JukeDrummer: Conditional Beat-aware Audio-domain Drum Accompaniment Generation via Transformer VQ-VAE

Yueh-Kao Wu
Academia Sinica

Ching-Yu Chiu
National Cheng Kung University

Yi-Hsuan Yang
Taiwan AI Labs

Introduction

- **Drum Accpaniment** Generation: Generating coherent accompaniment drum tracks based on drumless audio tracks.
- **Audio Domain:** Instead of dealing with composition in the symbolic domain, our work directly deals with "sounds" and thereby has more difficulty and complexity.
- **Transformer VQ-VAE:** Similar to Jukebox purposed by OpenAI, we use vector-quantization variational autoencoders and the Transformer to implement our model.
- **Conditional Beat-aware:** In addition to the VQ-VAE and the Transformer, we purpose brand-new beat-aware modules to extract beat information as conidtions in order to improve the generation.

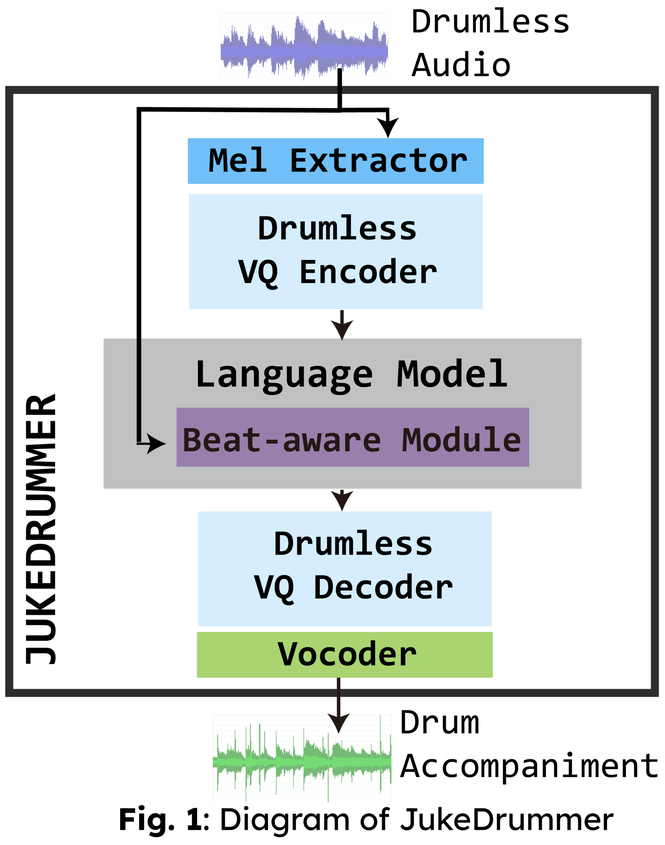

Fig. 1: Diagram of JukeDrummer

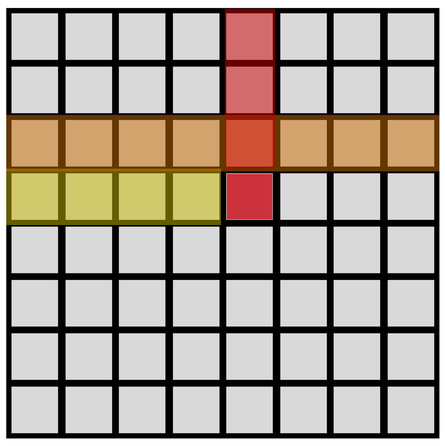

Fig. 2: Sparse attention pattern of Scalable Transformer.

VQ-VAE

- **Input:** Mel spectrograms (80 Mel-filter banks x 4096 sequence length)
- **Two VQ-VAE models:**
  - Drum VQ-VAE: train on drum data with 32 codebook size
  - Drumless VQ-VAE: train on drumless data with 1024 codebook size

Transformer

- **Input:** Drumless tokens are placed in the Transformer encoder, and drum tokens are placed in the Transformer decoder.
- **Scalable Transformer (Sparse Transformer purposed by OpenAI):** Transformer with sparse attention pattern. (Fig. 2)

