GENERALIZED MULTI-SOURCE INFERENCE FOR TEXT CONDITIONED MUSIC DIFFUSION MODELS

Emilian Postolache[1], Giorgio Mariani [1], Luca Cosmo [2], Emmanouil Benetos [3], Emanuele Rodolà [1]

[1] Sapienza University of Rome, Italy
[2] Ca' Foscari University of Venice, Italy
[3] Centre for Digital Music, Queen Mary University of London, UK

Motivation

- State-of-the-art generative models for music [1] typically output a single "final" mixture, which is difficult to manipulate.
- A new class of *compositional* generative models for music operates on sub-constituents of musical tracks (stems).

- The first compositional model in continuous domain (as opposed to symbolic) is the Multi-Source Diffusion Model (MSDM) [3].
  - Generate all stems.
  - Perform accompaniment of stems based on others.
  - Separate stems.
- **Problem:** MSDM requires stem-separated datasets containing considerably less data than mixture datasets.
- **Objective:** Develop a method for compositional music generation called *Generalized Multi-Source Diffusion Inference (GMSDI)* that does not require stem-separated datasets.

Preliminaries

- In the absence of separated sources $\{\mathbf{x}_k\}$ for mixed tracks $\mathbf{y}$, we resort, for training, to a dataset with mixes $\mathbf{y}$ and text embeddings $\mathbf{z}$ describing the constituent stems.
- A text embedding $\mathbf{z}$ can be obtained:
  - By mapping a textual description $\mathbf{q}$ with a text-only encoder E_ϕ^{text}:
$$\mathbf{z} = E_\phi^{\text{text}}(\mathbf{q})$$
  - Via a text-audio contrastive encoder with independent branches E_ϕ^{contr} mapping the mixture itself:
$$\mathbf{z} = E_\phi^{\text{contr}}(\mathbf{y})$$
- We assume the embeddings have the form: $\mathbf{z} = \mathbf{z}_1 \otimes \cdots \otimes \mathbf{z}_K$, with each $\mathbf{z}_k$ describing a source $\mathbf{x}_k$ in $\mathbf{y}$.
- We train a (score-based) diffusion model [4] with such data:

$$\nabla_{\mathbf{y}(t)} \log p(\mathbf{y}(t) \mid \mathbf{z}) \approx S_\theta(\mathbf{y}(t), \mathbf{z}, \sigma(t)) \quad (1)$$

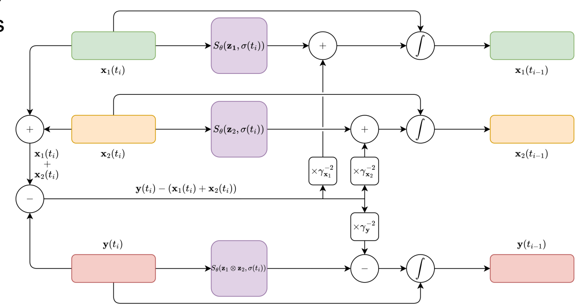
Method

- The idea is that, by leveraging (1), we can parameterize the score of individual sources:
$$\nabla_{\mathbf{x}_k}(t) \log p(\mathbf{x}_k(t) \mid \mathbf{z}_k) \approx S_\theta(\mathbf{x}_k(t), \mathbf{z}_k, \sigma(t))$$
- With this, we can set up an inference procedure where we sample in parallel both the candidate sources $\mathbf{x}_k$ and a mix $\mathbf{y}$, linking them with a Gaussian likelihood at each step. This inference procedure is defined by:

$$\begin{cases} S_\theta(\mathbf{x}_k(t), \mathbf{z}_k, \sigma(t)) + \frac{1}{\gamma^2_{\mathbf{x}_k}}\left(\mathbf{y}(t) - \sum_{l=1}^K \mathbf{x}_l(t)\right) \\ S_\theta(\mathbf{y}(t), \mathbf{z}_1 \otimes \cdots \otimes \mathbf{z}_K, \sigma(t)) + \frac{1}{\gamma^2_{\mathbf{y}}}\left(\sum_{l=1}^K \mathbf{x}_l(t) - \mathbf{y}(t)\right) \end{cases} \quad (2)$$

- The method separates the sources while generating them! Generating a mixture is necessary for informing the sources about a shared context.
- For accompaniment generation we simply add the conditioning mixture (perturbed at time t) $\sum_{i \in \mathcal{I}} \mathbf{x}_i(t)$ to the sum $\sum_{j \in \mathcal{J}} \mathbf{x}_j(t)$ of the sources we are generating:

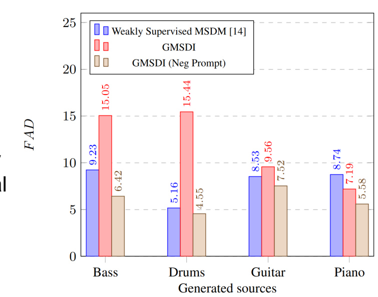
$$\sum_{l=1}^K \mathbf{x}_k(t) = \sum_{i \in \mathcal{I}} \mathbf{x}_i(t) + \sum_{j \in \mathcal{J}} \mathbf{x}_j(t)$$

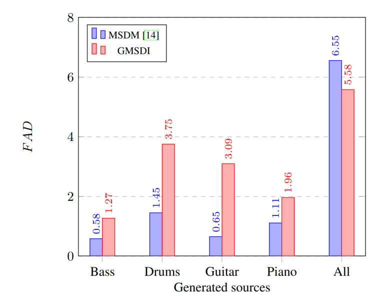


Experiments

- Quantitative experiments on Slakh2100 [4], trained with E_ϕ^{text} using supervised text data (tags) and mixtures. Qualitative experiments on MTG-Jamendo, trained with E_ϕ^{contr}.
- **Right-Top:** We study how well we can parameterize single sources having trained over mixtures, comparing with the weakly supervised MSDM [3] using FAD. Negative prompting is essential for good parametrizations.
- **Right-Bottom:** We use the sub-FAD protocol of [3] to test the coherence of the generated accompaniments and the FAD for unconditional generation.
- **Bottom:** We use the Dirac likelihood of [3] with our parameterized model for source separation (separating all or extracting one source). We obtain non-negligible results on separation despite training only with mixtures.

Model	Bass	Drums	Guitar	Piano	All
Demucs + Gibbs (512 steps) [27]	17.16	19.61	17.82	16.32	**17.73**
Weakly Supervised MSDM [14]	19.36	20.90	14.70	14.13	17.27
MSDM [14]	17.12	18.68	15.38	14.73	16.48
GMSDI Separator	9.76	15.57	9.13	9.57	11.01
GMSDI Extractor	11.00	10.55	9.52	10.13	10.30
Ensamble	11.00	15.57	9.52	10.13	**11.56**





References

1. Evans, Zach, et al. "Fast Timing-Conditioned Latent Audio Diffusion." arXiv preprint arXiv:2402.04825 (2024).
2. Mariani, Giorgio, et al. "Multi-Source Diffusion Models for Simultaneous Music Generation and Separation." *ICLR*. 2024.
3. Song, Yang, et al. "Score-Based Generative Modeling through Stochastic Differential Equations." *ICLR*. 2020.
4. Manilow, Ethan, et al. "Cutting music source separation some Slakh: A `dataset to study the impact of training data quality and quantity." *WASPAA*. IEEE, 2019.
5. Bogdanov, Dmitry et al. "The MTG-Jamendo Dataset for Automatic Music Tagging." ICML Machine Learning for Music Discovery Workshop, 2019.