

MIREX 2026 SUBMISSION: CONTEXT-AWARE SEMI-MARKOV POST-PROCESSOR FOR BEAT AND DOWNBEAT ESTIMATION

Zhanhong He¹

Hanyu Meng²

Jiahao Zhao³

Ganghui Ru⁴

Yaolong Ju⁵

¹University of Western Australia

²University of New South Wales

³Kyoto University

⁴Fudan University

⁵Great Bay University

zhanh.he.uwa@gmail.com; juyao-long@gbu.edu.cn

ABSTRACT

Our MIREX 2026 submission is a system to estimate beats and downbeats with either of three neural backbones and four event-decoding settings. We compare the Direct peak picking, the Context-Aware Semi-Markov (CASM) post-processor, and dynamic Bayesian networks (DBNs) with two tempo ranges. Unlike DBN which heavily depends on predefined temporal constraints, CASM adapts temporal constraints based on the local evidence (beat activations). All models and configurations are developed without SMC or GTZAN. Our preliminary experiments show improved continuity while preserving beat F1 across backbones.

1. INTRODUCTION

Neural beat trackers estimate framewise beat and downbeat activations, which requires post-processing into discrete events. Direct peak picking follows local evidence but can retain spurious peaks or miss weak beats. Dynamic Bayesian networks (DBNs) impose temporal regularity through predefined tempo, metre and transition constraints, making their behaviour sensitive to these settings [1].

We apply the Context-Aware Semi-Markov (CASM) post-processor [2], which conditions temporal regularity on local periodic evidence. CASM accounts for ambiguity between competing periods, including half- and double-tempo alternatives, strengthening regularity when one interpretation is clear and relaxing it otherwise. We describe its offline integration with three backbones and the MIREX-specific development protocol.

2. METHOD

2.1 Task Formulation

Figure 1 shows the shared pipeline. Each invocation takes one WAV file and an output path. A backbone produces beat and downbeat logits at 50 Hz; the selected decoding setting returns event times. Separate beat/downbeat task flags write the requested events to an ASCII file, one timestamp in seconds per line.

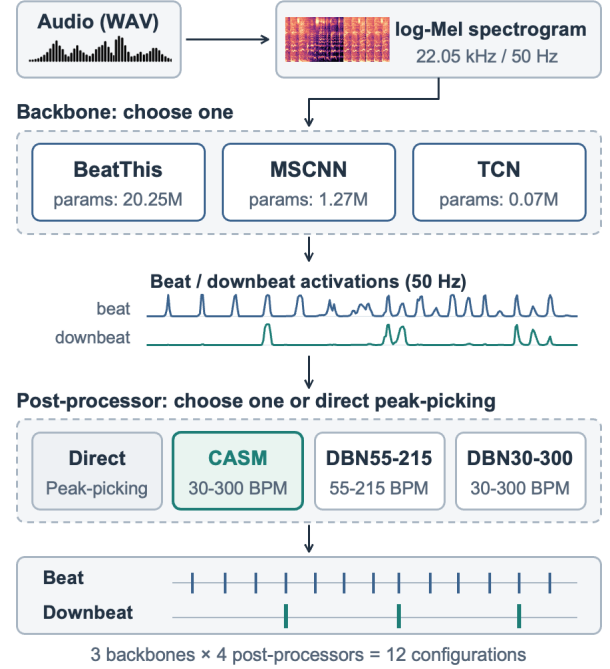


Figure 1. Shared pipeline with three backbones and four decoding settings considered in this report.

2.2 Backbone Configurations

Backbone beat trackers. We submit the BeatThis [3], MSCNN [4], and a TCN re-implementation derived from Böck et al. [5]. Minor modifications are applied to MSCNN and TCN: we remove all multitask heads except for the beat and downbeat estimation heads, and adapt them to accept a unified shared input (the BeatThis spectrograms). All backbones are retrained using the same pipeline (Sec. 2.4) and frozen prior to the post-processing.

Input log-Mel features. To match the BeatThis spectrogram format [3, 6], all input audio is downmixed to mono, resampled to 22.05 kHz, and segmented into 30 s excerpts. The frontend uses a 1,024-sample window, a 441-sample hop size (yielding 50 frames/s), and 128 Mel bands spanning 30–11,000 Hz, followed by $\ln(1 + 1000x)$ magnitude compression.

2.3 Post-Processing

Direct method is the minimal peak-picking strategy introduced in BeatThis [3], which extracts local maxima peak-



activations, and then snapping the downbeat peaks to the beat grid. No tempo constraint is imposed.

CASM is our context-aware semi-Markov decoder introduced in [2]. It estimates a preferred local period and its ambiguity from the activation context, enforcing stronger temporal regularity when the periodic evidence is clear and relying more on the neural activations when it is ambiguous. We use one frozen 30–300 BPM configuration across all backbones and recordings, with CASM serving as the default in our submitted system.

DBN is the joint beat/downbeat post-processor [1] provided in *madmom*. It imposes temporal coherence through predefined priors, including tempo support, meter states, transition dynamics, and observation weighting. Our ICASSP analysis shows that it is more sensitive than CASM to these configuration choices [2]. We submit two otherwise identical settings with tempo ranges of 55–215 BPM and 30–300 BPM (suggested by [7] because of SMC data specific), with other setting default in *madmom*.

