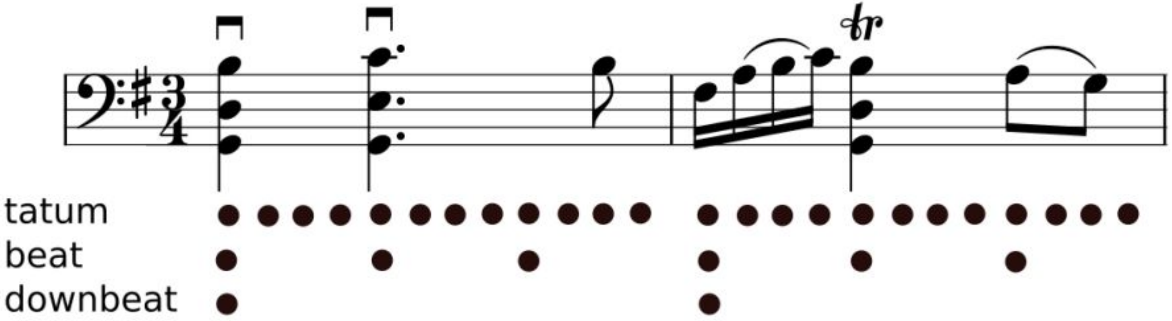
InSpecT

Beat Downbeat Estimation using Transformers

Introduction: Beat Downbeat Estimation

Beat Downbeat estimation is a critical task in Musical Information Retrieval (MIR). The task is to correctly predict the temporal position of beats and detect the downbeats from a given audio sample.

Given its significance, the task has seen a variety of approaches, from classical signal-processing techniques to more advanced deep-learning architectures comprising CNNs, CRNNs, TCN and LSTM, etc. More recently, Transformers have been used in the field, which we dive deeper into with our model.



Example of Beat and Downbeat in a music piece

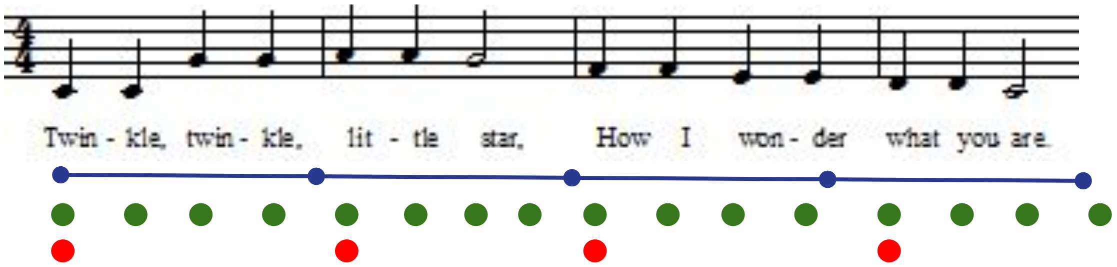
Background: Some Definitions

In music, a **beat** is a basic unit of time. It serves as a reference point for the speed of the music. Intuitively, the beat is the essential rhythm a human might (sometimes incorrectly) tap to when listening to audio.

A **bar** divides a beat into sets of recurring beats (Strong and weak ones). While tapping to a song, one might feel the tap has a rhythm to which it repeats.

A **downbeat** is the first beat of the bar, as simple as it gets.

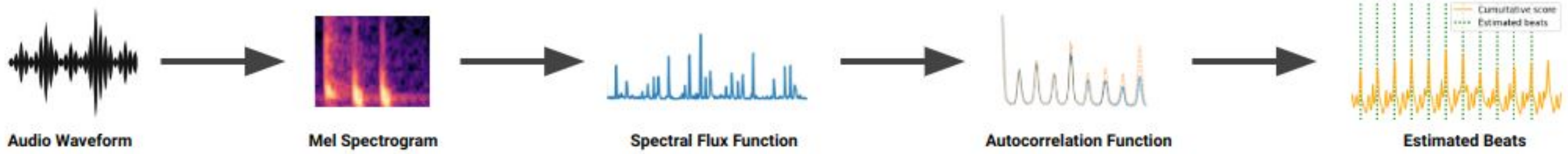
Time Signature of a musical piece is defined by how many beats are in a bar.


