IMPROVING INTERPRETABILITY IN GENERATIVE MULTITIMBRAL DDSP FRAMEWORKS VIA SEMANTICALLY-DISENTANGLED MUSICAL ATTRIBUTES

Francesco Ardan Dal Rì & Nicola Conci

DISI - Dept. of Information Engineering and Computer Science, University of Trento, Italy

ABSTRACT

We propose a generative multitimbral DDSP framework focusing on acoustic instruments modeling. Traditional DDSP architectures rely on frame-wise latent encoding, limiting semantic interpretability; the acoustic realm is further challenged by the inherently interdependency of musical descriptors. Our approach introduces disentangled global embeddings for timbre and dynamics, achieved via a Triple-VAE architecture, and a DDSP model to learn temporal dependencies from such reduced representation along with normalized envelope curves. Experimental results demonstrate optimal reconstruction accuracy while enabling flexible and realistic generation capabilities, with the disentangled representation promoting interpretable control over musical descriptors.

Index Terms— Differentiable Digital Signal Processing, Representation Learning, Audio Synthesis

1. INTRODUCTION

With the growing popularity of large models for audio, there is a growing interest towards smaller, specialized architectures, designed for fast and controllable real-time raw audio generation, relevant examples including autoencoders, latent diffusion, or adversarial frameworks [1, 2, 3].

A peculiar paradigm involves *parameter estimation*, where models are trained to estimate control parameters for a synthesizers to emulate input sounds [4, 5]. Notably, Differentiable Digital Signal Processing (DDSP) [6] expands on this approach by turning the synthesizer into a differentiable module, enabling end-to-end training on audio waveforms. DDSP frameworks have become popular in a variety of audio tasks [7, 8], and their compact size made them suitable for many real-time applications [9, 10]. Traditional DDSP architectures synthesize audio given pitch and loudness contours, along with a frame-wise latent representation $\mathbf{z}$ encoded from the input, capturing spectrotemporal nuances [6, 9]. However, this mostly requires the model to be trained on single instruments [11]. Moreover, while proved effective when synthesis is guided by external signals, as in timbre transfer or speech vocoding [7], frame-wise representations constitute a significant limitation in pure generative settings, where they may be difficult to manipulate as they lack global semantic meaning. Modeling acoustic instruments poses additional challenges, as key high-level musical descriptors such as timbre - i.e. instrumental class[1], and dynamics - i.e. how forcefully the instrument is played, are inherently interdependent and mutually influence each other [12, 13]. Therefore, a disentangled representation is critical for generation controllability, and especially multitimbral scenarios [14]. In this regard, frame-wise latent encoding is further problematic: for instance, distinguishing a *pp* from the decay of a *ff* requires broader temporal context, thereby forcing reliance on an encoder and, again, hindering interpretable control over latents $\mathbf{z}$.

Building on these motivations, we aim to improve interpretability in pure-generative multitimbral setting, and propose a decoder-only DDSP framework that exploits global embeddings for *timbre* and *dynamics*, trained on acoustic instruments sounds. We extract embeddings via a VAE-based architecture, enforcing latent disentanglement, reducing the representation to 8 dimensions for each descriptor - 16 in total. In addition to traditional pitch contours, we introduce normalized *envelope* curves, commonly used in sound synthesis, rather than raw loudness. This requires the model to infer amplitudes by internally combining information.

Using these well-established musical descriptors ensures interpretable representations, enabling flexible and controllable generation of novel audio samples.

2. ARCHITECTURES AND OBJECTIVES

The framework is composed by a VAE encoder and a DDSP decoder feeding an Harmonic-plus-Noise (HpN) audio module. The VAE encodes global embeddings for timbre (t) and dynamics (d) from spectrograms; despite not used by the decoder (see 2.2) in order to avoid representation complexity and training inconsistencies [15], we also disentangle pitch (p) preventing information leakage - benefits are proved in the literature [3, 12, 13]. Then, an improved DDSP decoder [6] elaborates such embeddings along with normalized envelope and pitch contours to synthesize audio waveforms. A summary of the framework is shown in Fig. 1; source code can be found in the online repository: https://github.com/return-nihil/MT-GEN_DDSP/.

