Zero-shot Crate Digging → DJ Tool retrieval using Speech Activity, Music Structure and CLAP embeddings

Iroro Orife

What are DJ Tools?

In genres like Hip-Hop, RnB, Reggae, Dancehall and every Electronic/Dance/Club style, DJ tools are a special set of audio files of short, simplified musical phrases retrieved, curated from existing music, or composted specifically to heighten the DJ's musical performance and creative mixing choices.

- Acapella loops, vocal chants, spoken word, one shot vocal samples
- Purely instrumental (no drums) piano, guitar, strings, horns
- Drum loops, solo percussion, drum breaks
- Sound effects: risers, sweeps, sirens, vinyl fx,
- Scratch and Battle loops

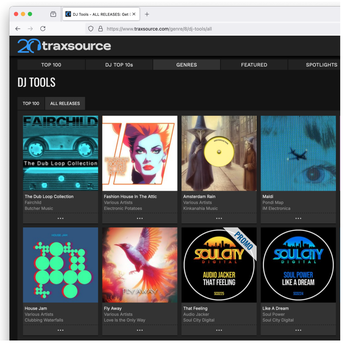

Figure 1: Traxsource DJ Tools shop

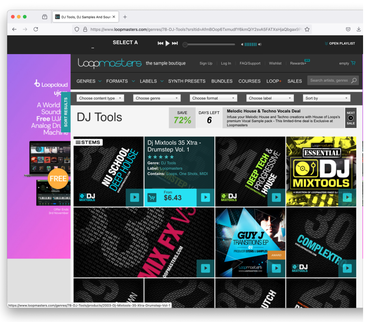

Figure 2: Loopmasters online shop. Note the specification of Loops, One-shots

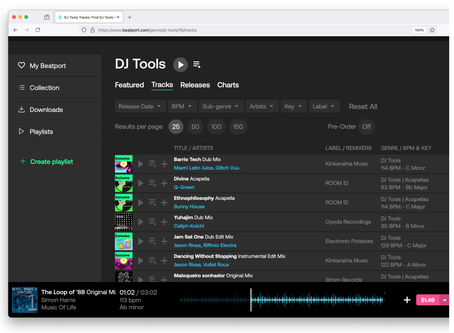

Figure 3: Beatport DJ Tools selection. Note the designated Genre, Key and BPM metadata

What is Zero-shot Audio Classification?