Twin - kle, twin - kle, lit - tle star, How I won - der what you are.

Example of **beat (green), bar (blue)** and **downbeats (red)**

Classical Approach: Signal Processing

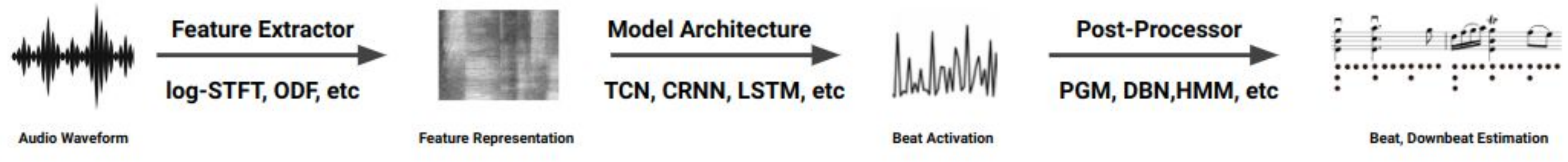
Before going to the state-of-the-art approaches, it is important to look at some classical solutions to this task involving signal processing techniques. A more detailed explanation is given here.



Concisely, the approach first converts the audio waveform to Mel spectrogram which is then converted into a spectral flux function which computes the change in Mel spectrogram over short periods. The spectral function is then passed through an autocorrelation function to estimate the tempo or speed of the audio. Finally, we get the beats tracked using a dynamic programming approach.

Deep Learning Approach: Pipeline

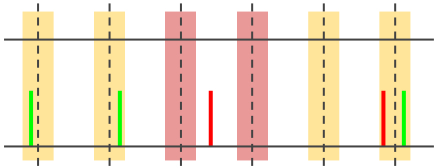
Newer approaches to the estimation task usually follow the following pipeline:



The input audio waveform or spectrogram is first used to extract some features using Short Fourier Time Transform, Onset Detection Function, etc. Then the model is trained over this set of features which outputs an activation function for beats. The beat and downbeat are estimated by post-processing the activation function using Probabilistic Graph Models or Dynamic Bayesian Networks.

Evaluation Metrics

F-measure



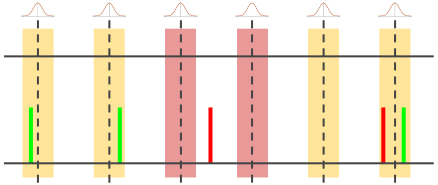
True Positives: Beats that falls in annotated tolerance window.

False Positives: Additional beats in a window or any beats outside window.

False Negatives: Window with no beats detected.

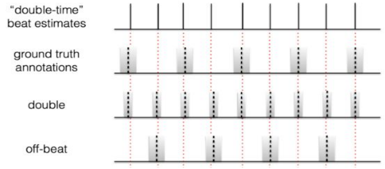
Precision, Recall and F-measure is calculated.

Cemgil



Cemgil is F-measure with a gaussian distribution applied over the tolerance window to reduce the impact of beats detected farther away from the center (annotated beat).

Continuity-Based



CMLx: Longest (x = c) or Total (x = t) contiguous segment/s of correct detections.

AMLx: Same as CMLx but with alternate metrical levels allowed.

