbf{Z}^{\mathrm{t,v}}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^{\mathrm{p}}_t|[\boldsymbol{\mu}^{\mathrm{p}}_{\phi^{\mathrm{p}}}(\mathbf{X},\mathbf{Z}^{\mathrm{t,v}})]_t, \mathrm{diag}([\boldsymbol{\sigma}^{\mathrm{p}2}_{\phi^{\mathrm{p}}}(\mathbf{X},\mathbf{Z}^{\mathrm{t,v}})]_t)), \tag{9}$$

where $\boldsymbol{\mu}^{\mathrm{t}}_{\phi^{\mathrm{t}}}(\mathbf{X})$ and $\boldsymbol{\sigma}^{\mathrm{t}2}_{\phi^{\mathrm{t}}}(\mathbf{X})$ are the D^{t}-dimensional outputs of the DNN with parameters ϕ^{t}, $\boldsymbol{\mu}^{\mathrm{v}}_{\phi^{\mathrm{v}}}(\mathbf{X}, \mathbf{z}^{\mathrm{t}})$ and $\boldsymbol{\sigma}^{\mathrm{v}2}_{\phi^{\mathrm{v}}}(\mathbf{X}, \mathbf{z}^{\mathrm{t}})$ are the $D^{\mathrm{v}}T$-dimensional outputs of the DNN with parameters ϕ^{v}, and $\boldsymbol{\mu}^{\mathrm{p}}_{\phi^{\mathrm{p}}}(\mathbf{X}, \mathbf{Z}^{\mathrm{t,v}})$ and $\boldsymbol{\sigma}^{\mathrm{p}2}_{\phi^{\mathrm{p}}}(\mathbf{X}, \mathbf{Z}^{\mathrm{t,v}})$ are the $D^{\mathrm{p}}T$-dimensional outputs of the DNN with parameters ϕ^{p}. The notation $[\mathbf{A}]_t$ indicates the t-th time-frame of $\mathbf{A}$. We approximately calculate the expectation term of $\mathcal{L}$ with the reparameterization trick [11] in the timbre-variation conditional pitch model as follows:

$$\mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})}[\log p_\theta(\mathbf{X}|\mathbf{Z})] \approx -\frac{1}{2}\sum_{f,t=1}^{F,T}\left\{\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(x_{ft} - \tilde{y}_{ft})^2\right\}, \tag{10}$$

where $\tilde{y}_{ft} = \mu_{\theta,ft}(\tilde{\mathbf{z}}^{\mathrm{t}}, \tilde{\mathbf{z}}^{\mathrm{v}}_{1:T}, \tilde{\mathbf{z}}^{\mathrm{p}}_{1:T})$ is the spectrogram that is reconstructed with the samples $\tilde{\mathbf{z}}^{\mathrm{t}}$, $\tilde{\mathbf{z}}^{\mathrm{v}}_{1:T}$, and $\tilde{\mathbf{z}}^{\mathrm{p}}_{1:T}$ from the variational posterior q. These samples are obtained by the following ancestral sampling:

$$\tilde{\mathbf{z}}^{\mathrm{t}} \sim q_{\phi^{\mathrm{t}}}(\mathbf{z}^{\mathrm{t}}|\mathbf{X}), \tag{11}$$

$$\tilde{\mathbf{z}}^{\mathrm{v}}_{1:T} \sim q_{\phi^{\mathrm{v}}}(\mathbf{z}^{\mathrm{v}}_{1:T}|\mathbf{X}, \tilde{\mathbf{z}}^{\mathrm{t}}), \tag{12}$$

$$\tilde{\mathbf{z}}^{\mathrm{p}}_{1:T} \sim q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\mathbf{X}, \tilde{\mathbf{z}}^{\mathrm{t}}, \tilde{\mathbf{z}}^{\mathrm{v}}_{1:T}). \tag{13}$$

2) Timbre-Variation Independent Pitch Model: In this model, we consider the pitch features to be independent of the timbral and variation features. Thus, the formulation of $q_\phi(\mathbf{Z}|\mathbf{X})$ is given as

$$q_\phi(\mathbf{Z}|\mathbf{X}) = q_{\phi^{\mathrm{t}}}(\mathbf{z}^{\mathrm{t}}|\mathbf{X}) q_{\phi^{\mathrm{v}}}(\mathbf{z}^{\mathrm{v}}_{1:T}|\mathbf{X}, \mathbf{z}^{\mathrm{t}}) q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\mathbf{X}), \tag{14}$$

$$q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\mathbf{X}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^{\mathrm{p}}_t|[\boldsymbol{\mu}^{\mathrm{p}}_{\phi^{\mathrm{p}}}(\mathbf{X})]_t, \mathrm{diag}([\boldsymbol{\sigma}^{\mathrm{p}2}_{\phi^{\mathrm{p}}}(\mathbf{X})]_t)), \tag{15}$$

where $\boldsymbol{\mu}^{\mathrm{p}}_{\phi^{\mathrm{p}}}(\mathbf{X})$ and $\boldsymbol{\sigma}^{\mathrm{p}2}_{\phi^{\mathrm{p}}}(\mathbf{X})$ are the $D^{\mathrm{p}}T$-dimensional outputs of the DNN with parameters ϕ^{p}, and the other terms are the same as (7) and (8). We calculate the expectation term of $\mathcal{L}$ as in (10) with the following samples:

$$\tilde{\mathbf{z}}^{\mathrm{t}} \sim q_{\phi^{\mathrm{t}}}(\mathbf{z}^{\mathrm{t}}|\mathbf{X}), \tag{16}$$

$$\tilde{\mathbf{z}}^{\mathrm{v}}_{1:T} \sim q_{\phi^{\mathrm{v}}}(\mathbf{z}^{\mathrm{v}}_{1:T}|\mathbf{X}, \tilde{\mathbf{z}}^{\mathrm{t}}), \tag{17}$$

$$\tilde{\mathbf{z}}^{\mathrm{p}}_{1:T} \sim q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\mathbf{X}). \tag{18}$$

C. Random Perturbation for Disentanglement

Our encoders transform the observed spectrogram $\mathbf{X}$ into the global timbral representations $\mathbf{z}^{\mathrm{t}}$, local variation representations $\mathbf{z}^{\mathrm{v}}_{1:T}$, and local pitch representations $\mathbf{z}^{\mathrm{p}}_{1:T}$. We can see that our model already has an inductive bias for disentanglement in terms of the global and local features in its formulation. The obtained representations, however, are not disentangled in terms of the timbral/variation and pitch features because all the encoders $q_{\phi^*}(\cdot)$ receive the same input feature $\mathbf{X}$, which has both the original timbral/variation and pitch contents. Since the VAE is trained such that the decoder can reconstruct an observation from the latent features extracted by the encoders, the encoders make the latent features keep as much information as they can. Thus, the timbral/variation and pitch contents naturally leak to the other latent spaces.

To make the latent features disentangled as much as possible, we introduce random perturbation techniques. We also enhance the disentanglement by cepstrum-domain spectrum separation. Specifically, we modify the inputs of the encoders to two types of partially-randomized spectrograms $\boldsymbol{f}(\mathbf{X})$ and $\boldsymbol{g}(\mathbf{X})$ as follows:

$$q_{\phi^{\mathrm{t}}}(\mathbf{z}^{\mathrm{t}}|\mathbf{X}) \rightarrow q_{\phi^{\mathrm{t}}}(\mathbf{z}^{\mathrm{t}}|\boldsymbol{f}(\mathbf{X})), \tag{19}$$

$$q_{\phi^{\mathrm{v}}}(\mathbf{z}^{\mathrm{v}}_{1:T}|\mathbf{X}, \mathbf{z}^{\mathrm{t}}) \rightarrow q_{\phi^{\mathrm{v}}}(\mathbf{z}^{\mathrm{v}}_{1:T}|\boldsymbol{f}(\mathbf{X}), \mathbf{z}^{\mathrm{t}}), \tag{20}$$

$$q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\mathbf{X}, \mathbf{Z}^{\mathrm{t,v}}) \rightarrow q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\boldsymbol{g}(\mathbf{X}), \mathbf{Z}^{\mathrm{t,v}}), \tag{21}$$

$$q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\mathbf{X}) \rightarrow q_{\phi^{\mathrm{p}}}(\mathbf{z}^{\mathrm{p}}_{1:T}|\boldsymbol{g}(\mathbf{X})), \tag{22}$$

where $\boldsymbol{f} : \mathbb{R}^{F\times T} \rightarrow \mathbb{R}^{F\times T}$ is a function that randomly shifts the pitches of $\mathbf{X}$ and then extracts its envelopes, and $\boldsymbol{g} : \mathbb{R}^{F\times T} \rightarrow \mathbb{R}^{F\times T}$ is a function that randomly distorts the timbres/variations of $\mathbf{X}$ and then extracts its fine structures. The envelopes and fine structures correspond to low and high quefrency regions of the cepstrum, respectively.