[1]The term *timbre* refers here to the tonal characteristics of instrument classes, rather than to broader perceptual aspects, in line with e.g., [12, 13].

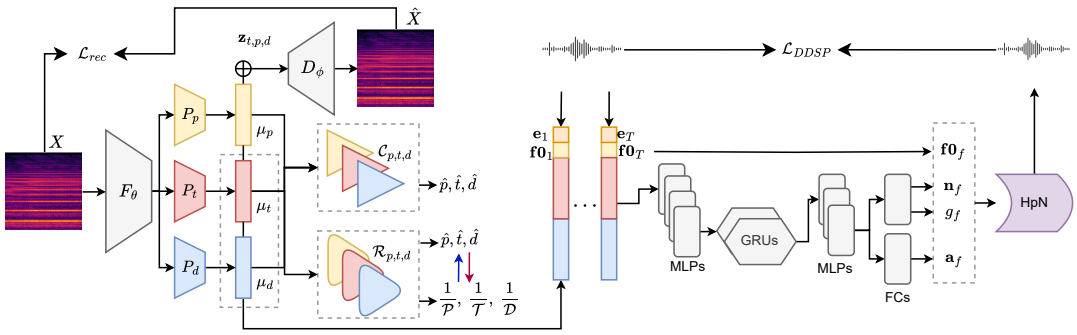
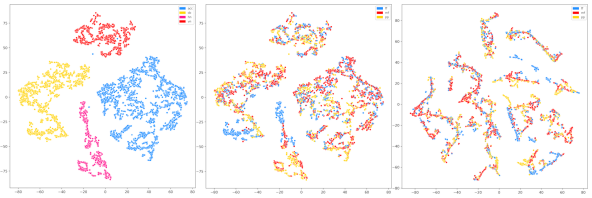


Fig. 1: Architecture and training scheme of the entire Triple-VAE + DDSP pipeline.



(a) (b) (c)

Fig. 2: Example of t-SNE of μ in t space: (a) w/ removers, color-coded by t; (b) w/ removers, color-coded by d; (c) w/o removers, color-coded by d.

2.1. Triple-VAE

A convolutional feature extractor F_θ maps the input spectrogram X to a shared feature vector $\mathbf{h}$, from which three parallel latent projectors $\mathcal{P}_{t,p,d}$ produce $(\boldsymbol{\mu}_i, \log \boldsymbol{\sigma}_i^2)$ for $i \in \{t, p, d\}$, sampling latent vectors $\mathbf{z}_i$ via reparameterization. The concatenated latent $\mathbf{z} = [\mathbf{z}_t, \mathbf{z}_p, \mathbf{z}_d]$ is passed to a decoder D_ϕ to reconstruct $\hat{X}$. The objective combines KL divergence with a weighted sum of ℓ_2 and ℓ_1 reconstruction losses.

To enforce attribute disentanglement, auxiliary classifiers $\mathcal{C}_{t,p,d}$ are trained on the respective latents; in addition, building on previous works [3, 12], "removers" $\mathcal{R}_{t,p,d}$ receive complementary latents (e.g., R_t takes $[\mathbf{z}_p, \mathbf{z}_d]$) to predict correspondent labels. Both networks classes are shallow 2-layer MLPs with ReLU activations. Following the method in [3, 12], supervised training alternates between two phases: (1) updating the removers to maximize their classification accuracy via cross-entropy loss, freezing the VAE and classifiers; and (2) updating the VAE and classifiers while freezing the removers, minimizing the following objective:

$$\mathcal{L}_{VAE} = (\text{MSE} + \text{L1}) + \mathcal{L}_{\mathcal{C}} + \beta(e)\mathcal{L}_{\text{kl}} + \lambda_{\mathcal{R}}\mathcal{L}_{\mathcal{R}} \quad (1)$$

where $\beta(e)$ is linearly annealed over epochs, $\lambda_{\mathcal{R}}$ is a scaling constants, and $\mathcal{L}_{\mathcal{R}}$ is a Jensen–Shannon divergence term pushing every remover output toward uniform categorical distribution $[1/K_i, \ldots, 1/K_i]$, with K_i being the number of respective classes, thus discouraging information leakage between different attributes' latent representation - Fig. 2. Phase 1 simply avoid removers' collapse; loss is expected to increase proportionally to the latent spaces' disentanglement.

