Modeling Beats And Downbeats With A Time-frequency Transformer

Yun-Ning Hung*, Ju-Chiang Wang, Xuchen Song, Wei-Tsung Lu and Minz Won

Speech, Audio and Music intelligence (SAMI) team, TikTok

We are hiring!

Introduction

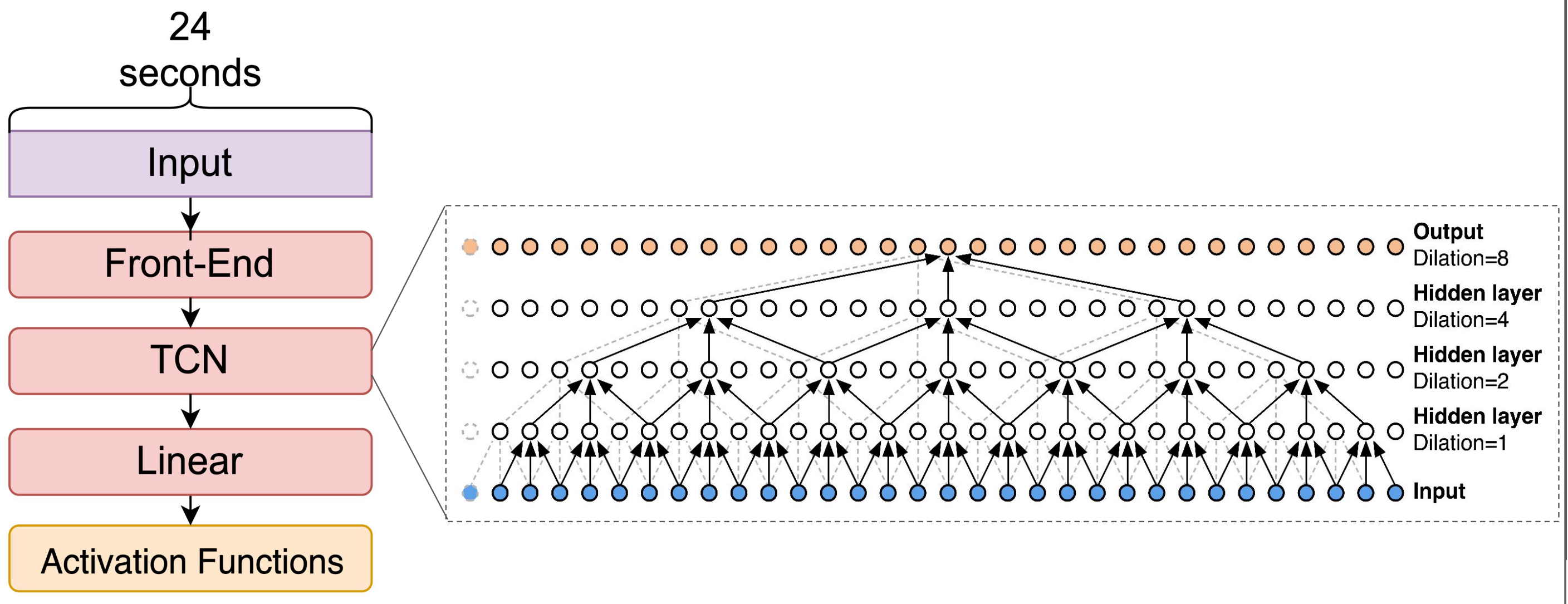
- **Goal**: detect beats and downbeats as pulse signals in music audio
- **Existing problems**:
  - Downbeat tracking performs inferior to beat tracking
    ➡ Lack of harmonic information
    (e.g. downbeat often happens during chord changes)
  - Not many publicly available datasets
  - ➡ Deep learning algorithms are data-hungry, especially the transformer
- **Our solution**: using a time-frequency transformer (SpecTNT [1])
  - SpecTNT is better at capturing both time and harmonic information
  - SpecTNT requires less data than traditional transformer

Methods

Baseline method: temporal convolutional networks (TCN)

👍 : Good at learning sequential/temporal structure

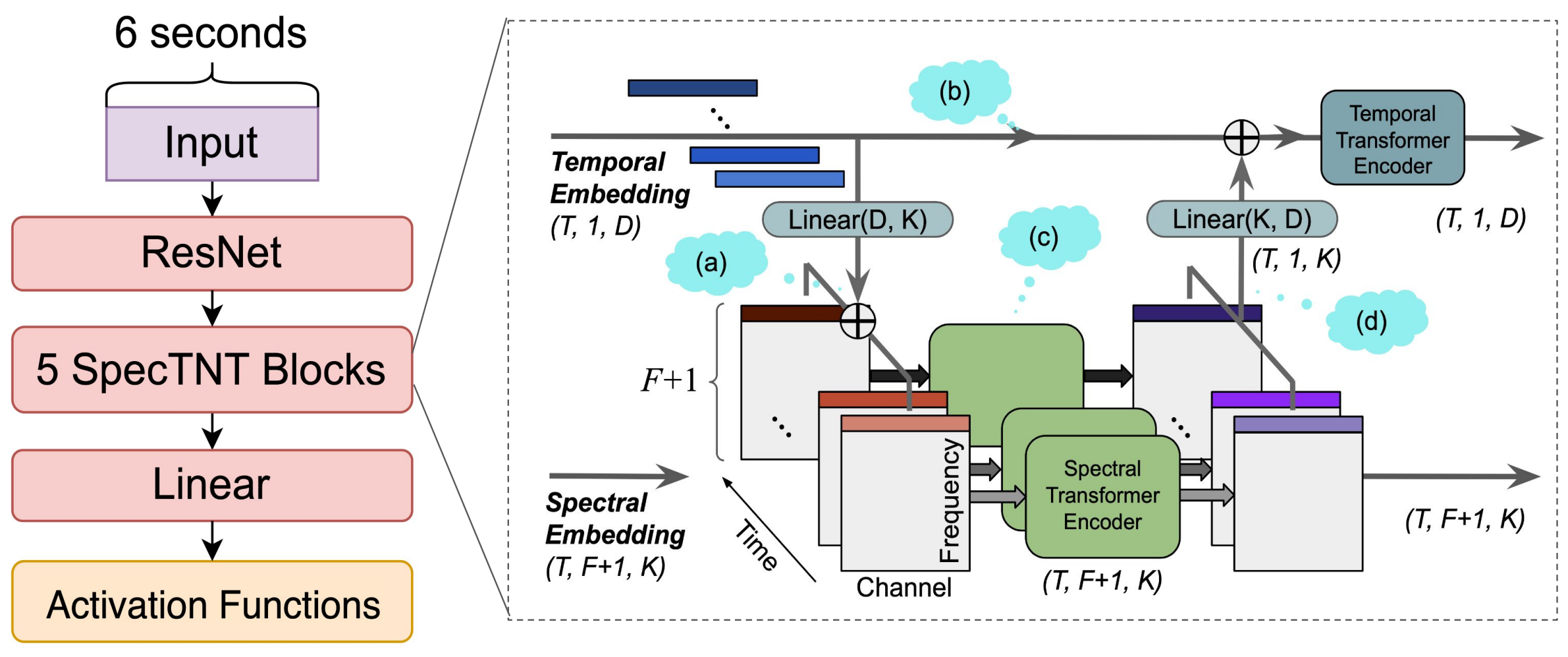
👎 : Potentially lacking harmonic information (for downbeat)



Transformer method: SpecTNT

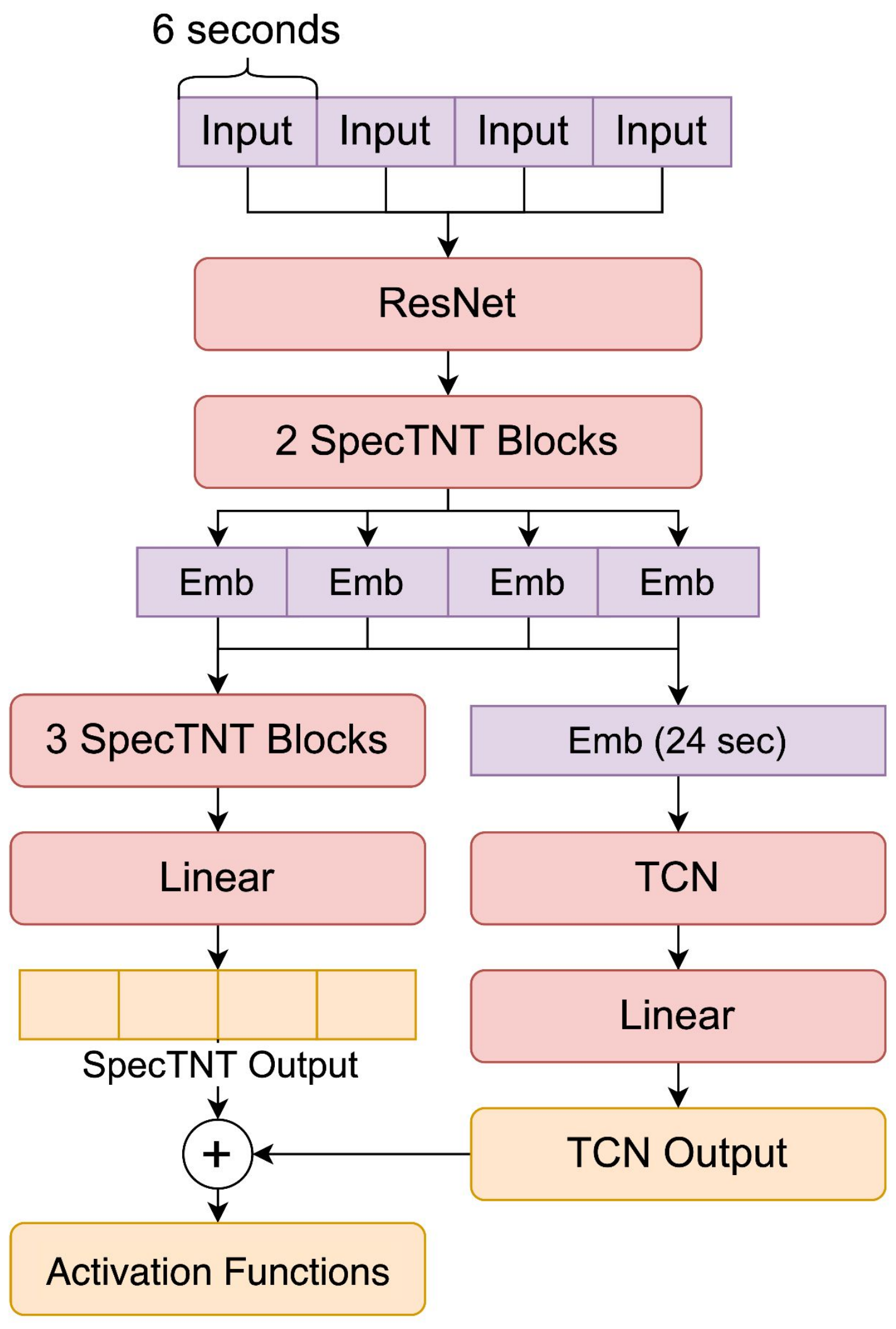
👍 : Better at capturing harmonic information (for downbeat)

