Exploring Multi-Descriptor Disentangled Representations of Acoustic Instrument Notes

Francesco Ardan Dal Rí, Gregorio Andrea Giudici, Luca Turchet, Nicola Conci
Department of Information Engineering and Computer Science,
University of Trento, Italy
[francesco.dalri-2 , gregorio.giudici , luca.turchet , nicola.conci]@unitn.it

***Abstract*—Comprehensive representation is key for improving controllability in generative neural networks. We present an approach for learning disentangled latent representations of individual instrumental notes, leveraging a Variational Autoencoder-based architecture designed to operate on spectrograms and explicitly capture four musical descriptors: timbre, pitch, dynamics, and duration. To achieve a structured and interpretable latent space, we exploit a combination of Gaussian Mixture priors, adversarial training, and auxiliary supervised clustering, promoting both compactness and semantic coherence in the learned representations yet preserving the ability to accurately reconstruct the original spectrograms. Experimental results and latent space explorations on the TinySol dataset show the effectiveness of the proposed approach, outperforming baseline models and existing methods in key metrics of reconstruction quality and classification accuracy.**

***Index Terms*—Representation Learning, Latent Space Disentanglement, Audio Generation**

I. Introduction

A musical note can be described via high-level attributes, the most common being:

- Timbre (T): the spectral characteristics of the instrument, or more generally, the instrumental class;
- Pitch (P): the fundamental frequency;
- Dynamics (V[1]): the intensity of the sound (e.g., *p* or *mf*);
- Duration (D): the temporal duration of the sound.

While in the synthetic realm these parameters can usually be modeled independently, in the acoustic domain they are more intertwined due to the physical and mechanical properties of musical instruments [1], [2]. For instance, P may influence D in plucked string instruments as higher-pitched notes decay faster; or V can affect T, as in bowed strings where increased bow pressure enhances higher overtones. Still, we are used to treating these parameters as separate entities: for instance, Western music notation evolved to represent these descriptors individually (Fig. 1). Among them, P, V, and D are relatively straightforward to represent and notate, with T remaining the most challenging to formalize, due to its multidimensional nature. Although high-level descriptive terminology for timbre exists [3], it is prone to subjective interpretation; thus, timbre is usually inferred indirectly via class labels [4]–[6].

Developing a structured and interpretable encoding of such descriptors is crucial in several applications [7], [8], including music-related tasks such as audio generation [4], [9], [10], domain adaptation [11], or computational musicology [12].

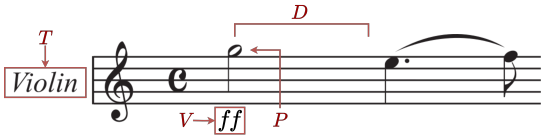

Fig. 1: Example of a Western classical musical score displaying four high-level descriptors: instrument / timbre (T), pitch (P), dynamics (V), and duration (D).

A *disentangled latent representation* refers to as an organized latent space, in which each dimension corresponds to a distinct and independent factor of variation in the data [13]. In such a representation, modifying a single variable influences only one specific attribute without affecting others.

Comprehensive latent representations are thereby fundamental for controllable audio generation, with Variational Autoencoders (VAEs) being particularly effective in modeling complex distributions in low-dimensional spaces. For instance, VAEs have been used for text-to-speech generation [14], timbre modeling [15], and large-scale music generation [16], demonstrating their ability to learn meaningful latent structures. However, such representation may lack direct interpretability, not allowing to find a straightforward and unique correspondence between the specific attributes and the generated output. This limitation motivates the need for latent disentanglement, fostering individual latent dimensions corresponding to distinct musical characteristics.

Furthermore, as VAEs in their standard formulation often struggle with multi-modal distributions due to the unimodal nature of the prior, Gaussian Mixture VAEs (GMVAEs) [17] have been introduced in this work; the idea behind the chosen architecture is to replace the simple prior with a mixture of Gaussian distributions, providing an inherent structure to their latent spaces. In the context of audio generation, this property has proven especially advantageous for timbre modeling [18].

Regarding the generation of instrumental samples, several approaches have been proposed, employing both supervised and unsupervised solutions. Unsupervised methods include multiple auxiliary losses [19] or Jacobian regularization [20]. Supervised methods, on the other hand, promote clusterization by leveraging classifier regularization [6], contrastive losses

[1]We denote dynamics as V - from *Velocity* - to avoid confusion with D (duration).