The random pitch shift and timbre distortion of the observed spectrogram, on the one hand, make the corresponding representations robust to the changed aspect of data. This is because if the randomly changed features are extracted by the encoders and used in reconstruction, they deteriorate the likelihood. The cepstrum analysis techniques, on the other hand, directly shut out the undesired information because the timbral and variation characteristics of a sound are present in the low quefrency regions, while the pitch characteristics are in the high quefrency regions [27]. These perturbations can be introduced without any label, and thus, we can train the VAE in an unsupervised manner.

IV. Evaluation

This section describes experiments conducted to evaluate the performance of the proposed method in terms of the disentanglement of timbral, pitch, and variation features.

A. Data

To evaluate the proposed method, we used musical instrument sounds from the RWC Music Database [28]. Each file is annotated with an instrument name and records the entire range of pitches that can be produced by the instrument at semitone intervals. We automatically split each file into sounds with individual pitches by silence detection and removed the silence regions at the beginning of each split sound by onset detection. Out of all 88,889 obtained files, we selected 62,704 files within the pitch range from A0 to C8 as the pitched sounds and 2,970 files played by percussive instruments as the unpitched percussive sounds. For evaluation, we randomly split the pitched sounds into three sets: a training set (43,892 files), a validation set (9,406 files), and a test set (9,406 files). Similarly, we split all selected sounds into three sets: a training set (45,972 files), a validation set (9,851 files), and a test set (9,851 files).

All sounds were sampled at 44.1 kHz, and we used only up to the first two seconds of each sound. During training, we randomly shifted the pitch and distorted the timbre of each sound on the fly. Specifically, the pitch shift was conducted with L semitones ($-7 \leq L \leq 7$), where L was selected randomly. The timbre distortion was performed using pedalboard[1], a Python library by Spotify. We randomly applied two of the nine presets (Chorus, Distortion, Phaser, LadderFilter, HighpassFilter, LowpassFilter, Reverb, GSMFullRateCompressor, and Bitcrush) to the sound. We used a short-time Fourier transform (STFT) with a Hann window of 4,096 samples and a shifting interval of 441 samples (10 ms) to obtain spectrograms with shapes of $F = 2049$ and $T \leq 201$. Each spectrogram was separated into its envelopes and its fine structures by liftering its real cepstrum, where the cut-off position was set to the 100th coefficient. We normalized the spectrograms such that the average amplitude of each spectrogram was one. We used the Librosa library [29] in our implementation.

B. Model Configuration

Our VAE-based method utilized the bidirectional gated recurrent unit (BiGRU) architecture for its encoders and decoder to capture the temporal characteristics of sounds, as shown in Fig. 2. All the BiGRU layers used in the model had 2×800 cells. The envelopes of the pitch-shifted sounds were fed into the encoder for global timbres and the encoder for local variations. The encoder for global timbres consisted of three layers of BiGRUs, an average pooling layer along the time-frame axis, and fully connected (FC) layers. Two FC layers independently transformed 1,600 dimensions into $D^{\mathrm{t}} = 64$ dimensions to represent the means and variances of the latent variables. The encoder for the local variations consisted of three layers of BiGRUs and FC layers. The global timbres were tiled along the time-frame axis and fed into the last BiGRU layer concatenated with the outputs of the second BiGRU layer along the spatial axis. Two FC layers were the same as those of the timbres (i.e., $D^{\mathrm{v}} = 64$).

We considered two types of encoders for the local pitch features, as we described in III-B. The architecture of the first one for the timbre-variation conditional pitch model was quite similar to that of the encoder for variations. The last BiGRU layer took as input the tiled global timbres and local variations concatenated with the outputs of the second BiGRU layer along the spatial axis. The architecture of the second one for the timbre-variation independent pitch model, in contrast, did not use the timbres and variations. Two FC layers independently transformed 1,600 dimensions into $D^{\mathrm{p}} = 32$ dimensions to represent the means and variances of the latent variables in both models. The decoder consisted of three layers of BiGRUs and an FC layer. The tiled global timbral, local variational, and local pitch features were all concatenated along the spatial axis and fed into the decoder. The three FC layers that represent variances of the latent variables were all passed through the softplus. We set σ^2 in (1) to 0.5.

