Rap2Beat: Generating Drum Grooves from A Cappella Rap Vocals

r14942082 賴奕錂 | r14942090 傅學惟 | r13942136 陳仲桓

Motivation

Original

+

Original

Mixed

- Rap vocals can fit multiple types of beats as long as the flow aligns with the downbeat.

Problem Formulation

- Given a rap vocal track as input, the task is to generate a corresponding drum accompaniment which is **rhythmically aligned and consistent with the vocal's flow and genre.**

Methodology

- Model Architecture
  - We fine-tune **Stable Audio Open** as our beat-generation backbone.
  - Instead of text prompts, the model is conditioned on **MIR features** extracted from the rap vocal.
- MIR Conditioning Signals (from vocal)
  - **Rhythmic**: downbeat, vocal onsets, syllable density(speed of words).
  - **Dynamic**: energy contour representing intensity changes.
  - **Structural**: verse/chorus segmentation and phrase boundaries.
- Structural Awareness
  - Different beat loops are generated for **verse** vs. **chorus** to create sectional contrast.
  - **Energy slope** and phrase **boundaries** guide transitions, allowing the model to produce build-ups or fill-ins at segment junctions.

Pipeline

- **Training:** Extract vocal + drum stems → compute MIR features → use them as conditioning to fine-tune **Stable Audio Open.**
- **Inference:** given a rap vocal, the model generates a drum track aligned with the flow and overall song structure.

Dataset here

- Scale: Comprises 92,371 collected rap songs from YouTube sources.
- Total Duration: Over 5,586 hours of raw audio content.
- Multilingual: Contains songs spanning 84 different languages, with English being the majority of the content.
- Core Content: The processed dataset provides distinct files for:
    - Separated Vocal Tracks
    - Accompaniment/Instrumental Tracks
    - Segmented Audio Clips (extracted from the full songs using Voice Activity Detection).
    - Lyrics Transcriptions (via Automatic Speech Recognition).

Dataset

Subset	DNSMOS Threshold	PPS Threshold	Primary Singer Threshold	Total Duration (h)	Average Segment Duration (s)
Orig Songs	-	-	-	5586.2	227.7
RapBank	-	-	-	4353.6	17.4
RapBank (English)	-	-	-	3830.1	17.3
Basic	2.5	12-35	0.8	1322.0	18.5
Standard	3.5	16-32	0.9	295.3	18.8
Premium	3.8	18-30	1.0	58.3	18.7

Expected Milestone

Stage 1:

- being able to generate two distinctive **repeating patterns** of drum loop for verses and choruses

Stage 2:

- being able to generate **variational patterns** within verses and choruses

Stage 3:

- non rule-base generation: let the model figure the structure out!

Expected Contribution

- Having variational beats for a specific rap verse to improve diversity
- Assisting music producers to visualize more rhythmic selections
- Enhancing personalization of the beat track for individual rappers