👎 : Performance decreases for a longer sequence



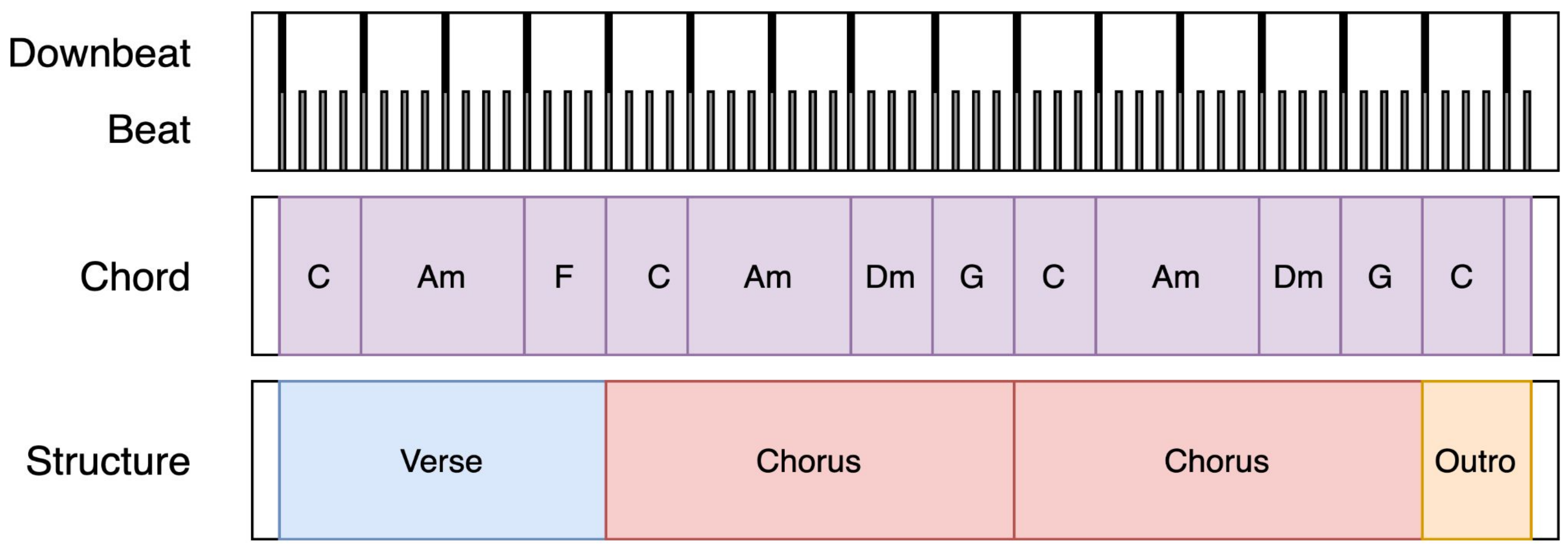
Fusion method: SpecTNT + TCN

- ResNet and 2 SpecTNT blocks to capture high-level information
- 👍: 3 SpecTNT blocks capture harmonic information (left)
- 👍: TCN captures sequential/temporal structure (right)
- Activation functions from each branch are added for the final activations