In our experiment, the batch size was set to 32. The dimensions D^{t}, D^{v}, and D^{p} were experimentally decided. We used Adam [30] optimizer with an initial learning rate of 0.0001, and it decayed exponentially by 0.01% per epoch. We applied cyclical annealing of KL regularization [31] from zero to one every ten epochs. The training was conducted for 200 epochs, and we used the model that achieved the best validation loss.

C. Evaluation Criteria

We evaluated the degree of disentanglement by calculating the instrument classification and pitch estimation accuracies in each latent space. If the latent spaces are ideally disentangled, one of the latent features (e.g., pitch features) should not include information on the other features (e.g., timbral/variation features). We created k-nearest neighbor (k-NN) classifiers using the two kinds of training sets along with their annotations of instrument names and semitone-level pitches. We adopt k-NN because DNN-based methods require additional model building and parameter tuning in a space- and task-dependent manner. In contrast, k-NN can directly reflect the structure of each latent space with sufficient accuracy. The accuracy should be high only in each corresponding space. We set k to 5 in all the experiments.

In addition, we also measured the quality of spectrogram reconstruction to monitor information loss via disentangled representations. For the quality of spectrogram reconstruction, we calculated the mean squared error (MSE) between the input log-amplitude spectrogram $\mathbf{X}$ and the output log-amplitude spectrogram $\mathbf{Y} \triangleq \mu_{\theta,ft}(\mathbf{Z})$ per time-frequency bin for the test data. Because we formulate the deep generative model as (1) with fixed $\sigma^2 = 0.5$, the MSE is equivalent to the negative log-likelihood for $\mathbf{X}$, and a lower MSE score indicates a better reconstruction quality.

On the basis of both criteria, we compared our proposed method and two conventional methods. Specifically, we compared the scores of the proposed three-factor model with those obtained for the local timbral and local pitch features and those obtained for the global timbral and local pitch features. We can see the proposed model as the first existing model that masks local variation features $\mathbf{z}^{\mathrm{v}}_{1:T}$ and as the second one that masks global timbral features $\mathbf{z}^{\mathrm{t}}$. We further conducted our experiments for both the timbre-variation conditional pitch model and the timbre-variation independent pitch model.

D. Experimental Results

Our method generally performed better in disentanglement. Table II shows comparative evaluations of the conventional

[1] https://github.com/spotify/pedalboard

TABLE II
COMPARISON OF TWO-FACTOR DISENTANGLEMENT (CONVENTIONAL) AND THREE-FACTOR DISENTANGLEMENT (OURS).

Factors			Input data			Instrument classification			Pitch estimation		
Timbre	Variation	Pitch	Pitched	Perc.	MSE ↓	**Timbre** ↑	Variation	Pitch ↓	Timbre ↓	Variation ↓	**Pitch** ↑
	✓	✓	✓		0.557	—	0.659 (↑)	0.445	—	0.066	0.813
✓		✓	✓		0.596	0.958	—	0.454	0.051	—	0.766
✓	✓	✓	✓		0.553	0.974	0.340 (↓)	0.395	0.107	0.024	0.751
	✓	✓	✓	✓	0.558	—	0.686 (↑)	0.516	—	0.109	0.816
✓		✓	✓	✓	0.599	0.938	—	0.422	0.092	—	0.729
✓	✓	✓	✓	✓	0.549	0.962	0.324 (↓)	0.416	0.145	0.054	0.777

↑ means higher is better, and ↓ means lower is better.

TABLE III
COMPARISON OF TIMBRE-VARIATION-CONDITIONED AND INDEPENDENT PITCH ENCODERS.

	Input data			Instrument classification			Pitch estimation		
Pitch encoder	Pitched	Perc.	MSE ↓	**Timbre** ↑	Variation ↓	Pitch ↓	Timbre ↓	Variation ↓	**Pitch** ↑
Conditional	✓		0.553	0.981	0.334	0.476	0.096	0.024	0.770
Independent	✓		0.553	0.974	0.340	0.395	0.107	0.024	0.751
Conditional	✓	✓	0.548	0.954	0.320	0.514	0.132	0.053	0.776
Independent	✓	✓	0.549	0.962	0.324	0.416	0.145	0.054	0.777