[5], or adversarial losses [21]. Despite these advancements, the majority of the literature (e.g., [5], [19]–[24]) deals with timbre-pitch disentanglement, implicitly assuming that other musical attributes are inherently encoded within the broader timbre space. Indeed, explicit control over dynamics and duration currently remains largely unexplored. Given the complex interdependencies among these attributes in acoustic instrumental notes, we deem that a neural-based architecture should explicitly incorporate all four descriptors.

To the best of our knowledge, only a few works extend beyond timbre and pitch, e.g. [25], [26], which introduce musical articulations[2] as an additional disentangled attribute, and [27], which shows how a pitch-, volume-, and duration-invariant representation improves audio quality assessment.

In this work, we present a method for encoding disentangled representations from audio spectrograms, alongside an exploration of the retrieved latent spaces, focusing on four common descriptors of timbre, pitch, dynamics, and duration, which musicians intuitively understand and can manipulate in performance and composition.

II. Proposed Method

An overview of the proposed method, including the model architecture and training method, is shown in Fig. 2. Supplementary material and code are available online[3].

A. Architecture Overview

We propose a *Multi-Descriptor Gaussian Mixture VAE* (MD-GMVAE) model, that consists of a decoder $\mathcal{D}$, and an encoder $\mathcal{E}$, comprising a main convolutional feature extractor ϕ, which maps an input spectrogram $\mathbf{X}$ to a shared feature vector $\mathbf{h} = \phi(\mathbf{X})$, and an ensemble of four variational encoders $\mathcal{V}_\xi$, one for each descriptor $\xi \in \{T, P, V, D\}$. The feature vector $\mathbf{h}$ is passed through the four encoders to generate the corresponding latent variables $\mathbf{z}_\xi = \mathcal{V}_\xi(\mathbf{h})$.

As in [18], priors follow a mixture of Gaussians $p(\mathbf{z}_\xi|y_\xi) \sim \mathcal{N}(\mu_{y_\xi}, \mathrm{diag}(\sigma_{y_\xi}))$, where y_ξ is K-way categorical variable - K being the number of classes in every ξ. The approximate posterior distribution $q(\mathbf{z}_\xi|\mathbf{X})$ is thus modeled as a Gaussian with learned mean and diagonal covariance, parameterized by the variational encoders $\mathcal{V}\xi$.

Finally, $\mathcal{D}$ reconstructs the spectrogram $\hat{\mathbf{X}}$ from the concatenated latent variables $\mathbf{z}_\mathcal{S} = \mathbf{z}_T \oplus \mathbf{z}_P \oplus \mathbf{z}_V \oplus \mathbf{z}_D$.

B. Objectives

The objective of the model $\mathcal{L}$ is to maximize the Evidence Lower Bound (ELBO), which balances the quality of the reconstruction while regularizing the latent spaces:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{z}_\mathcal{S}|\mathbf{X})}\left[\log p(\hat{\mathbf{X}}|\mathbf{z}_\mathcal{S})\right] - \sum_{\xi\in\{T,P,V,D\}} \beta D_{\mathrm{KL}}\left(q(\mathbf{z}_\xi|\mathbf{X})\|p(\mathbf{z}_\xi|y_\xi)\right) \quad (1)$$

[2]Articulation describes how notes are performed, defining unique effects according to instrumental peculiarities (*staccato*, *legato*, etc.)

[3]https://github.com/gregogiudici/multidescriptor-vae

The first term involves the reconstruction loss, where we compute the log-likelihood using a combination of a weighted Mean-Squared Error (MSE) loss and a Huber loss; the second one is the Kullback-Leibler (KL) divergence, computed for each descriptor, scaled by a factor β [28].