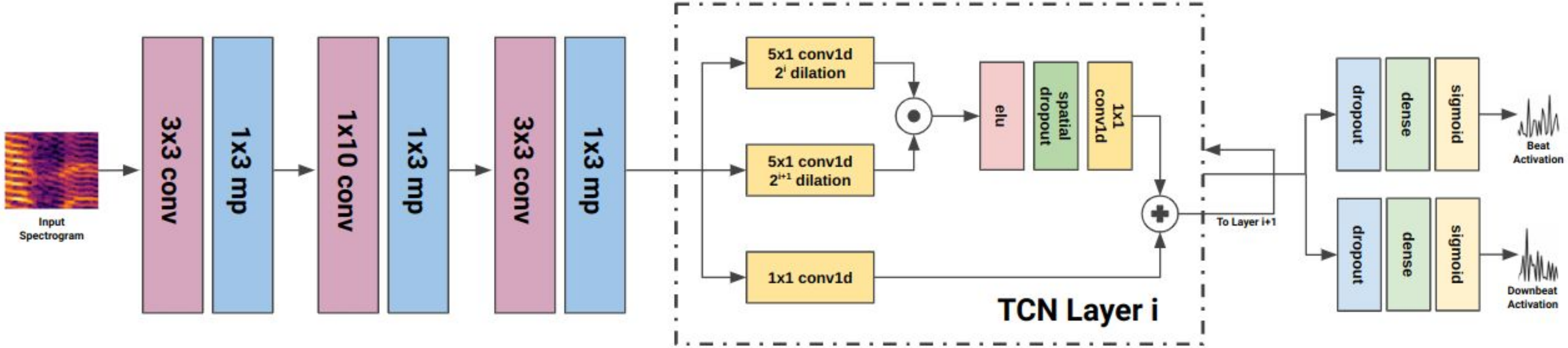
Dataset: GTZAN

The **gtzan8** audio dataset contains 1000 tracks of 30 second length. Containing 100 tracks each of 10 genres, all in **22050Hz** Mono **16-bit** audio files in **.wav** format. The genres are:

- Blues
- Classical
- Country
- Disco
- Hip-hop
- Jazz
- Metal
- Pop
- Reggae
- Rock

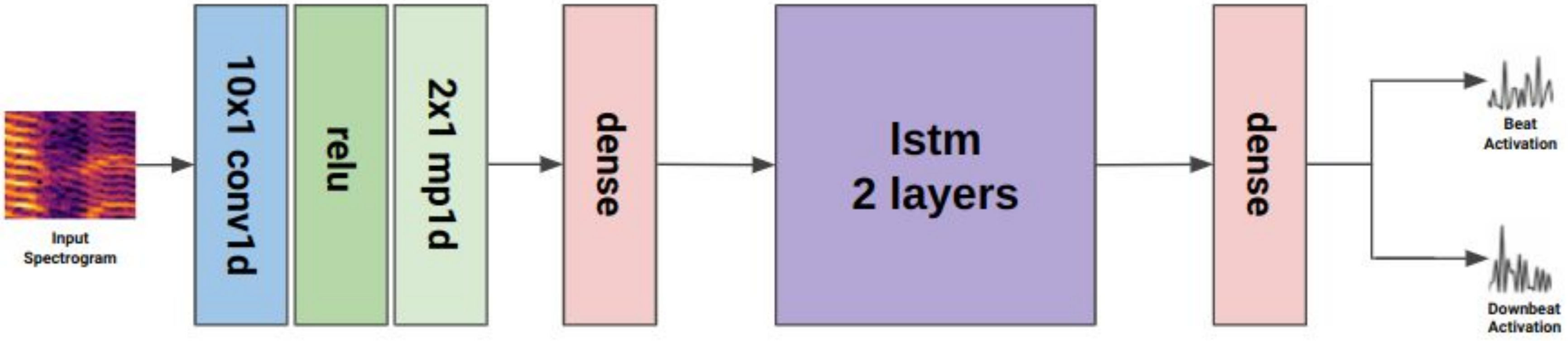
Baseline-1: Deconstruct, Analyse, Reconstruct

Paper: Deconstruct, Analyse, Reconstruct: How to improve tempo, beat and downbeat estimation.



