Learning Interpretable Features in Audio Latent Spaces via Sparse Autoencoders

Nathan Paek
Stanford University
nathanjp@stanford.edu

Yongyi Zang
Smule Labs
zyy0116@gmail.com

Qihui Yang
University of California, San Diego
qiy009@ucsd.edu

Randal Leistikow
Smule Labs
randal.leistikow@smule.com

Abstract

While sparse autoencoders (SAEs) successfully extract interpretable features from language models, applying them to audio generation faces unique challenges: audio's dense nature requires compression that obscures semantic meaning, and automatic feature characterization remains limited. We propose a framework for interpreting audio generative models by mapping their latent representations to human-interpretable acoustic concepts. We train SAEs on audio autoencoder latents, then learn linear mappings from SAE features to discretized acoustic properties (pitch, amplitude, and timbre). This enables both controllable manipulation and analysis of the AI music generation process, revealing how acoustic properties emerge during synthesis. We validate our approach on continuous (DiffRhythm-VAE) and discrete (EnCodec, WavTokenizer) audio latent spaces, and analyze DiffRhythm, a state-of-the-art text-to-music model, to demonstrate how pitch, timbre, and loudness evolve throughout generation. While our work is only done on audio modality, our framework can be extended to interpretable analysis of visual latent space generation models.

1 Introduction

As powerful neural networks become more integrated into society, their lack of interpretability raises a significant concern [11]. To address this challenge, sparse autoencoders (SAEs) have emerged as a key tool in mechanistic interpretability research [19, 4, 17]. They are motivated by the polysemantic hypothesis [19, 7, 15]: that neurons encode more features than dimensions by superposing multiple concepts. SAEs work by finding sparse directions in activation space to isolate these underlying, disentangled features. This approach has proven effective in large language models (LLMs), where SAEs can extract highly monosemantic features that are automatically characterized by using the model itself to summarize the results of token-level perturbations [5].

However, extending this approach to audio generative networks presents fundamental challenges. Unlike text, audio is inherently dense [24], and thus typically requires learned compression through autoencoders before tokenization [14]. This compression step, whether producing continuous or discrete latent codes, obscures the semantic meaning of individual "tokens," making perturbation-based analysis less interpretable [24, 28]. Moreover, while language models excel at summarizing textual patterns, current audio understanding models are not yet capable of providing an equally robust automatic characterization of SAE feature behaviors [23, 27]. These limitations necessitate new approaches for interpretable feature discovery in audio generative systems.

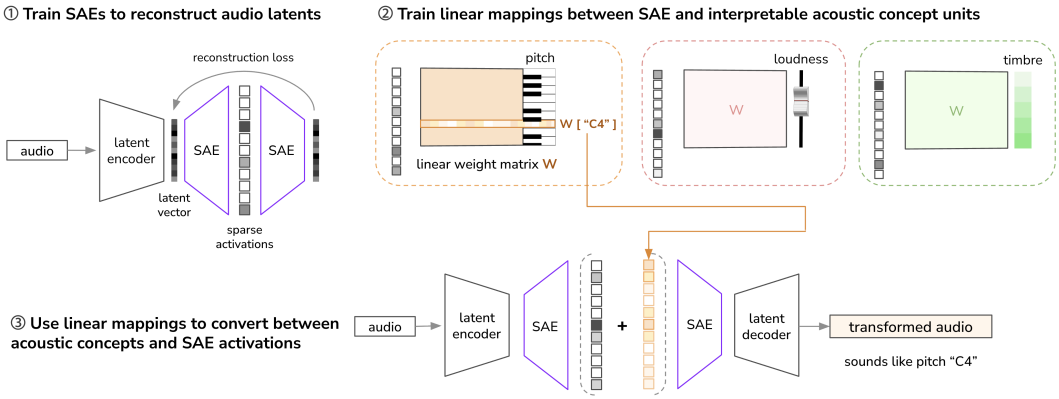

Figure 1: Framework for interpreting and controlling audio generative models through sparse features learned on their generation space. Sparse autoencoders extract interpretable features from audio latents, which are then linearly mapped to acoustic concepts. Control vectors extracted from these linear mappings can then be used to transform audio.