C. Supervised Clustering

To encourage structured latent representations, we introduce four classifiers $\mathcal{C}_\xi$, each of them corresponding to one of the four musical descriptors $\xi \in \{T, P, V, D\}$. These classifiers are implemented as shallow Multi-Layer Perceptrons (MLPs) with two layers. Their primary function is to predict the categorical label associated with each descriptor from the respective latent variables. In addition to the primary objective defined in Eq. (1), we introduce a supervised classification loss $\mathcal{L}_\mathcal{C}$ to enhance latent space clustering. This loss is formulated as a Cross-Entropy loss for each classifier, encouraging the latent space embeddings to align with their corresponding categorical labels:

$$\mathcal{L}_\mathcal{C} = \sum_{\xi\in\{P,V,D,T\}} \mathrm{CE}(\mathbf{y}_\xi, \mathcal{C}_\xi(\mathbf{z}_\xi)) \quad (2)$$

By incorporating this classification loss, we enforce the model to encode information in specific dimensions and learn structured latent spaces, where each latent variables z_ξ effectively captures characteristics related to the corresponding musical descriptor.

D. Adversarial Disentanglement

Inspired by the study reported in [22], we also implement a 2-stage adversarial training, to promote each of the four latent representation to discard information related to all other descriptors. Indeed, let $\mathcal{R}_\xi$ be a *Remover* - shallow 2-layers MLPs - for each of the four descriptors. During training, we alternate two stages: in the first one, we freeze the Removers and add the following term to the loss functions:

$$\mathcal{L}_{\mathcal{R}_1} = \sum_{\xi\in\{T,P,V,D\}} \lambda D_{\mathrm{KL}}\left(\frac{1_\xi}{|\Xi|}\|\mathcal{R}_\xi\left(\bigoplus_{\substack{\rho\in\{T,P,V,D\}\\ \rho\neq\xi}} \mathbf{z}_\rho\right)\right) \quad (3)$$

In (3), 1_ξ is the all-one vector, $|\Xi|$ is the number of classes in each descriptor, and λ a scaling factor. This hampers the model to predict a descriptor given the other $\mathbf{z}_{\rho\neq\xi}$ and promote the removal of residual information.

In the second stage, we instead freeze the model and simply optimize the Removers through the following summary of Cross-Entropy losses to avoid the collapse of the Removers:

$$\mathcal{L}_{\mathcal{R}_2} = \sum_{\xi\in\{T,P,V,D\}} \mathrm{CE}\left(\mathbf{y}_\xi, \mathcal{R}_\xi\left(\bigoplus_{\substack{\rho\in\{T,P,V,D\}\\ \rho\neq\xi}} \mathbf{z}_\rho\right)\right) \quad (4)$$

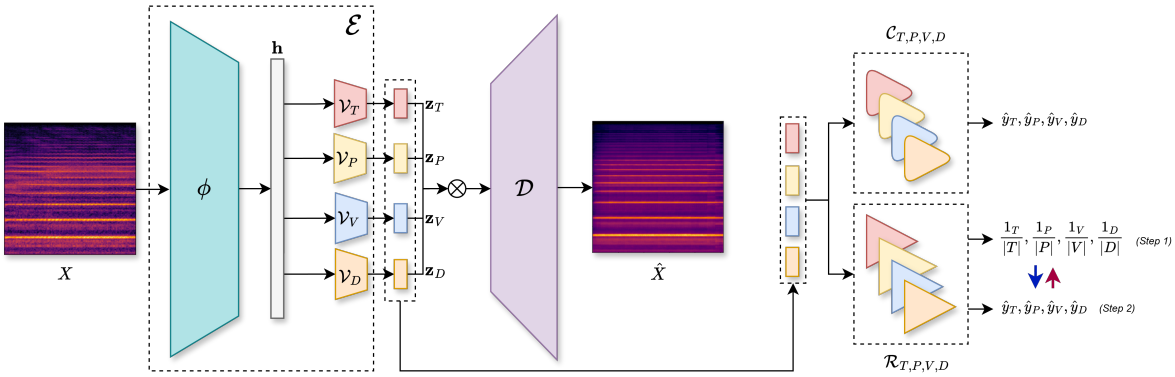

Fig. 2: The proposed MD-GMVAE includes an encoder $\mathcal{E}$, consisting of a feature extractor ϕ and an ensemble of 4 variational encoders $\mathcal{V}_\xi$, one for each descriptor $\xi \in \{T, P, V, D\}$, and a shared decoder $\mathcal{D}$. To improve the performance of the model, we also employed four Classifiers C_ξ for supervised clustering, and four Removers R_ξ for adversarial disentanglement.