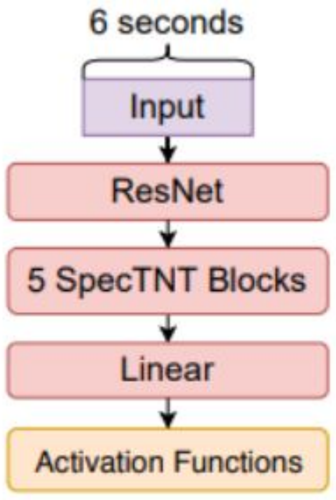
Baseline-2: BeatNet for Offline Inference

Paper: BeatNet: CRNN and Particle Filtering for online joint beat downbeat and meter tracking.

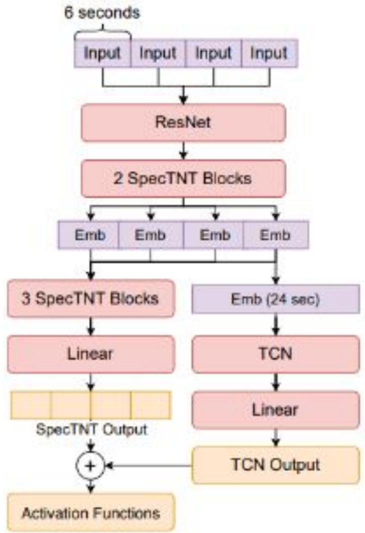


Transformers: Time-Frequency Transformer

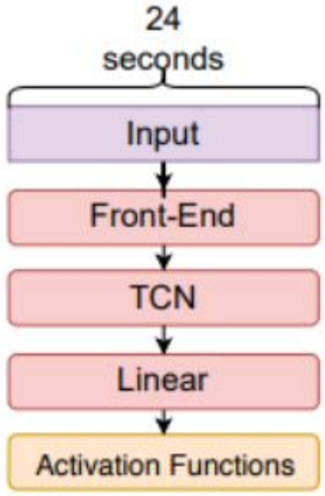
Paper: Modeling beats and downbeats with a time-frequency transformer