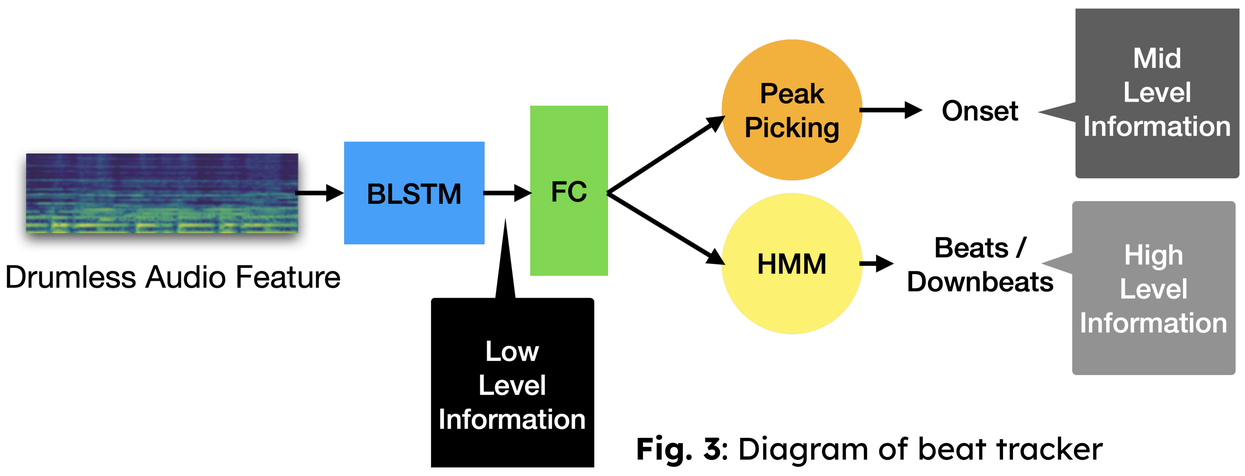

Fig. 3: Diagram of beat tracker

Beat-aware Module

- **Input:** 23.8-second wave clips
- We use pre-trained non-drum beats/downbeats tracker to extract beat information.
- The modules are placed right after position embedding.
- We study three different levels of information from different parts of the pre-trained tracker as conditions. (Fig. 3)

Experiment

- We purpose customized objective metrics and use them to compare rhythmic features between generated drum tracks and ground truth.
- **TrackEmd-MSE**: mean square error for the low-level tracker embedding. (Low-level information in Fig. 3)
- **Act-Entropy:** cross entropy for mid-level activation. (Mid-level information in Fig. 3)
- **B/DB-F1:** F1 score of beats/downbeats estimation using ground-truth drums as the reference.

Model	Objective metrics			Subjective MOS ($\in [1,5]$)				
	TrackEmd-MSE↓	Act-Entropy↓	B/DB-F1↑	$\mathbf{R}^{m/d}$	$\mathbf{S}^{m/d}$	$\mathbf{Q}^{d}$	$\mathbf{R}^{d}$	$\mathbf{O}^{d}$
seq2seq+beat (low)	**.068**±.023	**.928**±.189	**.340**	**3.27**	**3.44**	**3.39**	**3.37**	**3.20**
seq2seq+beat (mid)	.077±.022	.963±.200	.212	—	—	—	—	—
seq2seq+beat (high)	.080±.024	.938±.160	.132	—	—	—	—	—
seq2seq w/o beat	.081±.022	.968±.233	.110	1.78	2.34	2.95	2.58	2.05
decoder+beat (low)	**.068**±.023	**.931**±.200	**.339**	**3.05**	**3.34**	**3.33**	**3.07**	**3.17**
decoder w/o beat	.087±.025	.987±.240	.114	1.59	1.83	2.56	1.88	1.73
Real data (not vocoded)	—	—	—	4.39	3.95	3.85	4.61	4.17

- In subjective metrics, listeners are asked to rate mixtures (drumless + drum) on a 5-point Likert scale in following aspects: rhythmnic consistency $R^{m/d}$, stylisitc consistency $S^{m/d}$, audio quality Q^d , and rhythmn stablity R^d .
- The mixtures in the listening test include drum samples generated by four different models plus human-made drum samples.

Conclusion

- Beat-aware modules are essential to generate coherent drum accompaniments.
- Low-level beat information could be most conducive to generation.
- The effectiveness of the Transformer encoder is ambiguous.

Limitation

- **Generalizability**: generated accompaniments are worse when using recordings of drumless music outside our joined dataset.
- **Stability**: the model struggles to change its tempo going through different sections of a song.
- **Dependency**: insufficient clues for locating beats and tempo would lead to bad accompaniment generation.

Demo page:

code: