151st AES Convention • October 2021

WaveBeat: End-to-end beat and downbeat tracking in the time domain

Christian J. Steinmetz
c.j.steinmetz@qmul.ac.uk

Joshua D. Reiss

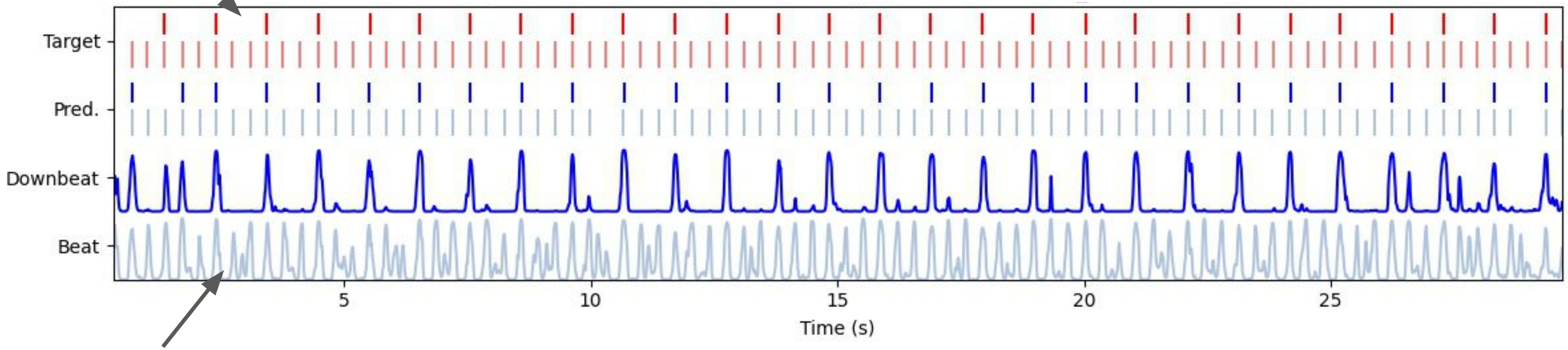
Centre for Digital Music, Queen Mary University of London

Queen Mary University of London

AI + MUSIC

Beat and downbeat tracking

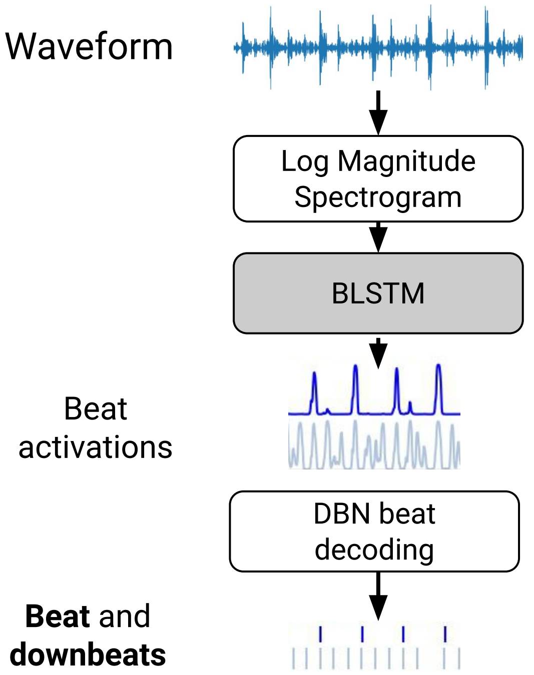
Human beat annotations



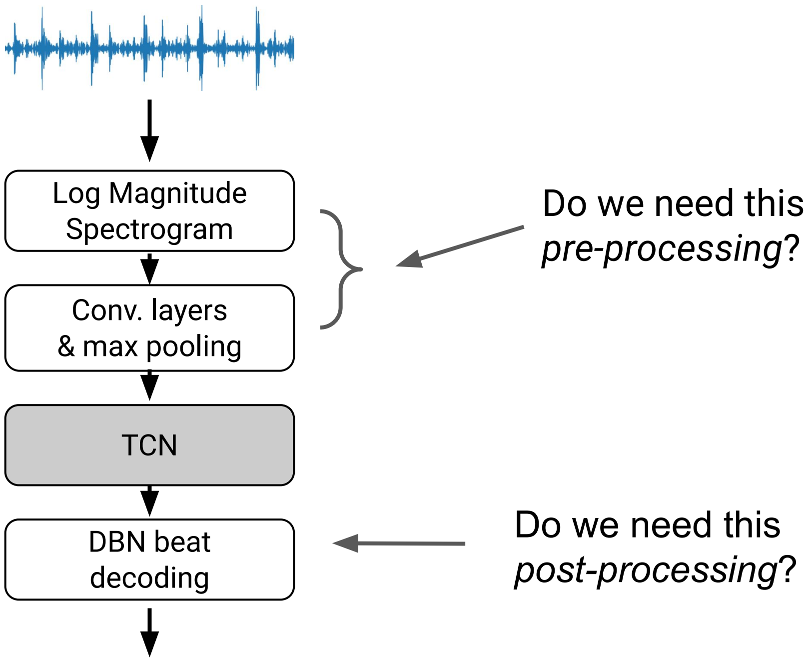
Model estimates beat likelihood

Estimating a sequence of time instants that reflect how a human listener may tap along with a musical piece.

Deep learning-based beat tracking systems

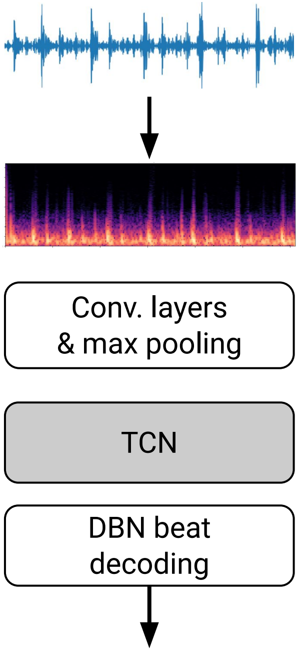


Böck, Sebastian, Florian Krebs, and Gerhard Widmer. "Joint Beat and Downbeat Tracking with Recurrent Neural Networks." ISMIR. 2016.



Matthew E. P. Davies and Sebastian Böck. "Temporal convolutional networks for musical audio beat tracking." EUSIPCO. 2019.

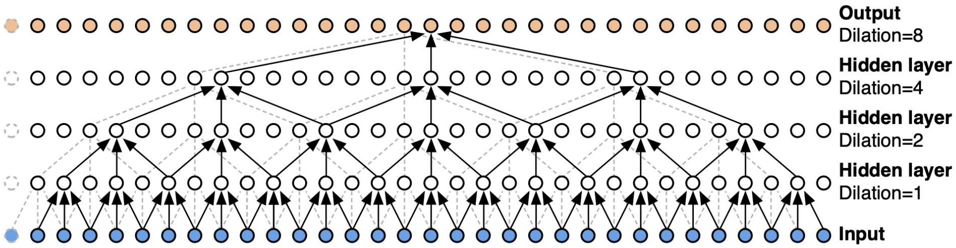
Spectral vs. Time domain



- Considers only spectral magnitude (and ignores phase)
- Subsampling (pooling) spectral frames discards temporal information

*Time domain approaches (phase information) was used in traditional beat tracking systems, so why not have an **end-to-end time domain beat tracking** model?*

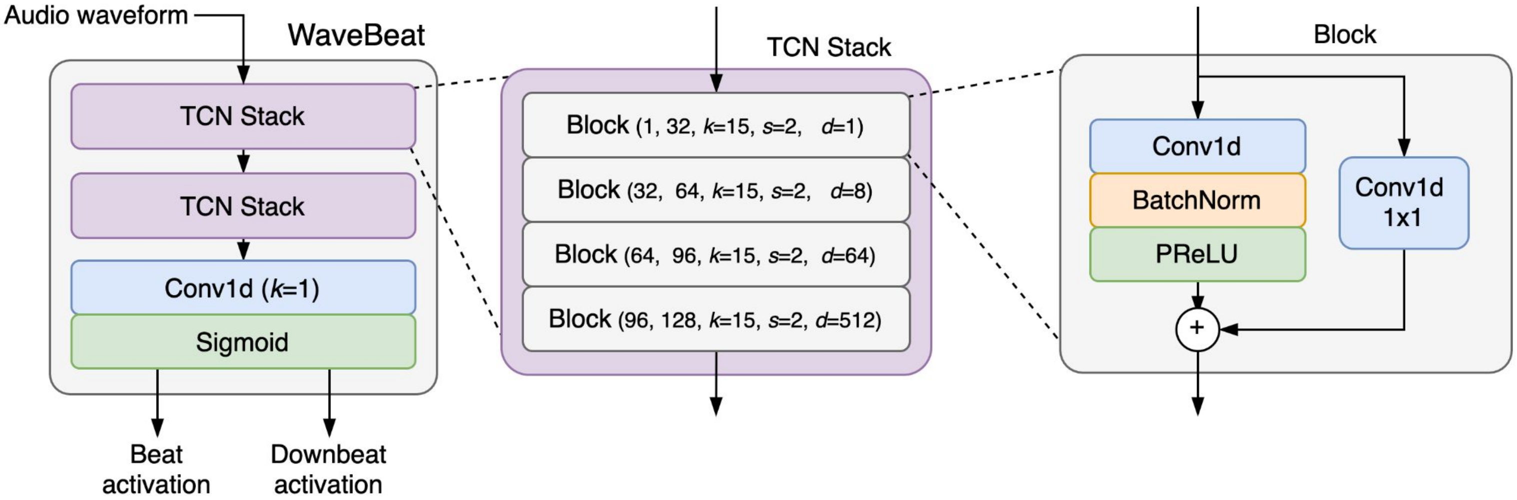
Time domain architecture for beat (downbeat) tracking