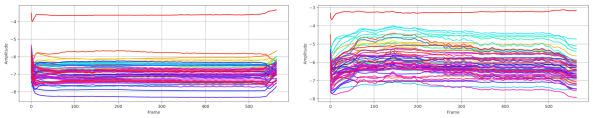


Fig. 3: Example of resulting predicted noise bands, by single-layer (left) and two-layer (right) GRU models.

2.2. DDSP Decoder

The decoder overall follows a standard DDSP implementation [6], with key modifications tailored to our setting. Instead of processing a flattened input vector, as in [9, 16], we construct the decoder input for the entire sequence by concatenating, at each frame $f \in F$, the normalized envelope $\tilde{e}_f$ and fundamental frequency $\tilde{f0}_f$ contours with the learned latent vectors $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_d$ replicated across all frames, thus obtaining the matrix $\mathbf{X} \in \mathbb{R}^{F \times D}$ where $D = \dim(\boldsymbol{\mu}_t) + \dim(\boldsymbol{\mu}_d) + 2 = 18$:

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{\mu}_t & \boldsymbol{\mu}_d & \tilde{e}_1 & \tilde{f0}_1 \\ \boldsymbol{\mu}_t & \boldsymbol{\mu}_d & \tilde{e}_2 & \tilde{f0}_2 \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\mu}_t & \boldsymbol{\mu}_d & \tilde{e}_F & \tilde{f0}_F \end{bmatrix} \quad (2)$$

Four consecutive MLP blocks, each comprising linear layers followed by layer normalization and LeakyReLU activations, expand D to 256 producing transformed features $\mathbf{x}'_f$. Contrarily to most DDSPs using a single recurrent layer [6, 9], here we use a two-layer GRU stack - Fig. 3 - to better learn detailed temporal dependencies [17] from a compact space:

$$\mathbf{h}_f^{(1)} = \text{GRU}_1(\mathbf{x}'_f, \mathbf{h}_{f-1}^{(1)}), \quad \mathbf{h}_f^{(2)} = \text{GRU}_2(\mathbf{h}_f^{(1)}, \mathbf{h}_{f-1}^{(2)}) \quad (3)$$

The hidden state $\mathbf{h}_f$ is computed as $\mathbf{h}_f = \text{LN}(\mathbf{h}^{(2)} f + \alpha, \mathbf{h}^{(1)} f)$ via a residual connection with scaling $\alpha = 0.5$ and LayerNorm, and processed through three additional MLP blocks and a final linear projection layer, which outputs a vector of size $n_{\text{harm}} + n_{\text{noise}} + 1$. The decoder is thus trained to predict harmonic amplitudes $\mathbf{a}_f \in \mathbb{R}^{n_{\text{harm}}}$, noise band magnitudes $\mathbf{n}_f \in \mathbb{R}^{n_{\text{noise}}}$, and a global amplitude control $g_f \in \mathbb{R}$. Softplus or exponential–sigmoid activations ensure parameter positivity, components exceeding Nyquist are masked out, and $\mathbf{a}_f$

are scaled by g_f. All parameters are linearly upsampled to sampling rate before being passed to the HpN synthesizer.