SpecTNT



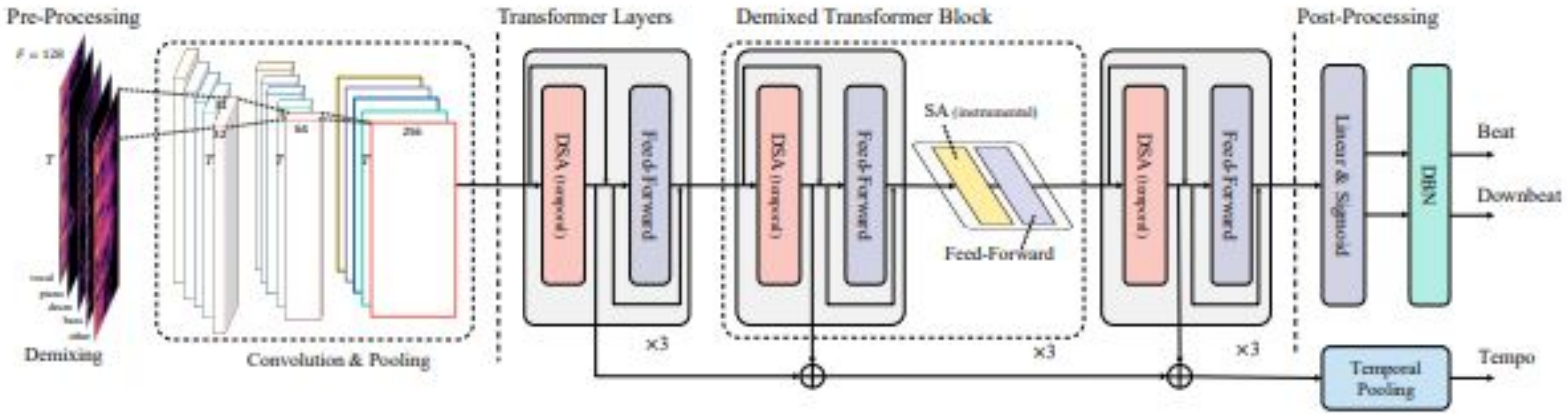
Model Architecture



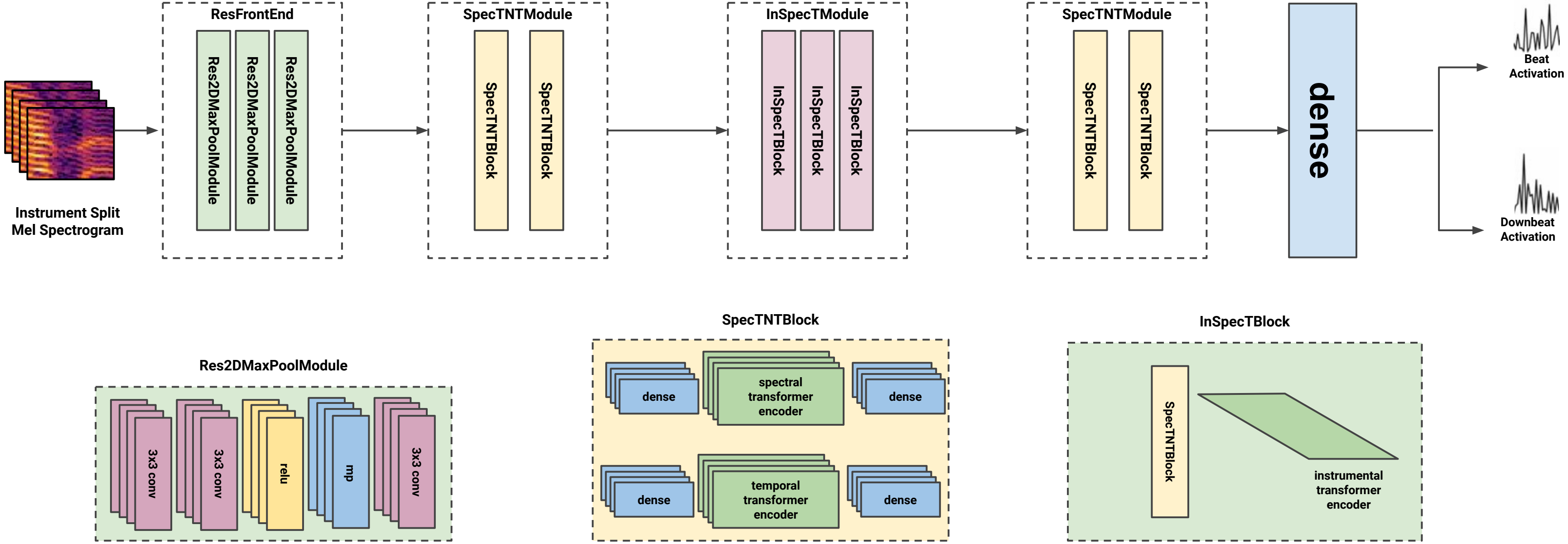
TCN

Transformers: Time-Instrument Transformer

Paper: Beat Transformer: Demixed beat and downbeat tracking with dilated self-attention.