- Often requires receptive field of over 30 seconds
- Equates to over **1/2 million time steps** at 22.05 kHz
- Most time domain models have a receptive field of around a few seconds at most

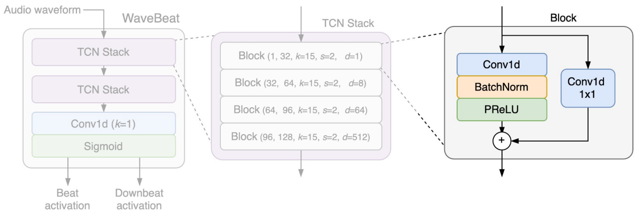
*We use a TCN (feedforward WaveNet) with **rapidly growing dilation factors** for a large receptive field.*

WaveBeat Architecture

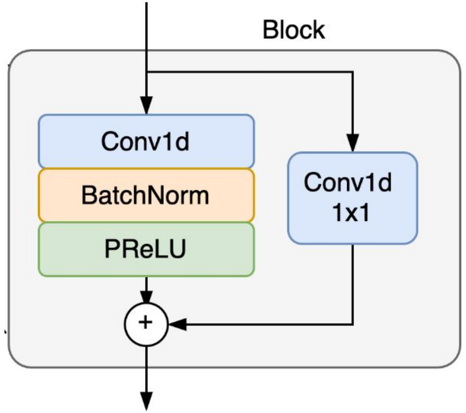


Architecture

Convolutional block

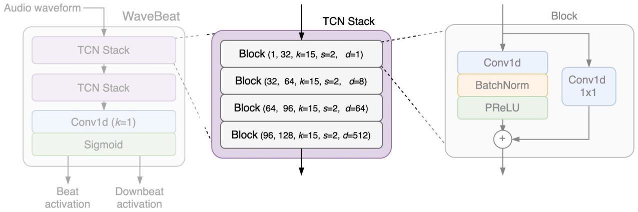


- Residual 1D convolutional layers (kernel=15)
- Use stride=2 to downsample in time
- Batch normalization for stability
- PReLU (learnable activation function)

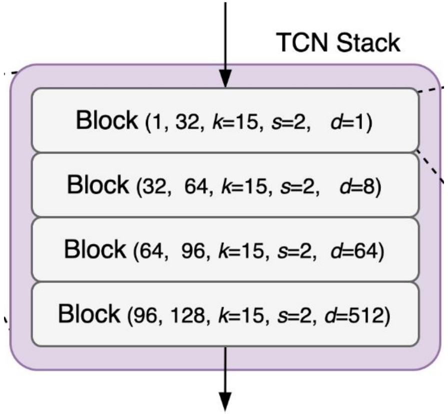


Architecture

TCN Stack

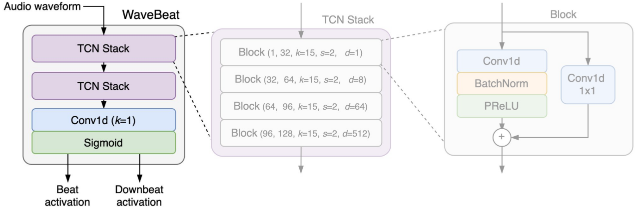


- Each TCN stack is composed of 4 blocks
- Convolutional channels increases by 32 at each block (more parameters, deeper)
- Dilation increases by factor of 8^n (1, 8, 64, 512)
- This pattern repeats at each stack

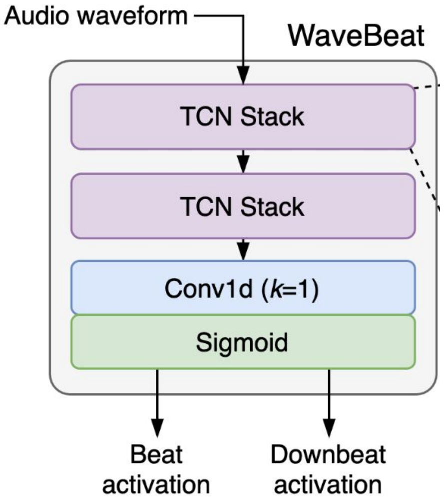


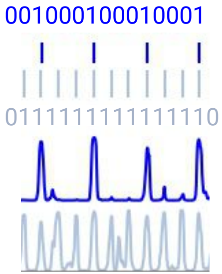
Architecture

Complete WaveBeat



- Complete architecture is composed of 2 stacks (total of 8 layers)
- Final 1x1 convolution downmixes channels to 2 outputs (beat and downbeat activations)
- Sigmoid forces outputs between 0 and 1





Use BCE loss where target is 1 at beat location and 0 elsewhere.