III. Experimental Setup

A. Dataset

We conducted our esperiments on TinySOL [29], a publicly available dataset widely applied in the literature - e.g., [18], [30], [31]. It contains about 2900 monophonic, 44.1kHz/16-bit audio files from 14 classical instruments belonging to different orchestral families: strings, winds and brass. Files last about 6s and cover the entire range of each instrument, with annotations for pitch (in semitones) as well as three dynamics (*pp*, *mf*, *ff*). Actual duration of each audio sample, though not directly notated, can be easily retrieved (see Section III-B) and stored as an additional label for each sample.

B. Preprocessing

The audio samples in the dataset exhibit different lengths. Thus, at first we trim every audio file, so as to remove silence with a threshold of -40dB and to retrieve the actual duration of the samples. We quantize the durations every 250ms to achieve discrete classes. Then, we resample at 22050Hz, pad each sample to a fixed length of 5.94s, so as to extract Mel-Spectrograms with $n\_bins = 256$, $n\_fft = 1024$, $hop\_size = 512$; in this way we obtain spectrograms of size (256, 256). Spectrograms are normalized 0-1, to guarantee that the network focuses on spectral relationships rather than absolute values. Finally, we split the dataset into 80% train, 10% validation, and 10% test.

C. Hyperparameters and Training

As we aim at minimizing the latent space dimensionality, we initialize our model with latent dimensions equal to $L_T = 8$ for $\mathcal{V}_T$, and $L_{P,V,D} = 4$ for the other encoders. Linear layers in $\mathcal{E}$, $\mathcal{D}$ are initialized with Xavier initialization, while we used the Kaiming one in $\mathcal{C}_\xi$ and $\mathcal{R}_\xi$ [32]. The model is trained on a single Nvidia RTX 4090 GPU for a maximum of 1000 epochs, with Early Stopping in validation to prevent overfitting. We use Adam optimizer with batch size $BS = 64$ and an initial learning rate $LR = 1 \times 10^{-3}$, adjusted via a plateau LR scheduler. Similarly, the scaling factors β and λ in Eq. (1) and (3) are dynamically incremented using a fixed scheduler over epochs.

As baselines, we also readapt the architecture as a unimodal VAE and a MD-VAE, keeping the same amount of layers. In the former, the classifiers receive the whole $\mathbf{z}_{VAE}$ of latent dimension $L = 8 + 4 + 4 + 4$; in the latter, each classifier receives separate unimodal $\mathbf{z}_T$, $\mathbf{z}_P$, $\mathbf{z}_V$, and $\mathbf{z}_D$.

IV. Results and Discussions

Our method overall returns optimal spectrogram reconstruction, while also proving robust in classification tasks on unseen data for all the four descriptors. Among them, V exhibits the worst classification results. Despite amplitude-related features are generally overlooked in the literature as they are easy to compute [5], we argue that in the context of acoustic instrument their close relationship with timbre supports the need of a multidescriptor approach. In addition, with respect to the unimodal baselines, the use of multiple distributions promoted clear clustering in the latent spaces, with the inclusion of the four Removers slightly improve overall results in classification tasks (see Table I).

	$T_{acc} \uparrow$	$P_{acc} \uparrow$	$V_{acc} \uparrow$	$D_{acc} \uparrow$	$R_{loss} \downarrow$
VAE + $\mathcal{C}_\xi$	0.938	0.931	0.876	0.891	0.0156
MD-VAE + $\mathcal{C}_\xi$	0.945	0.865	0.803	0.909	0.0174
MD-GMVAE + $\mathcal{C}_\xi$	0.996	0.993	0.883	**0.989**	0.0113
MD-GMVAE + $\mathcal{C}_\xi$, $\mathcal{R}_\xi$	**1.000**	**0.996**	**0.898**	**0.989**	**0.0111**

TABLE I: Performance of VAE with Multi-Descriptor (MD), Gaussian Mixture (GM), Classifiers ($\mathcal{C}_\xi$), and Removers ($\mathcal{R}_\xi$).

	Dataset	L_T / L_P	$T_{acc} \uparrow$	$P_{acc} \uparrow$
Luo et al. [18]	TinySol	16 / 16	**1.000**	**0.996**
Luo et al. [19]	TinySol	8 / -	0.892	–
Tanaka et al. [25]	RWC [33]	64 / 32	0.981	0.816
Our	TinySol	8 / 4	**1.000**	**0.996**

TABLE II: T and P accuracy comparison with similar methods in literature.

Despite the reduced latent dimensionality and the increased latent complexity due to the increased attributes to be disentangled, accuracies on T and P are in line with similar methods in the literature (see Table II).

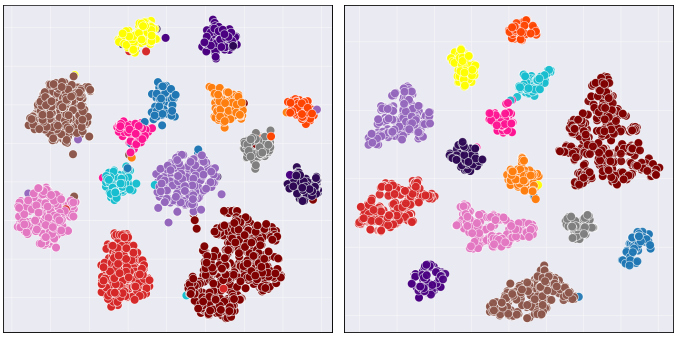

Fig. 3: t-SNE on MD-VAE's z_T (left) and on full MD-GMVAE's z_T (right), both color-coded by T labels.

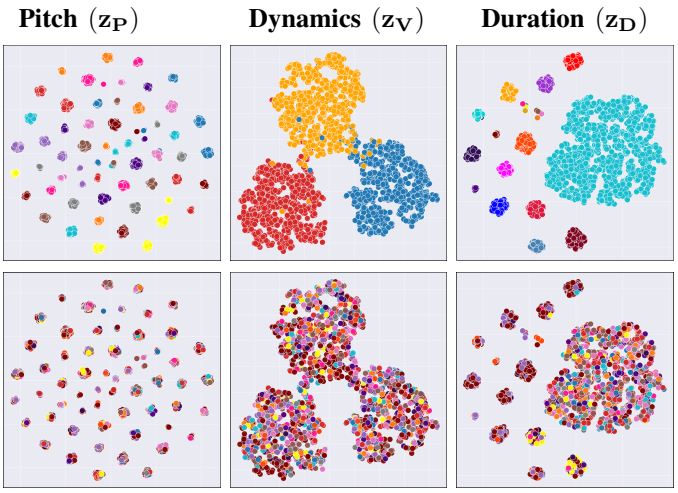

Fig. 4: t-SNE on full MD-MGVAE's z_P, z_V, and z_D: (top) color-coded by respective labels, with P being reduced to octaves; (bottom) color-coded by T labels.

A. Latent Space Exploration

We perform a set of experiments to verify the effectiveness of our approach.

At first, we aim at exploring the overall structure of the learned latent representation. We apply dimensionality reduction using t-SNE on the latent spaces retrieved by the MD-VAE and full MD-GMVAE encoders. The introduction of classifiers on the latents naturally forces the models to separate information; still, for each descriptor, we observe better-defined clusters in the latter, indicating that the MD-GMVAE has learned more meaningful structures in the data with respect to the MD-VAE, where clusters are overall less defined. (see Fig. 3).

Next, we verify the disentanglement of the four descriptors. By color-coding each dimensionality reduction plot by labels belonging to a different descriptor, e.g., P color-coded by T, then the plots appear unorganized without any notable label cluster, indicating that each latent representation is effectively disentangled and no relevant information of a given descriptor leaks into others latents (see Fig. 4).

Then, we perform stepped sample-to-sample interpolation: given two latent representations $\mathbf{z}_{\mathcal{S}}^1$ and $\mathbf{z}_{\mathcal{S}}^2$, extracted from two samples, we first compute full interpolation over latent representations as $\mathbf{z}(\alpha) = (1-\alpha)\mathbf{z}_{\mathcal{S}}^1 + \alpha\mathbf{z}_{\mathcal{S}}^2, \quad \alpha \in \{-1, -0.5, 0, 0.5, 1\}$. Secondly, for every descriptor $\xi \in \{T, P, V, D\}$, we interpolate only along $\mathbf{z}_\xi$, keeping other dimensions fixed as $\mathbf{z}_\xi(\alpha) = (1-\alpha)\mathbf{z}_\xi^1 + \alpha\mathbf{z}_\xi^2, \quad \mathbf{z}_{\mathcal{S}\setminus\neg\xi}(\alpha) = \mathbf{z}_{\mathcal{S}\setminus\neg\xi}^1$. Each interpolated latent representation $\mathbf{z}(\alpha)$ is then decoded back into the data space, producing a sequence that smoothly transitions between the original samples (see Fig. 5). In this experiment, we occasionally observed some artifacts in the reconstructed spectrograms: we argue that this is motivated by the fact that operating over discrete classes does not always allows for meaningful in-between scenarios.

