

MULTI-SCALE EXPERT REVERIFICATION FOR JOINT BEAT AND DOWNBEAT TRACKING

Tianle Yang

ytyoytyooo@gmail.com

ABSTRACT

We describe a shared system for MIREX Audio Beat Tracking and Audio Downbeat Estimation. Deterministic audio experts retain dense onset, spectral, harmonic, structural and periodicity evidence instead of reducing a recording to its largest peaks. Three neural stages reinterpret this evidence at frame, candidate-sequence and beat-synchronous scales. Multiple time-varying tempo and bar-phase hypotheses remain available until constrained joint decoding; the system does not assume one constant tempo or restrict musical bars to three fixed time signatures. The local candidate retains a 4.46-million-parameter joint checkpoint trained on GuitarSet. Later bounded fine-tuning on four datasets did not improve the saved aggregate validation objective, so its weights were not selected. We disclose both the retained lineage and the later development data. Results on a small, repeatedly used validation cohort indicate weak cross-domain performance, especially for downbeats. Windows engineering acceptance verifies actual model loading and timestamp behavior, but is not an accuracy benchmark or evidence of Linux compatibility. The submission owner has selected Windows-only acceptance and omitted Linux validation.

1. TASK AND INPUT CONTRACT

The MIREX tasks request all beat times or only the first beat of each bar, respectively [1, 2]. We provide two command-line entries sharing one model. Each accepts a single 44.1 kHz mono PCM16 WAV and an explicit output filename. Results are ASCII, one time in seconds per line. Outputs must be finite, nonnegative, strictly increasing and bounded by the original source duration. Invalid inputs fail without replacing an existing output.

Exact-zero sources and one-source-sample inputs produce empty predictions and unknown rhythmic states. This is a narrow input-degeneracy rule, not a loudness threshold or a minimum music duration. Other valid inputs traverse the full architecture. A missing or inconsistent declared model remains an error even when the audio would otherwise yield an empty result.

© 2026 T. Yang. This technical report is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Attribution: T. Yang, "Multi-scale Expert Reverification for Joint Beat and Downbeat Tracking," MIREX 2026 technical report. This notice does not license the software, model or third-party datasets.

2. MULTI-SCALE REVERIFICATION

2.1 Shared Representation and Expert Evidence

The retained configuration uses 96 Mel bands with three mixture Log-Mel resolutions, harmonic and percussive HPSS Log-Mel maps, and Chroma: 492 raw feature channels in total. Seventeen dense expert curves represent onset and phase changes, energy, frequency-band accents, harmonic novelty and structure. Availability masks and confidence accompany complete curves; local peaks are additional candidates, not a filter that removes all non-peak frames. Autocorrelation, Fourier tempograms, comb filtering and PLP-style evidence form four local-period experts. These are implementation-specific evidence channels, not independent ground-truth votes or claims of exact equivalence to every similarly named published algorithm.

2.2 Three Neural Stages

Stage 1 encodes raw features and expert evidence through separate branches and reliability-aware fusion. Six dilated temporal residual blocks and Transformer processing model local and longer context. The retained network has hidden width 96, four attention heads and three configured Transformer layers. Its frame-level Beat/Downbeat probabilities and local rhythmic-state evidence remain provisional.

Traditional multi-hypothesis decoding combines these outputs with periodicity evidence. Alternative paths preserve metrical levels, phases and local tempo changes instead of committing to one global BPM. Stage 2 resamples the evidence around each path, validates candidates and ranks sequences. Candidate editing supports insertion, deletion, timing offsets and local-period corrections; later stages are not limited to selecting among a permanently pruned peak list.

Stage 3 summarizes the evidence synchronously with candidate beats and reasons across events and bars. Unconditional bar-state prediction is separate from compatibility with a proposed condition. The vocabulary contains Unknown and all phases for bar lengths 2–12, giving 78 joint states. These lengths count tracked beats per bar; they are not a complete vocabulary of notated time signatures. State changes over time are permitted.