During training, we minimize a weighted sum of losses:

$$\mathcal{L}_{DDSP} = \mathcal{L}_{\mathrm{MR-STFT}} + \beta(e)\mathcal{L}_{\mathrm{MFCC}} + \mathcal{L}_{\mathrm{E}} \quad (4)$$

where $\mathcal{L}_{\mathrm{MR-STFT}}$ is a multi-resolution STFT loss following the original implementation [6], $\beta(e)$ is linearly annealed over epochs, $\mathcal{L}_{\mathrm{MFCC}}$ computes cosine similarity between MFCC features, and $\mathcal{L}_{\mathrm{E}}$ is a scale-invariant energy loss penalizing RMS differences.

3. EXPERIMENTAL SETUP

3.1. Dataset

We perform all experiments on TinySOL [18], a publicly available dataset of monophonic instrumental samples, commonly used in the literature - e.g., [12, 13]. It contains ca. 2900 files of classical instruments with annotation for instrument (14 labels), pitch (entire range), and dynamics (*pp*, *mf*, *ff*), sampled at 44.1kHz. We acknowledge the limited size of the dataset; however, literature shows that DDSPs can be trained on small dataset [7, 19, 20]. Still, to partially address such problem, we applied light augmentations - time-shift, time-stretch, pitch-shift, low-pass filtering, noise injection, and tanh saturation. From each sample, we computed spectrograms for training the Triple-VAE, and normalized per-frame envelope (RMS) and pitch (YIN) contours for the DDSP.

3.2. Hyperparameters and Training

We train the Triple-VAE and the DDSP separately on the same configuration, 70/30 train/test splits. We thus train three full pipelines, considering d (*pp*, *mf*, *ff*), a 2-octaves p range (48-71), and increasing sets of t:

- $t = 2$: accordion (*acc*), violin (*vn*);
- $t = 4$: *acc*, *vn*, double-bass (*cb*), horn (*hn*);
- $t = 8$: *acc*, *vn*, *cb*, *hn*, flute (*fl*), oboe (*ob*), trombone (*tbn*), viola (*va*);
- $t = 14$: full dataset.

Training is performed on a single NVIDIA GeForce RTX 4090 GPU, for 250 epochs each, using Adam optimizer with initial learning rate $\eta = 1e^{-3}$ (linearly decayed to $1e^{-6}$) and a batch size of 48. For DDSP, we randomly extract 2 seconds chunks at every iteration. Consistently with the literature [6, 7, 9], envelopes and pitch contours are extracted with a resolution of 10 ms, and we use $n_{\text{harm}} = 100$ and $n_{\text{noise}} = 65$.

4. RESULTS

Table 1 summarizes the performance of both models.

Regarding the Triple-VAE, accuracy results are in line with the state-of-the-art [12, 13]. We provide $\mathcal{R}$ accuracies as the metric for disentanglement: p accuracies tend to be low and stable, signing few leakage of information between semantic spaces. Also, t's improve proportionally to the number of t, and d shows an opposite trend. The former is an expected behavior, as the model is forced to focus on primary attributes for each descriptors, while the latter confirms d's strong dependence from t [12].

DDSP was evaluated by reconstructing the whole samples in the test split. MR-STFT is mostly similar in every configuration, while MSE increases proportionally to t, signing that the model performs well on a perceptual level but struggles at waveform-level when data diversity becomes large. In addition, we compute Fréchet Audio Distance (FAD) [21] using both VGG (spectral-related) and PANN (temporal-related) models. Similar to MSE, VGG FAD shows a slight increase proportional to t, though less pronounced; this aligns with DDSPs notoriously struggling in multitimbral scenarios [11, 22]. In contrast, PANN FAD decreases noticeably as t increases, signing that models' ability to reconstruct envelopes benefits from increasing the variety in the training data. Overall, the results indicate that our method can generate high-quality audio from a very condensed space - 16 dimensions compared to the traditional DDSP's 64 [6]. While this does not dramatically improve efficiency, we argue that a reduced dimensionality leads to a more interpretable representation.

To the best of our knowledge, the literature does not report multi-timbral instrumental DDSPs frameworks for straightforward comparison. Thus, to validate out method, we first provide a baseline: a VAE with no disentanglement (removing terms 2 and 4 in Eq. 1) with $ls = 16$ and $t = 2$, along

Table 1: Performance summary on unseen data (in every configuration, the whole pipeline shares the same test split).

