SampleMatch: Drum Sample Retrieval by Musical Context

Stefan Lattner

Sony Computer Science Laboratories (CSL), Paris

https://sites.google.com/view/samplematch

stefan.lattner@sony.com

Try it yourself!

Modern digital music production typically involves combining numerous acoustic elements to compile a piece of music. Important types of such elements are drum samples, which determine the characteristics of the percussive components of the piece. Artists must use their aesthetic judgement to assess whether a given drum sample fits the current musical context. However, selecting drum samples from a potentially large library is tedious and may interrupt the creative flow. In this work, we explore the **automatic drum sample retrieval based on aesthetic principles learned from data**. As a result, artists can rank the samples in their library by fit to some musical context at different stages of the production process (i.e., by fit to incomplete song mixtures). To this end, we use contrastive learning to maximize the score of drum samples originating from the same song as the mixture. We conduct a listening test to determine whether the human ratings match the automatic scoring function. We also perform objective quantitative analyses to evaluate the efficacy of our approach.

Method

We train the encoders using a **contrastive loss** called NT-Xent, where $\text{sim}(\cdot,\cdot)$ is the cosine similarity:

$$\mathcal{X}(Z) = -\log \frac{\exp\left(\text{sim}(\mathbf{u}_i, \mathbf{v}_j)/\tau\right)}{\sum_{l \neq j} \exp\left(\text{sim}(\mathbf{u}_i, \mathbf{v}_l)/\tau\right)}, \tag{1}$$

where $\{\mathbf{u}_i, \mathbf{v}_j\}$ is the encoding of a positive pair, $Z \in \mathbb{R}^{n \times d}$ are all representations of a training batch, τ is the temperature parameter, and we adopt the decoupled contrastive learning variant, that has shown to work better for smaller batch sizes, by removing the positive pair from the denominator (i.e., $l \neq j$).

Regularizations

We combine the contrastive loss with the **variance and covariance regularization** used in VICReg. The variance regularization term is defined as a hinge function that penalizes variances of latent features along the batch dimension that are smaller than 1 as

$$\mathcal{V}(Z) = \frac{1}{d} \sum_{j=1}^{d} \max\left(0, 1 - S(\mathbf{z}_{:,j}, \epsilon)\right), \tag{2}$$

where Python slicing notation is used, and S is the regularized standard deviation

$$S(x, \epsilon) = \sqrt{\text{Var}(x) + \epsilon}. \tag{3}$$

The covariance regularization penalizes non-zero off-diagonal entries in the covariance matrix of each batch, leading to a decorrelation of the latent dimensions:

$$\mathcal{C}(Z) = \frac{1}{d} \sum_{i \neq j} [C(Z)]_{i,j}^2, \tag{4}$$

where C is the covariance matrix.

Data

We used a dataset of **electronic music** (4830 "remix packs") and 885 **pop/rock songs** of 44.1 kHz sample rate for training and evaluation. From every percussion track in the dataset, we extract so-called "one-shots", single hits with the respective percussion instrument (63042 in total). Based on their filenames, we categorize the extracted drum samples into 6 categories which are {kick, snare, hi-hat, ride, crash, toms}.

Training

As encoders, we use the **EfficientNet-B4** (pre-trained on the ImageNet dataset). We input **log-mel spectrograms** with an STFT window length of 2048, a hop length of 512, and 128 resulting mel bins, considering the whole frequency range (fmax $= 22050$). The encoders are trained by the ADAM optimizer, with a batch size of 190, a learning rate of $3\text{e-}4$, and a weight decay factor of $3\text{e-}5$. **Data augmentation**: Gaussian noise, time-stretch, reducing the gain, and time shift.

Results

User Study

Participants prefer the highly scored samples over random samples approximately twice as often.

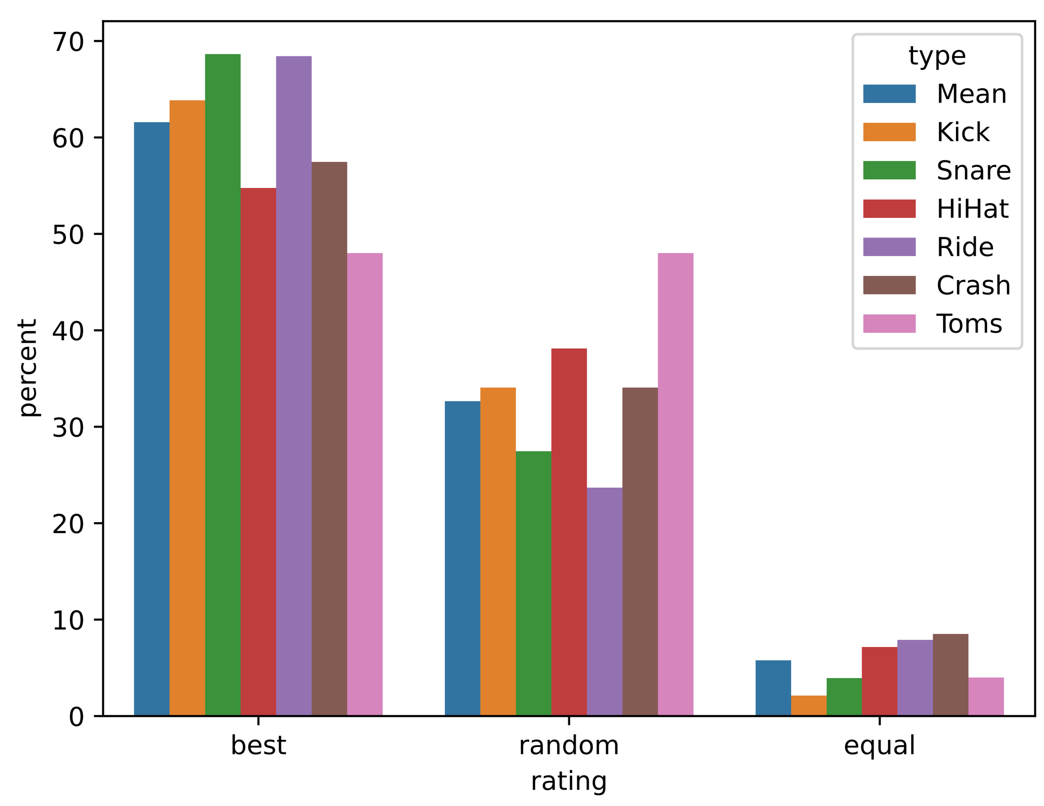

Figure 1: Preference ratings of participants in the user study, separated by percussion type (the blue bar "Mean" shows the mean of all ratings). "best" are mixtures with samples that *scored highest* by our method, and "random" denotes mixtures with *random samples* from the data set. An "equal" rating means no particular preference.

Correlation Analysis

For interpretability, we perform correlation analyses between samples that are close in the learned space.

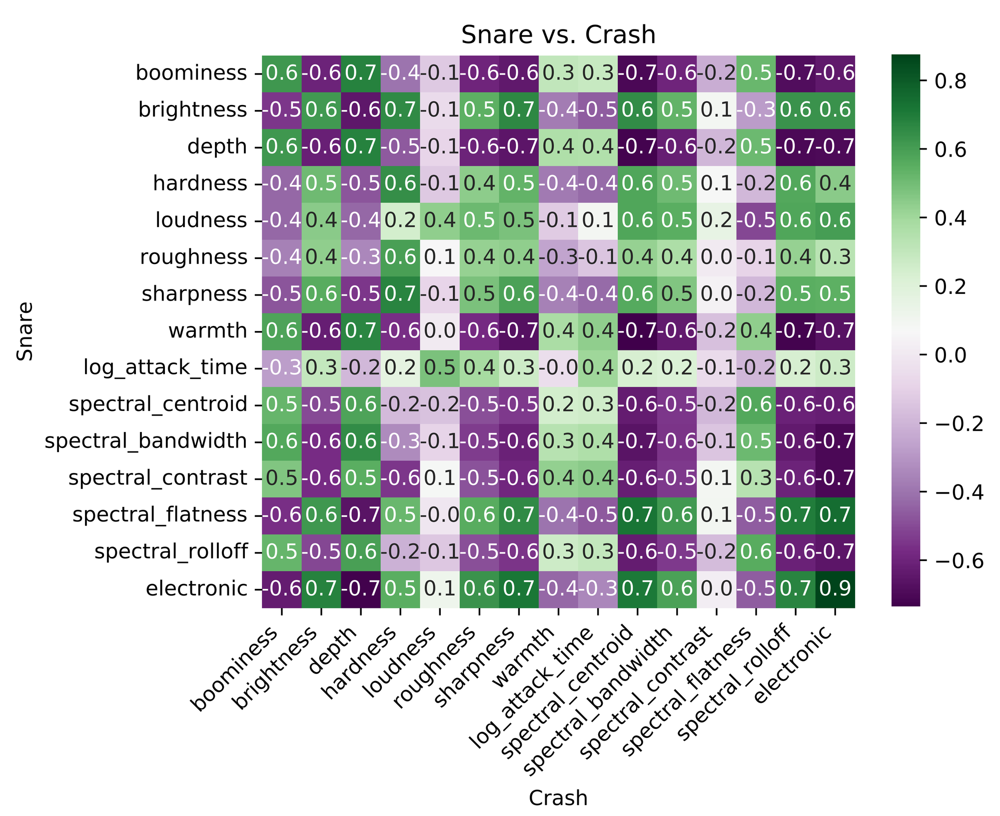

Figure 2: Correlations between perceptual and spectral features (and electronic / acoustic indicator) of Snare and Crash drum samples that are close in the latent space (i.e., scored to fit well in the same musical context).

Quantitative Evaluation

The main evaluation metric is the **Mean Normalized Rank** summarized as

$$R_{\text{mn}} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{\text{rank}_i}{N}. \tag{5}$$

Ablation studies of different model configurations unveils that VICReg regularization, pre-training, data augmentation and sparse mixing improves results.

	Queries: full mixtures				
Variant	$\mathcal{X}$	R_{mn}	R_{md}	L_q	L_k
`2Enc+PTrain+Aug+VCReg+SMix`	3.614	**0.105**	**0.032**	0.9940	0.9767
`2Enc+PTrain+Aug+VCReg+SMix+QSInv`	3.718	0.120	0.037	0.9900	0.9782
`2Enc+PTrain+Aug+VCReg`	4.001	0.124	0.061	0.9825	0.9732
`2Enc+PTrain+VCReg+SMix`	3.742	0.128	0.037	0.9945	**0.9895**
`2Enc+PTrain+Aug+SMix`	**3.575**	0.116	**0.032**	**0.9946**	0.9774
`2Enc+Aug+VCReg+SMix`	5.235	0.458	0.432	0.7268	0.7629
`2Enc+Aug+VCReg+SMix+QSInv`	4.174	0.164	0.079	0.9826	0.9566
`PTrain+Aug+VCReg+SMix`	3.853	0.121	0.047	0.9812	0.9809
`2Enc+PTrain`	3.883	0.140	0.053	0.9925	0.9795

	Queries: sparse mixtures				
Variant	$\mathcal{X}$	R_{mn}	R_{md}	L_q	L_k
`2Enc+PTrain+Aug+VCReg+SMix`	**3.761**	**0.124**	0.043	**0.9905**	0.9763
`2Enc+PTrain+Aug+VCReg+SMix+QSInv`	3.818	0.136	0.047	0.9862	0.9768
`2Enc+PTrain+Aug+VCReg`	4.389	0.183	0.100	0.9635	0.9724
`2Enc+PTrain+VCReg+SMix`	3.780	0.137	**0.042**	0.9901	**0.9898**
`2Enc+PTrain+Aug+SMix`	3.812	0.135	0.043	0.9893	0.9790
`2Enc+Aug+VCReg+SMix`	5.237	0.470	0.451	0.7188	0.7480
`2Enc+Aug+VCReg+SMix+QSInv`	4.387	0.181	0.091	0.9768	0.9585
`PTrain+Aug+VCReg+SMix`	4.000	0.137	0.058	0.9768	0.9819
`2Enc+PTrain`	4.399	0.205	0.089	0.9821	0.9803

Table 1: Ablation study for different architectures and training scenarios tested on queries from full mixtures and queries from sparse mixtures (a sparse mixture is based on a random number n of stems, where $n > 1$).

PCA

When performing a PCA on the latent space, we see that acoustic and electronic samples are well-separated.

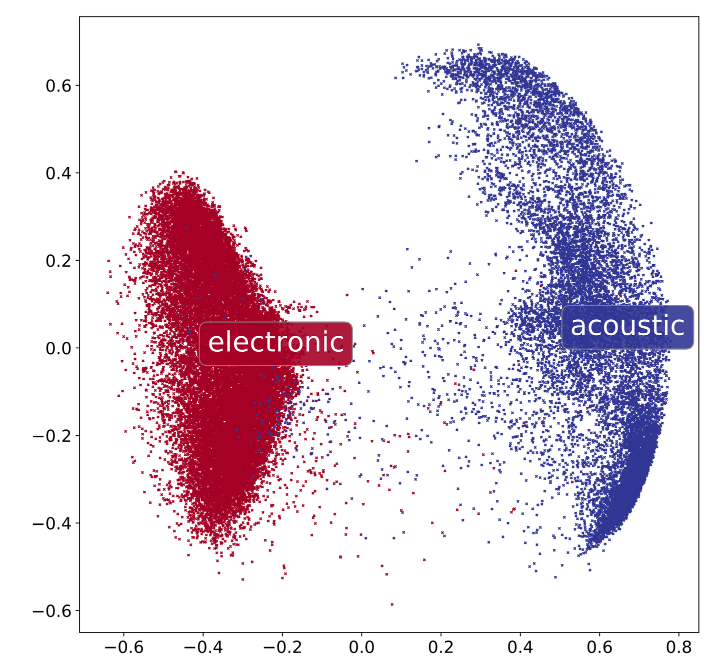

Figure 3: Principal Component Analysis (PCA) of drum sample encodings. Red dots indicate samples originating from electronic music and blue dots indicate samples originating from acoustic music.

Conclusion

Contrastive learning for drum samples and song mixes.

- VICReg regularizations, pre-training, augmentation and sparse mixing helps
- Users prefer automatically selected samples twice as often as random samples
- Correlation analysis unveils "rules"
- Electronic and acoustic samples well-separated

Future Work

- Extend to other instrument combinations