2.3 Joint Decoding and Bounded Search

Final decoding selects compatible beat and bar-phase trajectories using neural probabilities, expert evidence and

Scale	Use of the shared evidence
Frame	Dense curves, raw features and periodicity maps predict initial events and local rhythmic states.
Candidate	Windows around alternative paths assess event validity, missing beats, offsets and sequence quality.
Beat / bar	Beat-synchronous summaries assess bar phase, state changes and Beat/Downbeat compatibility.

Table 1. Evidence is reinterpreted at three scales, not copied into three identical votes. All stages remain in the retained candidate.

sequence scores. Downbeats must be a subset of the selected beats. The default local tempo grid spans 25–320 BPM in 0.5-BPM increments. Finite tempo bins, hypothesis budgets and bar-length bounds are computational approximations, not assertions that all music lies inside these ranges. Long-context processing uses overlapping local attention and chunk-memory tokens rather than unbounded full-recording self-attention.

3. TRAINING, DATA AND MODEL SELECTION

Training proceeds through Stage 1, Stage 2, Stage 3 and joint fine-tuning. Supervision covers Beat/Downbeat events, rhythmic states, candidate ranking, edits and consistency where targets are available. Missing labels are masked, not invented. The retained deployment-best checkpoint records completion of all stages and joint selection at epoch 2. It is genuinely selected for the joint task, not relabeled as two task-specific best models. Its recorded training lineage is GuitarSet 1.1.0 [3], with 286 unique training-audio hashes and 48 recording-group identifiers. These are provenance counts, not 286 independent compositions. Original selection used 36 GuitarSet validation tracks. Checkpoint loading validates dataset identity, protocol versions, training lifecycle, configuration and validation metadata.

Subsequent development used a frozen cohort from GuitarSet, Ballroom annotations [4], Candombe [5] and ASAP/MAESTRO [6, 7]. Training/validation counts were respectively 48/8, 96/16, 12/2 and 8/4, totaling 164/30 tracks. Two bounded fine-tuning rounds of two epochs were completed. The later development use is disclosed even though those weights were not retained. It must not be conflated with the selected model’s original training lineage.

Recorded versions include Ballroom annotation revision 1db08914 (full identity in the accompanying disclosure), Candombe Zenodo record 6533068, ASAP v1.1 and MAESTRO v2.0.0. Recorded licenses are CC BY 4.0 for GuitarSet/Candombe and CC BY-NC-SA 4.0 for the cited Ballroom annotation and ASAP/MAESTRO records. Annotation licenses do not substitute for original audio terms, and metadata strings are not a legal determination about

Validation dataset	Tracks	Beat F1	Downbeat F1
ASAP/MAESTRO	4	0.2628	0.0927
Ballroom	16	0.4417	0.2369
Candombe	2	0.2451	0.0496
GuitarSet	8	0.6420	0.5564

Table 2. Retained joint checkpoint on 30 development-validation tracks. Scores are fractions, not percentages; sample sizes are small.

model redistribution. Audio, private paths and feature caches are excluded from the software package; dataset attribution and rights remain separately reviewable.

All GTZAN and SMC splits are excluded from shared-model development, following the stricter combined task restriction. Hidden competition test data and pre-trained checkpoints of unknown training origin are not used. Downloaded collections without reviewed audio and Beat/Downbeat targets are not claimed as training data.

4. PRELIMINARY RESULTS AND LIMITATIONS

Table 2 reports saved dataset-mean F-measures with a 70 ms matching tolerance on the fixed mixed validation cohort. These are not official MIREX results or an untouched held-out test. The repository’s composite joint selection score was 0.25604 for the retained model and 0.21973 for the best new joint candidate; this composite is not the mean of the table’s F1 columns. The original GuitarSet-only score, 0.52097, uses a different validation corpus and is not directly comparable.

Cross-domain accuracy is limited, particularly for downbeat phase. Repeated selection on this small validation cohort prevents an unbiased generalization claim; Candombe has only two validation examples. Bounded tempo/bar search can miss rhythms outside the configured representation. Exact-zero abstention does not establish rejection of quiet noise, tones or other ambiguous non-musical input. No architecture ablation or competition-ranking claim is made.