Config	**Triple-VAE**							**DDSP**			
t	MSE ($\times 10^{-3}$) ↓	$\mathcal{C}$ Acc. ↑			$\mathcal{R}$ Acc. ↓			MR-STFT ↓	MSE ($\times 10^{-3}$) ↓	FAD ↓	
		t	*p*	*d*	*t*	*p*	*d*			VGG	PANN ($\times 10^{-4}$)
2	5.664 ± 3.970	1.00	1.00	1.00	0.74	0.21	0.44	1.208 ± 0.056	4.168 ± 0.974	2.256	0.500
4	5.488 ± 3.698	1.00	1.00	0.99	0.50	0.19	0.52	1.209 ± 0.077	6.263 ± 1.873	2.448	0.132
8	5.556 ± 3.797	1.00	1.00	0.99	0.37	0.13	0.54	1.153 ± 0.075	10.310 ± 4.550	2.618	0.019
14	5.733 ± 4.808	0.99	1.00	0.99	0.29	0.14	0.54	1.133 ± 0.078	13.622 ± 6.220	2.743	0.019
Baseline	5.574 ± 4.879	1.00	0.99	0.99	//	//	//	1.292 ± 0.072	4.728 ± 1.662	4.556	1.688

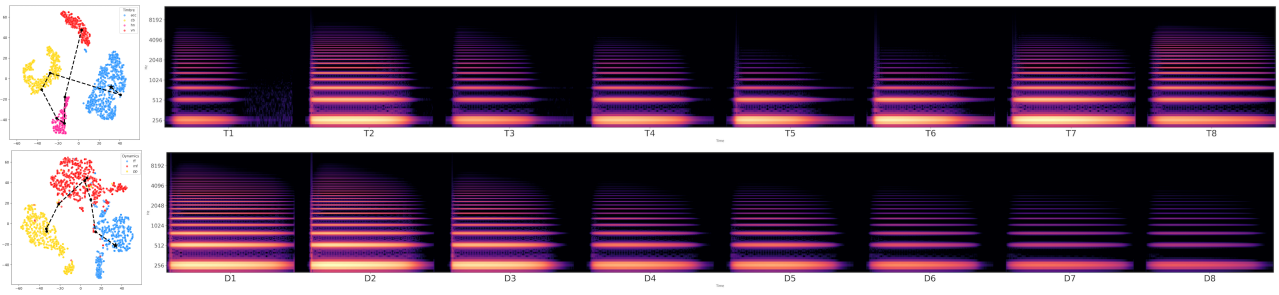


Fig. 4: Example of latent trajectories in $t = 4$ configuration, t space (top) and d space (bottom), with a fixed envelope and pitch contours of 2.5 sec, 261 Hz. The t-SNE plots on the left represent the actual trajectories (★ = starting point).

with a traditional DDSP with single GRU layer and loudness contour instead of envelopes. The baseline metrics are overall comparable for both the Triple-VAE and our improved DDSP in terms of MR-STFT and MSE. However, FAD metrics in the baseline (see Tab. 1, last row) are almost twice the values observed for the same configuration in our framework ($t = 2$, first row in 1), namely the one performing lowest in our experiments. Moreover, we still achieved VGG FAD results in line with the state-of-the-art considering similar approaches. For instance, Liu et al. reported FADs spanning from 1.529 to 7.150 for diverse sound effects reconstructed by their model [23]; Giudici et al. obtained comparable results on individual instruments, from a best FAD of 0.54 on trumpet to a worst one of 2.80 on flute, however exploiting individual models trained separately on each instrument [9]; similarly, Caspe et al. achieved FAD values of 2.73, 1.62, and 1.67 on flute, violin, and trumpet, also with a monotimbral DDSP with FM synthesizer [24]; finally, Ye et al. operate frame-wise timbre-pitch disentanglement, but achieved worst values of of about ~5 and ~12 on individual saxophone and violin models [25].