Future work

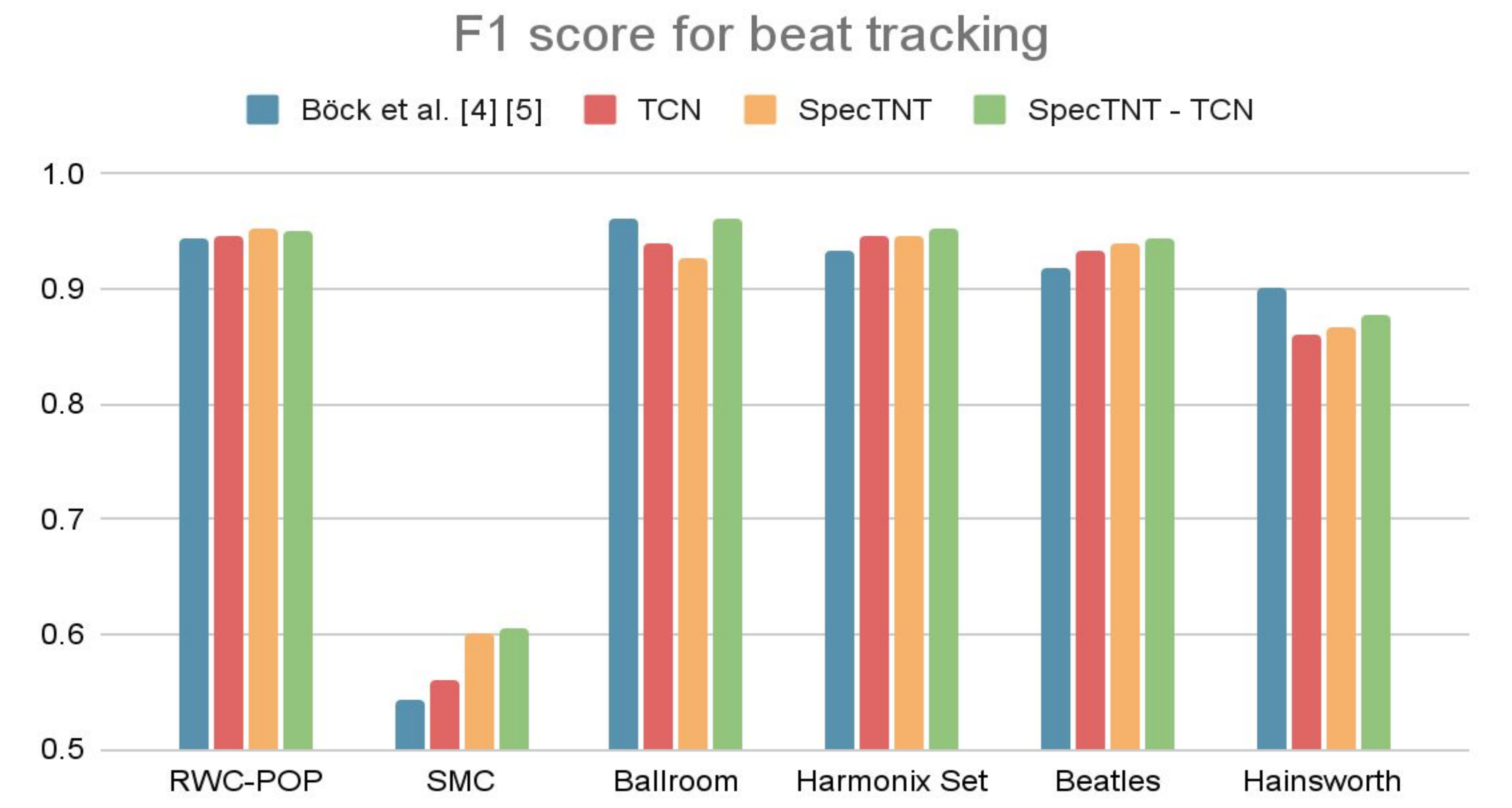
Modeling Hierarchical Structure with Multi-Task Learning