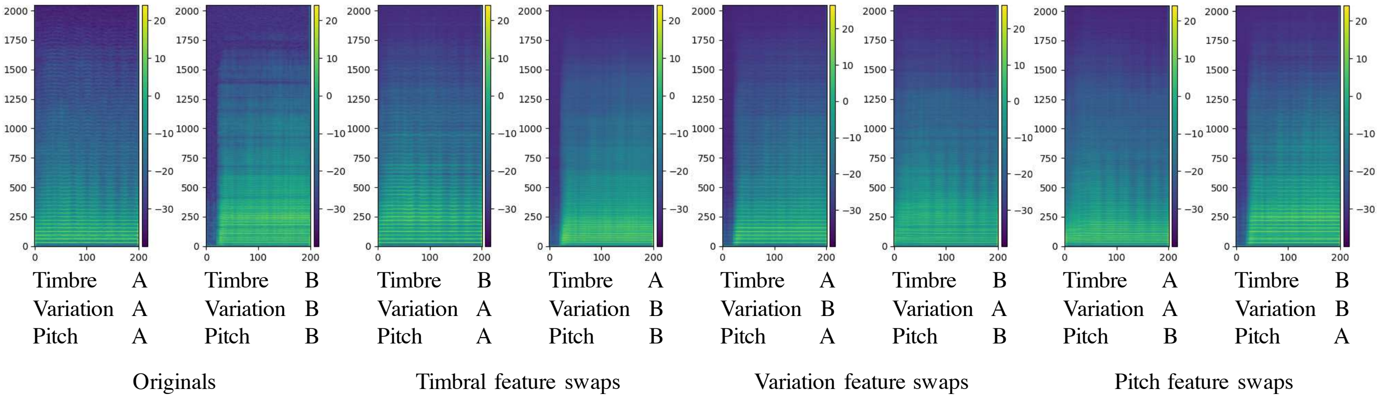

Fig. 3. Examples of feature-swapped spectrograms.

two-factor disentanglement and our three-factor disentanglement methods. We used the timbre-variation independent model for this comparison. The performance of the conventional disentanglement of local timbral (depicted as variation) and local pitch features was far inferior in instrument classification compared with the other disentanglement methods. In contrast, our method performed better in instrument classification and comparatively (for the pitched sounds) or better (for the unpitched percussive sounds) in pitch estimation compared with the two-factor method with global timbral and local pitch features. We can also see that our method achieved the best reconstruction qualities in terms of MSE score.

An interesting difference between the two-factor methods and the proposed method is that our method succeeded in utilizing percussive sounds much better than the others. Specifically, the pitch estimation score improved in the latent pitch space. This suggests that our model was capable of recognizing unpitched sounds in addition to pitched sounds. From the results so far, we can declare that our method disentangled the three latent features more attractively with less information loss compared with the existing methods with two latent features.

Table III shows comparative evaluations of the timbre-variation conditional pitch and independent pitch versions of the proposed method. Overall, their performances were quite similar except for the score of instrument classification in the latent pitch spaces. This result suggests that too much information other than pitch characteristics was included in the latent pitch space of the conditional version. Thus, we can say that the independent version is superior to the conditional version in disentanglement.

To investigate the effect of each feature in the proposed three-factor disentanglement, we swapped corresponding features between different musical instrument sounds (Fig. 3). The global timbres seemed to control the time-invariant distributions of the amplitude densities along the frequency axis abstractly, while the local pitch features controlled time-varying peak positions and sparseness along the same axis more concretely. In contrast, the local variation features seemed to control time-varying expressions, playstyles, and also volumes along the time-frame axis, as we expected.

V. Conclusion

This paper presented an unsupervised disentangled representation learning method for disentangling musical instrument sounds into global timbral, local pitch, and local variation features. Our model can handle various musical instrument sounds, including unpitched percussive sounds. We introduced random perturbation with pitch shift and timbre distortion to achieve disentanglement in an unsupervised manner, and we enhanced the disentanglement by cepstrum analysis. We experimentally confirmed that our method disentangled the three latent features more attractively with less information loss compared with the existing methods with two latent features. Our future work includes extending the proposed method to treat longer-duration sounds. Such sounds can include silent frames or pitch transitions and are thus much more challenging to tackle. We also plan to expand our method to a semi-supervised one to be able to utilize existing annotations during training.

Acknowledgment

This work is supported in part by JST PRESTO No. JPMJPR20CB and JSPS KAKENHI Nos. 19H04137, 20K21813, 21K02846, 21K12187, 22H03661, 22J22424.