5. ENGINEERING ACCEPTANCE AND RESOURCES

The network contains 4,460,649 parameters and an 18,220,791-byte checkpoint. The saved Windows RC2 acceptance completed 42 intended outcomes, including 18 deliberate rejections and a 308-test regression suite. Direct output review confirmed empty silence/single-sample outputs and unchanged ordinary outputs relative to the preceding candidate. A later relocated Windows package passed 15 checks, including both entries on six inputs, and independent archive/model review. These are bounded software-acceptance results, not accuracy scores.

Measurements used an AMD Ryzen 7 7735HS, Windows 11, Python 3.13.12, PyTorch 2.13.0+cpu and four CPU threads. On a 22.1472-second GuitarSet validation recording, fresh-process Beat/Downbeat runs took

169 47.399/47.106 seconds, with 605.94/607.38 MiB peak 197
170 working set. Elapsed time includes imports and model 198
171 loading. Across four complete validation songs, each task 199
172 invocation required 1.57–2.18 times audio duration. Both
173 entries independently perform joint inference; these obser- 200
174 vations are neither full-corpus upper bounds nor measure-
175 ments on the organizer’s hardware. 201

176 The submitted execution route is Windows 11 x64, 202
177 CPython 3.13.12 and CPU PyTorch, using the observed 203
178 versions above. Native CMD launchers enforce strict
179 model resolution and four CPU threads before calling the
180 unchanged Python entries. Linux validation is explicitly
181 omitted at the owner’s request. The repository retains an
182 unvalidated Debian 12 / CPython 3.10 container recipe 207
183 for reference; it is not the accepted environment for this
184 Windows-only package. Observed Windows versions are 208
185 not a complete hash-locked clean-install certification. 209

186 A strict sealed bundle binds source, configuration and 210
187 model hashes; inference does not download weights. Miss- 211
188 ing or modified assets fail rather than silently falling back
189 to the development baseline. Windows-only verification 213
190 does not establish compatibility with organizer hardware
191 or confirm organizer acceptance of the platform. The final
192 package records its actual Windows checks separately. 215

193 6. AI USAGE STATEMENT 217

194 OpenAI Codex assisted design discussion, implementa- 218
195 tion, testing, documentation and manuscript preparation 219
196 under author direction. Numerical claims are drawn from 220

saved experiments and audits, not generated benchmark re-
sults. The author is responsible for final verification and
submission.

7. REFERENCES

- [1] MIREX, “2026: Audio Beat Tracking,” 2026. [On-
line]. Available: [https://music-ir.org/mirex/wiki/2026:
Audio_Beat_Tracking](https://music-ir.org/mirex/wiki/2026:Audio_Beat_Tracking).
- [2] MIREX, “2026: Audio Downbeat Estimation,” 2026.
[Online]. Available: https://music-ir.org/mirex/wiki/2026:Audio_Downbeat_Estimation.
- [3] GuitarSet, version 1.1.0, Zenodo, record 3371780.
[Online]. Available: [https://doi.org/10.5281/zenodo.
.3371780](https://doi.org/10.5281/zenodo.3371780).
- [4] CPJKU, “BallroomAnnotations,” revision 1db08914.
[Online]. Available: [https://github.com/CPJKU/Bal
roomAnnotations](https://github.com/CPJKU/BalroomAnnotations).
- [5] Candombe dataset, Zenodo, record 6533068. [Online].
Available: <https://zenodo.org/records/6533068>.
- [6] F. Foscarin et al., “ASAP dataset,” version 1.1. [On-
line]. Available: [https://github.com/fosfrancesco/asap
-dataset](https://github.com/fosfrancesco/asap-dataset).
- [7] Magenta, “The MAESTRO Dataset,” version 2.0.0.
[Online]. Available: [https://magenta.tensorflow.org/
datasets/maestro](https://magenta.tensorflow.org/datasets/maestro).