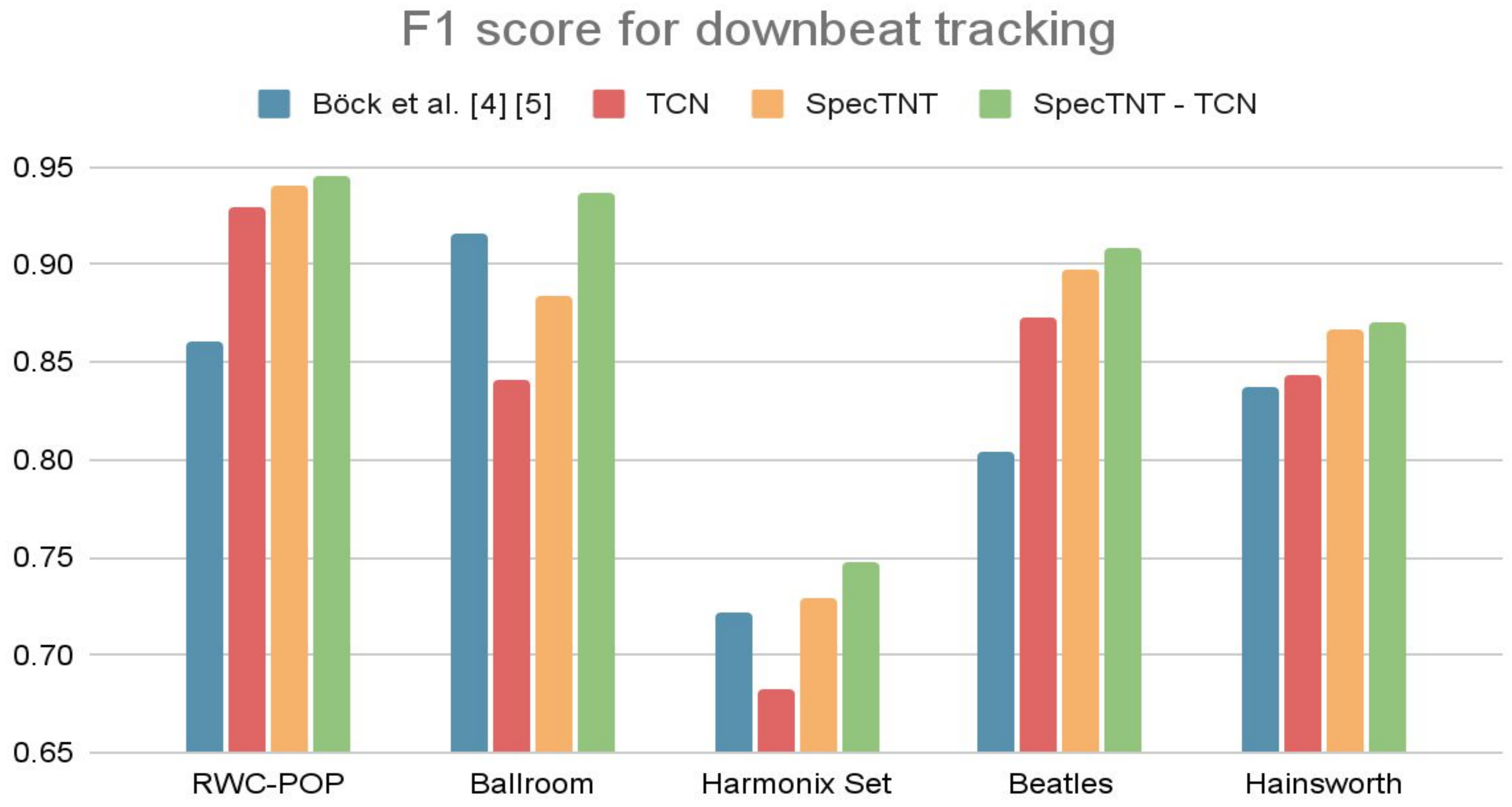
Experiments

- Please see paper for full results!

- **Dataset**: like [5], we use RWC-POP, SMC*, Ballroom, Harmonix Set, Beatles, Hainsworth, Simac*, HJDB*, GTZAN (* dataset with beats only)
- **Data split**: 8-Fold Cross-Validation
- **Evaluation metrics**: F Measure, CMLt, AMLt (see [3])