These results show that our model is capable of handling the multitimbral setting, while leveraging a compact and global latent representation which enhances interpretability in pure-generative settings.

4.1. Experiments

Our framework unlocks a variety of creative applications. New sounds can be generated by sampling the latent and inferring the model with envelope/pitch curves, achieving perceptual realism. We argue that such realism is motivated by the model being forced to learn temporal dependencies in relation to global embeddings: for instance, observing the latent trajectories drawn in Fig 4, it is possible to notice how, despite all sounds being generated by a constant ASR envelopes, decays of individual harmonic partials are unique to each sample. Secondly, is it possible to smoothly interpolate between different samples relying solely on the respective μ_t and μ_d, guiding the generation with a unified envelope and pitch contour that, exploiting disentanglement, may be completely unrelated to those of the original samples - Fig 5a. Furthermore, sounds can be manipulated in unconventional ways, e.g. performing out-of-range transpositions, expand attack transients, or create effects that are not possible in the native instrumental asset - Fig. 5b. Finally, the model is also capable of handling sharp shifts in the control signals while still preserving the natural quality of the sounds: attack transients are often rendered as quick bursts of noise, mimicking the behavior of real playing - Fig. 6.

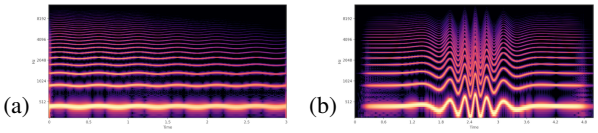


Fig. 5: Examples of (a) linear interpolation between a *va_ff* and *fl_pp*; (b) extreme $\pm 5th$ vibrato on an *acc_ff*.

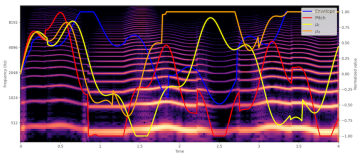


Fig. 6: Example of a 4 seconds sound generated by random curves on all inputs (μ_t and μ_d are averaged over $ls = 8$).

5. CONCLUSIONS

This work presents a novel approach towards controllable multitimbral audio synthesis within a DDSP framework, introducing disentangled timbre and dynamics embeddings. By employing a Triple-VAE architecture, we enforce semantic separation between key musical attributes. The replacement of loudness contours with normalized envelope curves, combined with a two-layer GRU architecture, enables the model to learn complex temporal dependencies from global, reduced representations. Experimental results on the TinySOL dataset demonstrates the framework's effectiveness across multiple configurations, achieving flexible controls while retaining audio realism, with potential applications on real-time sound synthesis. Future work involves expanding the framework to larger instrument vocabularies, performing individual dimensions analysis, and exploring synthesis techniques in challenging scenarios such as inharmonic/noise textures.

6. REFERENCES

[1] N. Devis, N. Demerlé, S. Nabi, D. Genova, and P. Esling, "Continuous descriptor-based control for deep audio synthesis," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[2] N. Demerlé, P. Esling, G. Doras, and D. Genova, "Combining audio control and style transfer using latent diffusion," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2024, pp. 721–728.

[3] G. Narita, J. Shimizu, and T. Akama, "Ganstrument: Adversarial instrument sound synthesis with pitch-invariant instance conditioning," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[4] O. Barkan, S. Shvartzman, N. Uzrad, M. Laufer, A. Elharar, and N. Koenigstein, "Inversynth ii: Sound matching via self-supervised synthesizer-proxy and inference-time finetuning," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2023, pp. 642–648.

[5] F. Bruford, F. Blang, and S. Nercessian, "Synthesizer sound matching using audio spectrogram transformers," in *Int. Conf. on Digital Audio Effects (DAFx24)*, 2024, pp. 135–138.

[6] J. Engel, C. Gu, and A. Roberts, "Ddsp: Differentiable digital signal processing," in *Int. Conf. on Learning Representations (ICLR)*, 2020.