In this work, we propose a novel framework for understanding audio generative models by analyzing their latent space representations through human-interpretable acoustic concepts. Our approach proceeds in three stages. First, we train SAEs on the latent representations of audio autoencoders to extract sparse features. Second, we learn linear mappings from these SAE features to human-interpretable acoustic concepts: pitch, amplitude, and timbre (represented here by spectral centroid as a simplified proxy [21, 20]). To enable discrete analysis, we quantize each acoustic property into interpretable "units": pitch is discretized according to the Western tonal system (e.g., C4, C#4), while amplitude and spectral centroid are binned with equal spacing within their physical ranges. The effectiveness of linear mappings suggests that SAE features already encode acoustic properties in a near-linear fashion, validating the hypothesis that these learned representations align with human-interpretable concepts. Finally, by decomposing the audio synthesis process into an interpretable feature hierarchy, our framework traces how specific acoustic properties emerge. We empirically validate this approach on DiffRhythm, a state-of-the-art text-to-music model. Although our experiments focus on audio, we believe this framework is generalizable to other generative models that operate within learned latent spaces, including those for image and video.

2 Methodology

2.1 Sparse Autoencoder Training

We train SAEs on latent representations from three pretrained audio encoders: the continuous VAE space of Stable Audio Open and DiffRhythm [8, 18], and the discrete latent spaces of EnCodec [6] and WavTokenizer [12]. To address the unique requirements of audio latents, we modify the standard SAE architecture by adding an RMS normalization layer after the ReLU activation. This modification maintains consistent activation magnitudes and, as we empirically found, prevents out-of-distribution artifacts during feature manipulation. Following standard practice [5], we optimize the SAEs using a composite loss function:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \lambda \|\mathbf{h}\|_1 \tag{1}$$

where the first term ensures reconstruction fidelity and the L_1 penalty promotes sparsity in the hidden activations $\mathbf{h}$. We conduct systematic grid searches over hidden dimensionalities (ranging from $4\times$ to $256\times$ the input dimension) and sparsity coefficients λ (ranging from 0.005 to 0.15) to identify optimal configurations for each latent space.

2.2 Linear Mapping to Acoustic Concepts

To connect SAE features to interpretable acoustic properties, we train linear probes that predict discretized audio attributes from sparse activations. Given a latent vector $\mathbf{x} \in \mathbb{R}^d$, our SAE produces sparse features:

$$\mathbf{h} = \text{ReLU}(W_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}), \quad \mathbf{f} = \text{RMSNorm}(\mathbf{h}) \tag{2}$$

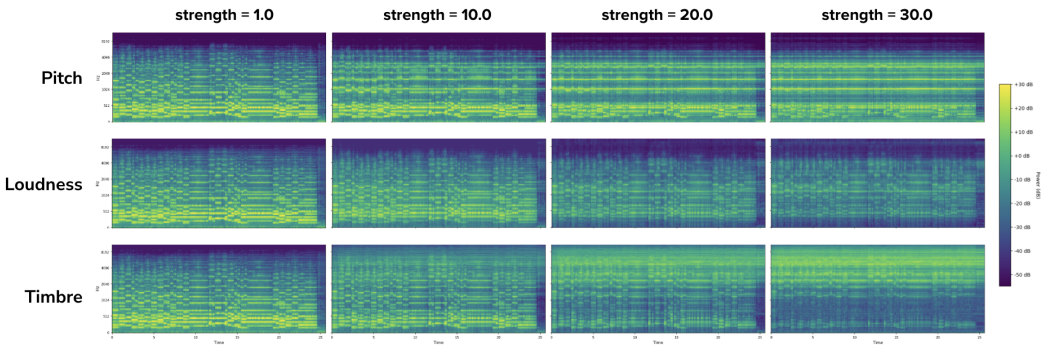

Figure 2: Controlled audio manipulation via control vectors. When α increases, isolated changes in pitch (imminent C5), amplitude (decreasing loudness), and timbre (brightening via high-frequency emphasis) can be observed. Corresponding audio samples can be found here: `https://anonymous.4open.science/r/audio_samples-A301/`

where $\mathbf{f} \in \mathbb{R}^m$ are the normalized features used for both reconstruction and interpretation.

For each acoustic attribute $a \in \{\text{pitch}, \text{amplitude}, \text{timbre}\}$, we first extract continuous measurements from the audio: pitch via CREPE [13], amplitude via windowed RMS energy using librosa [16], and timbre via windowed spectral centroid using librosa. We then discretize these continuous curves into K_a classes (pitch using logarithmic bins aligned with MIDI note numbers, and amplitude/timbre using linear bins) and train a linear classifier:

$$p^{(a)} = \text{softmax}(W^{(a)}\mathbf{f} + \mathbf{b}^{(a)}) \tag{3}$$

where $W^{(a)} \in \mathbb{R}^{K_a \times m}$ maps SAE features to class logits. The linearity provides bidirectional interpretability, as the contribution of SAE feature j to acoustic class k is simply $c^{(a)}_{j \to k} = W^{(a)}_{kj} \cdot f_j$, where the weights $W^{(a)}_{kj}$ reveal both which features encode specific acoustic properties and how acoustic concepts decompose into SAE features. For targeted intervention, we leverage this linearity directly by adding the scaled probe weight vector $\alpha \cdot \mathbf{w}k^{(a)}$ (*"control vectors"*) to the SAE features to shift the audio toward acoustic class k. After re-normalizing to maintain valid activation magnitudes, we decode through both SAE and audio decoders to generate the modified audio (shown in Figure 2).