References

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *Signal Processing Magazine (SPM)*, vol. 36, no. 1, pp. 20–30, 2019.

[2] Yu-Te Wu, Berlin Chen, and Li Su, "Polyphonic music transcription with semantic segmentation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 166–170.

[3] Carlos Hernandez-Olivan, Ignacio Zay Pinilla, Carlos Hernandez-Lopez, and Jose R. Beltran, "A comparison of deep learning methods for timbre analysis in polyphonic automatic music transcription," *Electronics*, vol. 10, no. 7, pp. 810, 2021.

[4] Keunwoo Choi, George Fazekas, and Mark Sandler, "Text-based lstm networks for automatic music composition," in *Proceedings of Conference on Computer Simulation of Musical Creativity (CSMC)*, 2016.

[5] Yoonchang Han, Jaehun Kim, and Kyogu Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 1, pp. 208–221, 2017.

[6] Siddharth Gururani, Mohit Sharma, and Alexander Lerch, "An attention mechanism for musical instrument recognition," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 83–90.

[7] Ethan Manilow, Prem Seetharaman, and Bryan Pardo, "Predominant musical instrument classification based on spectral features," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 771–775.

[8] Hugo Flores Garcia, Aldo Aguilar, Ethan Manilow, and Bryan Pardo, "Leveraging hierarchical structures for few-shot musical instrument recognition," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2021, pp. 220–228.

[9] Adrien Bitton, Philippe Esling, and Tatsuya Harada, "Vector-quantized timbre representation," in *Proceedings of International Computer Music Conference (ICMC)*, 2021.

[10] Shih-Lun Wu and Yi-Hsuan Yang, "MuseMorphose: Full-song and fine-grained music style transfer with one transformer VAE," in *arXiv:2105.04090*, 2021.

[11] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2014.

[12] Yin-Jyun Luo, Kat Agres, and Dorien Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 746–753.

[13] Yin-Jyun Luo, Kin Wai Cheuk, Tomoyasu Nakano, Masataka Goto, and Dorien Herremans, "Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 700–707.

[14] Yin-Jyun Luo, Sebastian Ewert, and Simon Dixon, "Towards robust unsupervised disentanglement of sequential data – a case study using music audio," in *Proceddings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 3299–3305.

[15] Keitaro Tanaka, Ryo Nishikimi, Yoshiaki Bando, Kazuyoshi Yoshii, and Shigeo Morishima, "Pitch-timbre disentanglement of musical instrument sounds based on VAE-based metric learning," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 111–115.

[16] Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman, "A universal music translation network," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

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

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

[19] Yun-Ning Hung, I-Tung Chiang, Yi-An Chen, and Yi-Hsuan Yang, "Musical composition style transfer via disentangled timbre representations," in *Proceddings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 4697–4703.

[20] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "DDSP: Differentiable digital signal processing," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[21] Yusong Wu, Ethan Manilow, Yi Deng, Rigel Swavely, Kyle Kastner, Tim Cooijmans, Aaron Courville, Cheng-Zhi Anna Huang, and Jesse Engel, "MIDI-DDSP: Detailed control of musical performance via hierarchical modeling," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2022.

[22] Masaya Kawamura, Tomohiko Nakamura, Daichi Kitamura, Hiroshi Saruwatari, Yu Takahashi, and Kazunobu Kondo, "Differentiable digital signal processing mixture model for synthesis parameter extraction from mixture of harmonic sounds," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 941–945.

[23] Kaizhi Qian, Yang Zhang, Shiyu Chang, David Cox, and Mark Hasegawa-Johnson, "Unsupervised speech decomposition via triple information bottleneck," in *Proceedings of International Conference on Machine Learning (ICML)*, 2020, pp. 7836–7846.

[24] Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Hwan Lee, Hoon Heo, and Kyogu Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," in *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[25] Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li, "Disentanglement of emotional style and speaker identity for expressive voice conversion," in *Proceedings of Annual Conference of the International Speech Communication Association (Interspeech)*, 2022 (to appear).

[26] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[27] B. P. Bogert, J. R. Healy, and J. W Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, 1963, pp. 209–243.

[28] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi

Oka, "RWC Music Database: Music genre database and musical instrument sound database," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2003, pp. 229–230.

[29] Brian McFee, Colin Raffel, Dawen Liang, Daniel P W Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of Python in Science Conference*, 2015, pp. 18–25.

[30] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[31] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating KL vanishing," in *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 240–250.