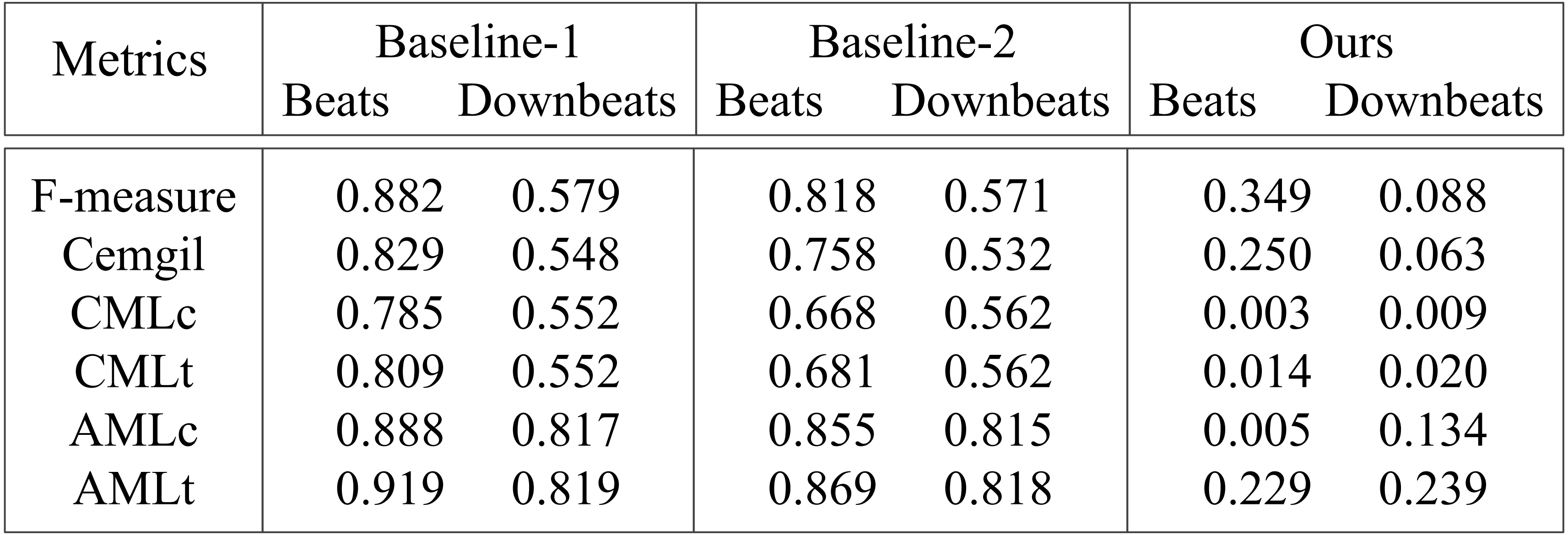


Ours: InSpecT



Proposed Architecture for InSpecT

Results



Metrics	Baseline-1		Baseline-2		Ours	
	Beats	Downbeats	Beats	Downbeats	Beats	Downbeats
F-measure	0.882	0.579	0.818	0.571	0.349	0.088
Cemgil	0.829	0.548	0.758	0.532	0.250	0.063
CMLc	0.785	0.552	0.668	0.562	0.003	0.009
CMLt	0.809	0.552	0.681	0.562	0.014	0.020
AMLc	0.888	0.817	0.855	0.815	0.005	0.134
AMLt	0.919	0.819	0.869	0.818	0.229	0.239

Results of Different Metrics over Architectures

Contributions

- **Aadit -** Implementation of Baseline-2 BeatNet. Formulation and implementation of InSpecT model. Presentation Design.
- **Sahas -** Evaluation Metrics. Report Work. Preprocessing and Postprocessing and Miscellaneous coding tasks. Presentation Material.
- **Niranjan -** Implementation of Baseline-1. Formulation and Implementation of InSpecT model.

References

- https://en.wikipedia.org/wiki/Beat_(music)
- https://en.wikipedia.org/wiki/Bar_(music)
- https://www.music-for-music-teachers.com/twinkle-twinkle.html
- https://tempobeatdownbeat.github.io/tutorial
- https://program.ismir2020.net/static/final_papers/223.pdf
- https://archives.ismir.net/ismir2021/paper/000033.pdf
- https://arxiv.org/pdf/2205.14701.pdf
- https://arxiv.org/pdf/2209.07140.pdf