Finally, we operate 5-step sweeps across every individual dimension of a $\mathbf{z}_{\mathcal{S}}$ extracted from a single sample, exploring how the model learned to organize information. Specifically, for each dimension $z_{\mathcal{S},i}$, we substitute its value with a given α, while keeping all other dimensions fixed, e.g., $\mathbf{z}_{\mathcal{S},1}(\alpha) = [z_0, \alpha, z_2, \dots z_n], \quad \alpha \in \{-1, -0.5, 0, 0.5, 1\}$. While some dimensions influence subtle details (e.g., small variations in frequency magnitudes or in the noise floor), others appear strongly related to specific attributes, such as envelope transients, high spectral components, or single-component decay (see Fig. 6).

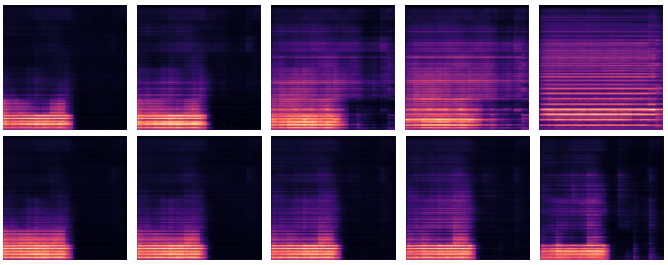

Fig. 5: Examples of interpolation for $\alpha \in \{-1, -0.5, 0, 0.5, 1\}$ (ordered left to right) over two samples, across $z_{\mathcal{S}}$ (top) and only z_T (bottom).

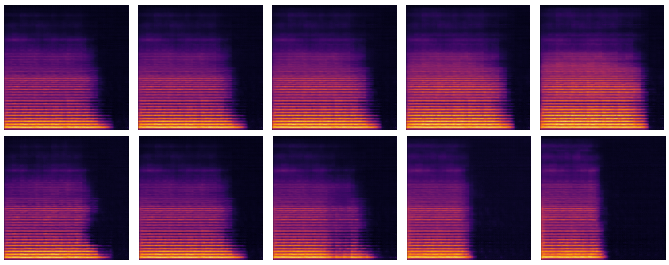

Fig. 6: Examples of single-dimension sweeps for $\alpha \in \{-1, -0.5, 0, 0.5, 1\}$ (ordered left to right) for $z_{T,1}$ (top) and $z_{D,1}$ (bottom).

V. Conclusions

In this paper we presented a novel approach to achieve a compact and disentangled latent representation of instrumental audio samples using four descriptors, commonly used to describe individual notes. Through a series of experiments, we explored the latent spaces produced by our model and verified the effective disentanglement between different attributes, as well as the reconstruction capabilities given a reduced - and therefore, easily interpretable - latent representation. We support that precise disentanglement over multiple descriptors could both provide a beneficial framework for creative applications and foster representation learning tasks.

Acknowledgment

G. A. Giudici's contributions are funded by the European Union (EU) under NextGenerationEU (M.D. 118/2023). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the EU or The European Research Executive Agency. Neither the EU nor the granting authority can be held responsible for them.

References

[1] M. Fabiani and A. Friberg, "Influence of pitch, loudness, and timbre on the perception of instrument dynamics," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 193–199, 2011.

[2] J. M. Hajda, "The effect of dynamic acoustical features on musical timbre," in *Analysis, synthesis, and perception of musical sounds: The sound of music*. Springer, 2007, pp. 250–271.

[3] J. Nistal, S. Lattner, and G. Richard, "Drumgan: Synthesis of drum sounds with timbral feature conditioning using generative adversarial networks," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2020, pp. 590–597.

[4] T.-W. Kim, M.-S. Kang, and G.-H. Lee, "Adversarial multi-task learning for disentangling timbre and pitch in singing voice synthesis," in *Interspeech*, 2022.