2.4 Dataset and Model Selection

Data splits. We use the BeatThis-format 30 s spectrograms released in [6]. GTZAN and SMC are strictly held-out as test sets. The remaining datasets (ASAP, Ballroom, Beatles, Candombe, Filosax, Groove MIDI, GuitarSet, Hainsworth, Harmonix, HJDB, JAAH, RWC, Simac, and TapCorrect) are split into 3,783 training and 556 validation tracks.

Checkpoint selection. To select the optimal checkpoints, we checked the direct peak-picking performance of each backbone on the validation set (over a total of 150 training epochs), choosing the ones that maximized the composite score S :

$$S = 0.6Q_{\text{beat}} + 0.4Q_{\text{downbeat}},$$

$$Q = 0.5F_1 + 0.25\text{CMLt} + 0.25\text{AMLt}.$$

The selected checkpoints correspond to epoch 119 for BeatThis, epoch 149 for MSCNN, and epoch 119 for TCN.

3. PRELIMINARY EXPERIMENT

Table 1 reports held-out diagnostics on 993 GTZAN and 217 beat-only SMC pieces (note that these are not official MIREX results). We report the **F1** (± 70 ms), **CMLt**, and **AMLt** metrics using *mir_eval* [8]. Compared to the Direct baseline, CASM improves beat F1 and all continuity scores, with only a marginal 0.3-point decrease in downbeat F1 for BeatThis on GTZAN. Conversely, while DBN often yields higher continuity, it reduces beat F1, particularly on SMC. Furthermore, the DBN’s performance is dramatically affected by its tempo range setting, highlighting a severe tuning burden (a vulnerability we extensively analyze in our ICASSP submission [2]).

Notably, the Table. 1 results are slightly diversified to our ICASSP submitted paper [2], because of the training data and evaluation protocol changed. In BeatThis benchmark, the 8-fold cross-validation is used, instead of the full dataset used in train and test.

Method	GTZAN (Beat / Downbeat)						SMC (Beat)		
	F1	CMLt	AMLt	F1	CMLt	AMLt	F1	CMLt	AMLt
BeatThis [3]	89.0	79.6	89.7	78.1	66.5	79.8	57.8	41.7	52.3
w. CASM	+0.0	+0.2	+0.4	−0.3	+4.5	+5.1	+0.2	+1.6	+2.4
w. DBN ^[55,215]	−0.6	+1.3	+2.0	−1.1	+6.0	+8.4	−3.2	+0.5	+6.0
w. DBN ^[30,300]	−0.5	+0.9	+2.6	−0.7	+6.1	+8.8	−3.1	+1.6	+6.1
MSCNN [4]	85.9	72.2	79.0	69.6	50.8	61.6	54.8	34.8	41.6
w. CASM	+0.4	+1.2	+1.5	+2.1	+10.4	+10.8	+0.9	+4.5	+5.4
w. DBN ^[55,215]	−0.4	+4.7	+7.8	+2.5	+18.0	+23.3	−4.4	+4.3	+11.9
w. DBN ^[30,300]	−1.4	+2.6	+6.9	+1.7	+16.5	+22.7	−7.9	−1.9	+8.2
TCN [5]	86.5	72.8	80.5	61.2	21.3	59.3	51.5	27.8	33.4
w. CASM	+0.3	+1.0	+1.2	+4.7	+30.5	+11.7	+0.3	+2.3	+4.5
w. DBN ^[55,215]	−2.0	+0.4	+6.0	+5.4	+38.5	+22.9	−9.5	−5.6	+6.3
w. DBN ^[30,300]	−2.7	−1.1	+5.8	+4.6	+37.7	+23.1	−14.8	−12.0	+3.3

Table 1. Fixed-activation comparison. Backbone rows use Direct; other rows report percentage-point changes. Bold marks the best of the four settings within each backbone. CASM uses 30–300 BPM. DBN ranges are indicated.

4. CONCLUSION

Our submission combines three neural backbones with Direct, CASM, and DBN post-processing variants for beat and downbeat estimation. BeatThis with CASM is our main system and default in our submitted packages.

5. REFERENCES

- [1] F. Krebs, S. Böck, and G. Widmer, “An efficient state-space model for joint tempo and meter tracking,” in *Proc. ISMIR*, 2015, pp. 72–78.
- [2] Z. He, H. Meng, and Y. Ju, “CASM: Context-Aware Semi-Markov post-processor for beat tracking,” 2026, submitted to ICASSP 2027. [Online]. Available at <http://arxiv.org/abs/2610.09318>.
- [3] F. Foscarin, J. Schlüter, and G. Widmer, “Beat this! accurate beat tracking without DBN postprocessing,” in *Proc. ISMIR*, 2024, pp. 962–969.
- [4] Z. He, H. Meng, D. Huang, and R. Togneri, “Joint estimation of piano dynamics and metrical structure with a multi-task multi-scale network,” in *Proc. ICASSP*, 2026.
- [5] S. Böck and M. E. P. Davies, “Deconstruct, analyse, reconstruct: How to improve tempo, beat, and downbeat estimation,” in *Proc. ISMIR*, 2020, pp. 574–582.
- [6] J. Schlüter and F. Foscarin, “Beat This! Spectrograms for Beat and Downbeat Tracking,” Oct. 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.13922116>
- [7] J. Ahn, T. G. Hwang, and M.-R. Jung, “The SMC blind spot: A failure mode analysis of state-of-the-art beat tracking,” 2026, [Online]. Available: <https://arxiv.org/abs/2605.12287>.
- [8] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, “mir_eval: A transparent implementation of common MIR metrics,” in *Proc. ISMIR*, 2014.