3 Experiments

3.1 Acoustic Concept Mapping Discovery

Dataset. We use a composite dataset of ∼31 hours of audio sampled from several sources: Coco-Chorales [25]—11.2 hours of four-part Bach chorales, DAMP-VSEP [22]—11.7 hours of pop/rock singing, the Extended Groove MIDI Dataset [3]—7.8 hours of drums, GuitarSet [26]—24 minutes of solo guitar, and MAESTRO [10]—33 minutes containing classical piano.

Training SAEs on audio latent spaces. We conduct grid searches over SAE hidden dimensions $\{2048, 4096, 8192, 12288, 16384\}$ and sparsity coefficients $\lambda \in \{0.005, 0.01, 0.05, 0.1, 0.15\}$ for each audio encoder. The resulting SAEs exhibit distinct characteristics across latent spaces. DiffRhythm achieves sparsity ratios ranging from 0.65 to 0.98. WavTokenizer produces the sparsest representations (0.993–0.999), suggesting its discrete tokens already encode highly disentangled features. EnCodec demonstrates the widest sparsity range (0.55–0.95). Across all models, larger hidden dimensions consistently improve reconstruction quality.

Training linear probes from SAE features to acoustic concepts. We train linear probes to predict pitch (with 66 bins spanning the pitch range present in our dataset), loudness (20 bins), and timbre (20 bins) from SAE features. Plotting the probe classification accuracy on a test set vs. the sparsity of its SAE in Figure 3 shows a hierarchy of linear decodability across acoustic properties. Pitch proves most linearly separable (0.75–0.87 accuracy) and remains stable across all sparsity levels, suggesting fundamental frequency encoding. EnCodec excels at loudness (0.56–0.63) compared to DiffRhythm and WavTokenizer (0.17–0.49). Timbre remains challenging across all models (0.17–0.46).

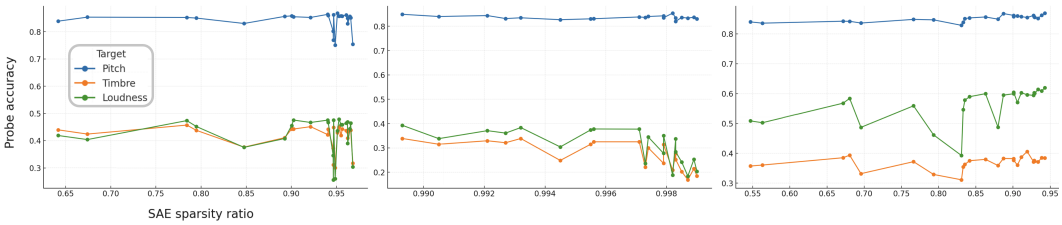

Figure 3: Linear probe accuracy for acoustic property classification across different sparsity levels. Left: Stable Audio Open/DiffRhythm VAE, Middle: WavTokenizer, Right: EnCodec.

Applying targeted interventions to audio samples. We test controllability on diverse audio sources (singing voice, drums, four-part harmony). Figure 2 shows a chordal audio sample from the CocoChorales dataset encoded with EnCodec and our highest-sparsity SAE (hidden_dim=16384, $\lambda = 0.1$). We apply control vectors targeting pitch (MIDI C5), timbre (spectral centroid class 17), and loudness (class 2) with strengths $\alpha \in \{1, 10, 20, 30\}$. As α increases, edits are isolated in the targeted attribute, while non-targeted properties remain largely preserved.

3.2 Generation Process Visualization

We demonstrate how our learned mappings can help us understand the audio generation process by analyzing DiffRhythm [18], a rectified flow model designed for full-length song synthesis. In this analysis, the model was configured to generate a 95-second audio segment, encompassing a verse and a chorus, over 32 inference steps. At each generation step $t \in \{0, ..., 31\}$, we extract the latent $\mathbf{X}_t \in \mathbb{R}^{C \times F}$, and decompose it through our SAE and linear probes to obtain acoustic concept activations $\mathbf{P}_t^{(a)} \in \mathbb{R}^{F \times K_a}$. After applying a mean pooling over frames (F), we obtain distributions $p_t^{(a)}$ for each attribute a. To quantify the evolution of acoustic properties, we track how these distributions interpolate from noise to the final audio. Specifically, for each attribute a and step t, we compute the per-class normalized L^1 distance:

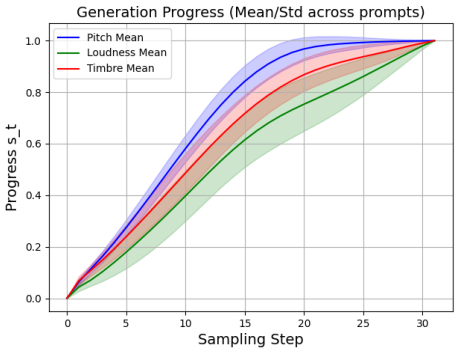

Figure 4: Probes Variation in Generation Progress

$$s_t^{(a)} = \frac{1}{K_a} \sum_{k=1}^{K_a} \frac{|p_{t,k}^{(a)} - p_{0,k}^{(a)}|}{|p_{T,k}^{(a)} - p_{0,k}^{(a)}|} \tag{4}$$

where $s_t^{(a)} \in [0, 1]$ measures the progression from initial noise ($t = 0$) toward the final acoustic structure ($t = T = 31$). This reveals when different acoustic properties emerge during generation. We sample 500 prompts from MusicCaps [1], then plot the mean and standard deviation of the generation progress in Figure 4, which indicates a clear hierarchy. Pitch converges first (around step 21), followed by timbre, while loudness converges last and remains unresolved by the final step. This coarse-to-fine progression suggests the model establishes fundamental frequency before refining textural and dynamic details.

4 Conclusion

We present a framework for interpreting audio generative models by mapping their latent representations to human-interpretable acoustic properties through sparse autoencoders and linear probes. Our experiments demonstrate that SAE features naturally align with acoustic properties, enabling both controllable manipulation and better understanding of music generative models.

In future work, we plan to apply our method to other generative architectures, such as RAVE [2], ACE-Step [9], and AudioLDM [14]. Beyond the three acoustic properties explored here, we will train

probes for richer audio features such as rhythm, harmony, and instrument identity. Finally, we aim to use these interpretable features to directly guide generation behavior during inference, potentially enabling fine-grained control over specific attributes while maintaining generation quality.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[2] Antoine Caillon and Philippe Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis, 2021.

[3] Lee Callender, Curtis Hawthorne, and Jesse Engel. Improving perceptual quality of drum transcription with the expanded groove midi dataset, 2020.

[4] Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, and Chris Olah. Curve circuits. *Distill*, 6(1):e00024–006, 2021.

[5] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

[6] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022.

[7] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[8] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[9] Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. Ace-step: A step towards music generation foundation model, 2025.

[10] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[11] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

[12] Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling, 2025.

[13] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 161–165. IEEE, 2018.

[14] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[15] Simon C Marshall and Jan H Kirchner. Understanding polysemanticity in neural networks through coding theory. *arXiv preprint arXiv:2401.17975*, 2024.

[16] B. McFee. librosa/librosa: 0.11.0, March 2025.

[17] Elhage Nelson, Nanda Neel, Olsson Catherine, Henighan Tom, Joseph Nicholas, Mann Ben, Askell Amanda, Bai Yuntao, Chen Anna, Conerly Tom, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[18] Ziqian Ning, Huakang Chen, Yuepeng Jiang, Chunbo Hao, Guobin Ma, Shuai Wang, Jixun Yao, and Lei Xie. Diffrhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion, 2025.

[19] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[20] Emery Schubert and Joe Wolfe. Does timbral brightness scale with frequency and spectral centroid? *Acta acustica united with acustica*, 92(5):820–825, 2006.

[21] Emery Schubert, Joe Wolfe, Alex Tarnopolsky, et al. Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of the international conference on music perception and cognition, North Western University, Illinois*, pages 112–116. sn, 2004.

[22] I. Smule. DAMP-VSEP: Smule Digital Archive of Mobile Performances - Vocal Separation. Zenodo, oct 2019.

[23] Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*, 2025.

[24] Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H Liu, and Hung-yi Lee. Towards audio language modeling–an overview. *arXiv preprint arXiv:2402.13236*, 2024.

[25] Yusong Wu, Josh Gardner, Ethan Manilow, Ian Simon, Curtis Hawthorne, and Jesse Engel. The chamber ensemble generator: Limitless high-quality mir data via generative modeling. *arXiv preprint arXiv:2209.14458*, 2022.

[26] Q. Xi, R. Bittner, J. Pauwels, X. Ye, and J. P. Bello. GuitarSet: A Dataset for Guitar Transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, sep 2018.

[27] Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. Towards holistic evaluation of large audio-language models: A comprehensive survey. *arXiv preprint arXiv:2505.15957*, 2025.

[28] Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705, 2025.