Training & Experiments

- **WaveBeat** with 8 layers and 2.9 M parameters
- Receptive field of over **1 million time steps** (~47 sec)
- Train for 100 epochs (1 epoch = 1000 segments 1.6 min long)
- Use Adam optimizer and learning rate 1e-3
- Decrease learning rate on plateau after 10 epochs (F-Measure)
- Model operates on waveforms with 22.05 kHz sample rate

Training datasets

- Beatles
- Hainsworth
- Ballroom
- RWC Popular

Held-out test sets

- GTZAN
- SMC

Data augmentation

- additive white noise (*p* = 0.05)
- tanh nonlinearity (*p* = 0.2)
- random phase inversion (*p* = 0.5)
- highpass and lowpass filters with random cutoff frequencies (*p* = 0.25)
- random pitch shifting between -8 and 8 semitones (*p* = 0.5)
- **shifting the beat locations by a random amount between ± 70 ms (*p* = 0.3)**
- dropping block of audio and beats no more than 10% of the input (*p* = 0.05)

Time domain beat tracking is competitive

Dataset	Size	Model	Beat			Downbeat		
			F-measure	CMLt	AMLt	F-measure	CMLt	AMLt
Ballroom	5 h 57 m	Spectral TCN [15]	**0.962**	**0.947**	**0.961**	0.916	0.913	**0.960**
		WaveBeat (Peak)	0.961	0.929	0.929	0.904	0.762	0.803
		WaveBeat (DBN)	0.925	0.829	0.937	**0.953**	**0.916**	0.941
Hainsworth	3 h 19 m	Spectral TCN [15]	0.902	0.848	0.930	0.722	0.696	0.872
		WaveBeat (Peak)	0.965	0.937	0.937	0.912	0.748	0.843
		WaveBeat (DBN)	**0.973**	**0.976**	**0.976**	**0.954**	**0.886**	**0.970**
Beatles	8 h 09 m	Spectral TCN [15]	-	-	-	**0.837**	**0.742**	**0.862**
		WaveBeat (Peak)	0.887	0.733	0.790	0.689	0.327	0.585
		WaveBeat (DBN)	**0.929**	**0.894**	**0.894**	0.732	0.509	0.724
GTZAN	8 h 20 m	Spectral TCN [15]	**0.885**	**0.813**	**0.931**	**0.672**	**0.640**	**0.832**
		WaveBeat (Peak)	0.825	0.682	0.767	0.563	0.279	0.515
		WaveBeat (DBN)	0.828	0.719	0.860	0.598	0.503	0.764
SMC	2 h 25 m	Spectral TCN [15]	**0.544**	**0.443**	**0.635**	-	-	-
		WaveBeat (Peak)	0.403	0.163	0.255	-	-	-
		WaveBeat (DBN)	0.418	0.280	0.419	-	-	-

- WaveBeat sometimes outperforms Spectral TCN but other times doesn't
- The benefit of the DBN is clear, sometimes it helps, sometimes not.

Improving time domain beat tracking

Time domain models are data hungry

- **More data augmentations**
  - Which augmentations change the audio without hurting performance?
- **Semi/Self-supervised learning**
  - Can we create “noisy” beat annotations from large music corpora
  - Pre-train a large time domain model with this data
  - Then fine-tune the model on our human labeled beat tracking datasts

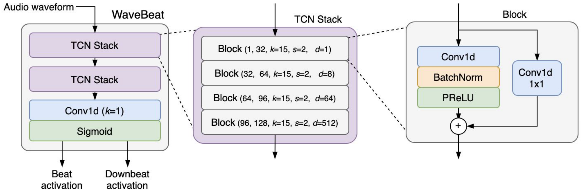
151st AES Convention • October 2021

WaveBeat: End-to-end beat and downbeat tracking in the time domain

WaveBeat

End-to-end beat and downbeat tracking in the time domain.

| Paper | Website | Video |



Setup

First clone the repo.

```
git clone https://github.com/csteinmetz1/wavebeat.git
```

Setup a virtual environment and activate it. This requires that you use Python 3.8.

```
python3 -m venv env/
source env/bin/activate
```

Christian J. Steinmetz
c.j.steinmetz@qmul.ac.uk

Joshua D. Reiss

https://github.com/csteinmetz1/wavebeat