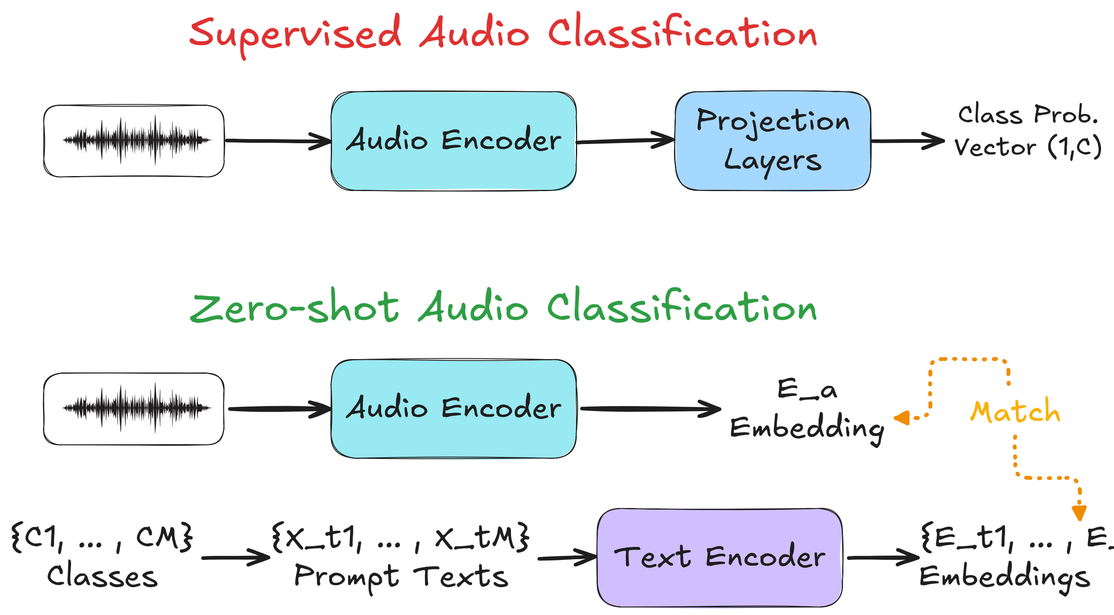

Figure 4

Speech Activity & Music Structure Analysis

Figure 5: A 5 minute Ragga Jungle song overlaid with detected speech, music activity and music-structural boundaries

DJ Tool class	Example text description
Acapella loops	"acapella, expressively sung vocal tracks, purely acoustic"
Sound effects	"siren, riser sound effects, whoosh, crash, synthetic"
Drums breaks	"drums, funky drummer, drum solo, Amen breakbeat"
Melodic hooks	"strings, solo guitar, piano melodies, synths instrumental"
Battle tracks	"vinyl scratch loop, turnatablist DJ battle sounds"

Table 1: In practise, text descriptions are more tortuous

Software Demo Implementation

The DJ-tool classification system and its dependencies are fully open-source software (OSS) written in Python. Firstly, we run MSAF[a] and SMAD[b] to generate raw activities and structural boundaries, saved as CSV files[2, 3]. Next, these are post-processed to yield lists of segment time ranges. Figure 5 shows the relationship between the boundaries and the speech activity for one song.

For zero-shot classification, we utilize the LAION version of CLAP[c] hosted on Huggingface[1, 4]. Finally, we manually create a list of descriptive text strings for each class, as listed in Table 1.

Algorithm 1: Zero-shot Crate Digging

Create text prompts for M classes $\{X_1^t, ..., X_M^t\}$
$\{E_1^t, ..., E_M^t\} \leftarrow TextEncoder(\{X_1^t, ..., X_M^t\})$
for *song* s_i *in the library* S_N **do**
  $W^s, W^m \leftarrow SMAD(s_i)$
  $B \leftarrow MSAF(s_i)$
  Adjust boundaries B_i to nearest W^s
  Use B to cut s_i to audio files $\{X_1^a, ..., X_N^a\}$
  for $j := 1$ *to* N **do**
    $E_j^a \leftarrow AudioEncoder(X_j^a)$
    $\vec{D} \leftarrow Similarity(E_j^a, \{E_1^t, ..., E_M^t\})$
    $\vec{z} \leftarrow softmax(\vec{D})$
    Store predicted class, the arg max of $\vec{z}$

[a] https://github.com/urinieto/msaf
[b] https://github.com/biboamy/TVSM-dataset/tree/master
[c] https://huggingface.co/laion/clap-htsat-unfused

Results: Making music with the system

We evaluated the system on 10+ albums from the author's DJ library →

- Vocal and drums tool classes perform better than short, burstier sound effects
- Performance is sensitive to *precise language*. Adding generic terms or entire genres (e.g. "hip-hop, drum and bass") to a class description led to markedly worse results overall.
- For best results, the class descriptions should not overlap
- One challenge was to determine optimal segmentation for CLAP classification, while slicing the song appropriately for use by DJs.
- Playing with the system and making music with the retreived DJ tools led to unexpectedly delightful mixing sessions.

References

[1] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang.
Clap: Learning audio concepts from natural language supervision, 2022.

[2] Y.-N. Hung, C.-W. Wu, I. Orife, A. Hipple, W. Wolcott, and A. Lerch.
A large tv dataset for speech and music activity detection.
EURASIP Journal on Audio, Speech, and Music Processing, 2022(1):21, 2022.

[3] O. Nieto and J. P. Bello.
Systematic exploration of computational music structure research.
In *ISMIR*, pages 547–553, 2016.

[4] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov.
Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation.
In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.