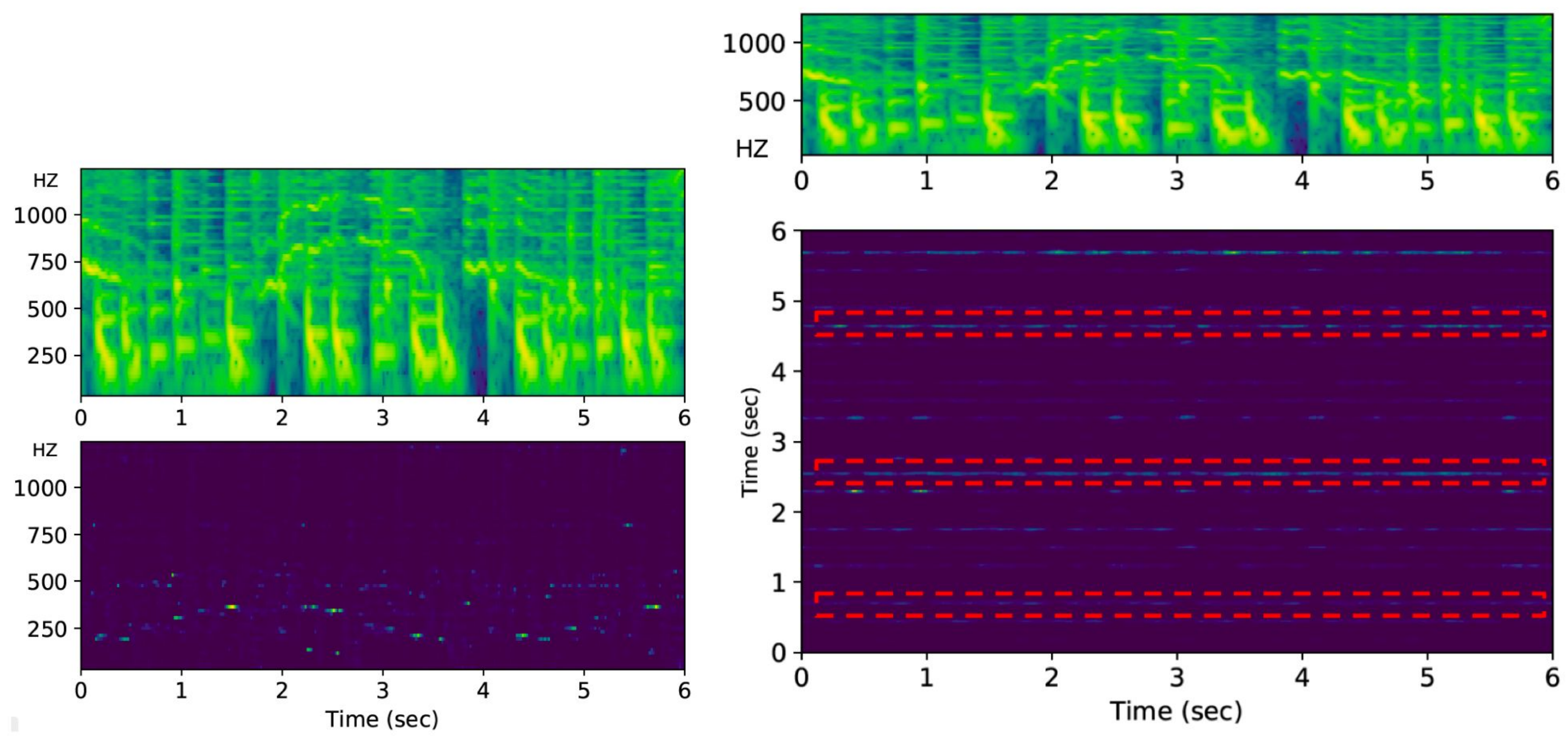


- SpecTNT performs similar to TCN
- SpecTNT - TCN performs similar to SpecTNT on most of the datasets, but especially well on Ballroom
- SpecTNT - TCN performs similar to existing models on most of the datasets, but especially well on SMC



- SpecTNT performs better than TCN
- SpecTNT - TCN performs better than SpecTNT on most of the datasets, but especially well on Ballroom
- SpecTNT + TCN performs better than existing models on most of the datasets

Attention Visualization



(a) Spectral attention (b) Temporal attention

- Spectral attention captures harmonic components (e.g. melody line).
- Temporal attention captures downbeat positions

Reference

[1] Böck et al., "Deconstruct, Analyse, Reconstruct: how to improve tempo, beat, and downbeat estimation", International Society for Music Information Retrieval Conference, 2020.
[2] Lu et al., "SpecTNT: a Time-Frequency Transformer for Music Audio", International Society for Music Information Retrieval Conference, 2021.
[3] Davies et al., "Evaluation methods for musical audio beat tracking algorithms," Centre for Digital Music, Queen Mary University of London, Tech. Rep. C4DM-TR-09-06, 2009.
[4] Böck et al., "Multi-task learning of tempo and beat: Learning one to improve the other.," in Proc. ISMIR, 2019.
[5] Böck et al., "Deconstruct,analyse,reconstruct: How to improve tempo, beat, and downbeat estimation.," in Proc. ISMIR, 2020.