[7] B. Hayes, J. Shier, G. Fazekas, A. McPherson, and C. Saitis, "A review of differentiable digital signal processing for music and speech synthesis," *Frontiers in Signal Processing*, vol. 3, 2024.

[8] G. Dal Santo, G. M. De Bortoli, K. Prawda, S. J. Schlecht, and V. Välimäki, "Flamo: An open-source library for frequency-domain differentiable audio processing," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[9] G. A. Giudici, F. Caspe, L. Gabrielli, S. Squartini, and L. Turchet, "Distilling ddsp: Exploring real-time audio generation on embedded systems," in *Audio Engineering Society Convention (AES)*, 2025, vol. 73, pp. 331–343.

[10] P. Agrawal, T. Koehler, Z. Xiu, P. Serai, and Q. He, "Ultra-lightweight neural differential dsp vocoder for high quality speech synthesis," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10066–10070.

[11] M. Kawamura, T. Nakamura, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Differentiable digital signal processing mixture model for synthesis parameter extraction from mixture of harmonic sounds," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 941–945.

[12] F. A. Dal Rí, G. A. Giudici, L. Turchet, and N. Conci, "Exploring multi-descriptor disentangled representations of acoustic instrument notes," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2025, pp. 416–420.

[13] 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 *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2019, pp. 746–753.

[14] Y.-J. Luo, K. W. Cheuk, W. Choi, W.-H. Liao, K. Toyama, T. Uesaka, K. Saito, C.-H. Lai, Y. Takida, and S. Dixon, "Disentangling multi-instrument music audio for source-level pitch and timbre manipulation," in *NeurIPS 2024 Workshop on Audio Imagination*, 2024.

[15] J. Turian and M. Henry, "I'm sorry for your loss: Spectrally-based audio distances are bad at pitch," in *NeurIPS Workshop: "I Can't Believe It's Not Better!"*, 2020.

[16] Y. Wu, E. Manilow, Y. Deng, R. Swavely, K. Kastner, T. Cooijmans, A. Courville, C.-Z. A. Huang, and J. Engel, "Midi-ddsp: Detailed control of musical performance via hierarchical modeling," in *Int. Conf. on Learning Representations (ICLR)*, 2022.

[17] A. Shewalkar, "Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, 2019.

[18] 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 *Int. Computer Music Conf. (ICMC)*, 2020.

[19] S. Li, S. Liu, L. Zhang, X. Li, Y. Bian, C. Weng, Z. Wu, and H. Meng, "Snakegan: A universal vocoder leveraging ddsp prior knowledge and periodic inductive bias," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2023.

[20] E. Gutiérrez, F. Font, X. Serra, and L. Wyse, "A statistics-driven differentiable approach for sound texture synthesis and analysis," in *Int. Conf. on Digital Audio Effects (DAFx)*, 2025.

[21] D. Roblek, K. Kilgour, M. Sharifi, and M. Zuluaga, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Interspeech: Annual Conf. of the Int. Speech Communication Association*, 2019, pp. 2350–2354.

[22] L. Renault, R. Mignot, and A. Roebel, "Ddsp-piano: A neural sound synthesizer informed by instrument knowledge," *AES Journal of the Audio Engineering Society*, vol. 71, no. 9, pp. 552–565, 2023.

[23] Y. Liu, C. Jin, and D. Gunawan, "DDSP-SFX: Acoustically-Guided Sound Effects Generation with Differentiable Digital Signal Processing," in *Int. Conf. on Digital Audio Effects (DAFx24)*, 2024, pp. 216–221.

[24] F. Caspe, A. McPherson, and M. Sandler, "DDX7: Differentiable FM Synthesis of Musical Instrument Sounds," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2022, pp. 608–616.

[25] L. Ye, G. Schuller, and M. Imran, "Instrumental timbre transfer based on disentangled representation of timbre and pitch," in *Asilomar Conf. on Signals, Systems, and Computers*. IEEE, 2024, pp. 1423–1426.