[5] K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii, and S. Morishima, "Pitch-timbre disentanglement of musical instrument sounds based on vae-based metric learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 111–115.

[6] A. V. Puche and S. Lee, "Caesynth: Real-time timbre interpolation and pitch control with conditional autoencoders," in *International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021, pp. 1–6.

[7] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," *Advances in neural information processing systems*, vol. 30, 2017.

[8] E. L. Denton *et al.*, "Unsupervised learning of disentangled representations from video," *Advances in neural information processing systems*, vol. 30, 2017.

[9] C. Gupta, P. Kamath, and L. Wyse, "Signal representations for synthesizing audio textures with generative adversarial networks," in *Sound and Music Computing Conference*, 2021.

[10] Y.-J. Luo, K. W. Cheuk, W. Choi, W.-H. Liao, K. Toyama, T. Uesaka, K. Saito, C.-H. Lai, Y. Takida, S. Dixon, and Y. Mitsufuji, "Disentangling multi-instrument music audio for source-level pitch and timbre manipulation," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[11] C.-H. Lin, C. Jones, B. W. Schuller, H. Coppock, and A. Akman, "Synthia's melody: A benchmark framework for unsupervised domain adaptation in audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 7450–7454.

[12] C. Weiß and M. Müller, "From music scores to audio recordings: Deep pitch-class representations for measuring tonal structures," *ACM Journal on Computing and Cultural Heritage*, vol. 17, no. 3, pp. 1–19, 2024.

[13] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," in *International Conference of Learning Representation (ICLR)*, 2017.

[14] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 5530–5540.

[15] P. Esling, A. Chemla-Romeu-Santos, and A. Bitton, "Generative timbre spaces with variational audio synthesis," in *International Conference on Digital Audio Effects (DAFx)*, 2018, pp. 175–181.

[16] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv:2005.00341*, 2020.

[17] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," *arXiv:1611.02648*, 2016.

[18] 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 *International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 746–753.

[19] Y.-J. 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.

[20] Y.-J. Luo, S. Ewert, and S. Dixon, "Unsupervised pitch-timbre disentanglement of musical instruments using a jacobian disentangled sequential autoencoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1036–1040.

[21] Z. Zhang and T. Akama, "Hyperganstrument: Instrument sound synthesis and editing with pitch-invariant hypernetworks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6640–6644.

[22] G. Narita, J. Shimizu, and T. Akama, "Ganstrument: Adversarial instrument sound synthesis with pitch-invariant instance conditioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[23] Y. Hashizume, L. Li, A. Miyashita, and T. Toda, "Learning multidimensional disentangled representations of instrumental sounds for musical similarity assessment," *arXiv:2404.06682*, 2024.

[24] S. Dutta and S. Ganapathy, "Zero shot audio to audio emotion transfer with speaker disentanglement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 371–10 375.

[25] K. Tanaka, Y. Bando, K. Yoshii, and S. Morishima, "Unsupervised disentanglement of timbral, pitch, and variation features from musical instrument sounds with random perturbation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 709–716.

[26] K. Tanaka, K. Yoshii, S. Dixon, S. Morishima *et al.*, "Unsupervised pitch-timbre-variation disentanglement of monophonic music signals based on random perturbation and re-entry training," *APSIPA Transactions on Signal and Information Processing*, vol. 14, no. 1, 2025.

[27] Y. Wang, X. Guan, Y. Du, C. Wang, X. Li, and Y. Pan, "Pivod: Pitch, volume and duration invariant representation for music tone quality evaluation," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[28] H. Sikka, W. Zhong, J. Yin, and C. Pehlevant, "A closer look at disentangling in β-vae," in *Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 888–895.

[29] C. E. Cella, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "Orchideasol: a dataset of extended instrumental techniques for computer-aided orchestration," in *International Computer Music Conference (ICMC)*, 2020.

[30] Y. Gonzalez and R. Prati, "Similarity of musical timbres using fft-acoustic descriptor analysis and machine learning," *Eng*, vol. 4, no. 1, pp. 555–568, 2023.

[31] M. Pasini, S. Lattner, and G. Fazekas, "Music2latent: Consistency autoencoders for latent audio compression," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[32] S. K. Kumar, "On weight initialization in deep neural networks," *arXiv:1704.08863*, 2017.

[33] 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.