MULTTIPOP: A MULTITRACK TRANSCRIPTION BENCHMARK FOR POP MUSIC

Nathan Pruyne *Benjamin Stoler* *William Chen*
Chien-yu Huang *Shinji Watanabe* *Chris Donahue*

Carnegie Mellon University

ABSTRACT

We present *MulTTiPop*, a benchmark for evaluating automatic music transcription systems, consisting of pop music segments and their associated multitrack MIDI transcriptions. MulTTiPop contains 572 segments of popular music totaling 3.5 hours of audio, and contains songs from diverse genres and decades from the 1930s to 2000s. To collect this benchmark, we perform 1) Metadata Matching on song segments from the Lakh MIDI and TheoryTab datasets, 2) Beat-based Alignment using audio beat-tracking to warp MIDI timing and 3) Human Anchor Beat Selection to identify alignment points between audio and MIDI. We evaluate state-of-the-art automatic music transcription models on MulTTiPop and find substantial room for improvement, with the best model achieving 42% Onset F1. More details and sound examples of MulTTiPop are available at `https://gclef-cmu.org/multtipop`.

***Index Terms*—** Music transcription, pop music, datasets

1. INTRODUCTION

Automatic music transcription (AMT) systems aim to recognize the symbolic music notes underlying a music recording. In recent years, AMT systems have evolved from transcribing solo piano music [6] to targeting performance on many instruments and musical styles [7]. However, AMT models' performance on real-world commercial pop music has lagged behind other domains. Models like YourMT3+ [8] are only trained on synthesized pop music, and perform poorly on commercially-produced songs. In contrast, Sheet Sage [4] performs well on commercial recordings, but only transcribes melodies and chords rather than full multitrack transcriptions.

Further impeding progress of AMT models on commercial music is the lack of a suitable evaluation benchmark: one that matches commercial audio with ground truth, time-aligned transcriptions. Datasets like MAESTRO [9] and MusicNet [1] provide transcription labels for the narrow domains of solo piano and classical music, respectively. Slakh2100 [2] contains multitrack pop MIDI, but only synthesized audio, which frequently does not closely match commercial recordings in timbre, especially for vocals. TheoryTab [4] contains original, fully-produced pop music audio, but only provides weakly-aligned melody and chord annotations, not full multitrack transcriptions. The closest existing dataset is RWC-Pop [5], which contains pop recordings and multitrack MIDI. However, it has limited genre diversity and uses songs composed specifically for RWC-Pop, rather than real-world pop music for which AMT systems may be deployed downstream.

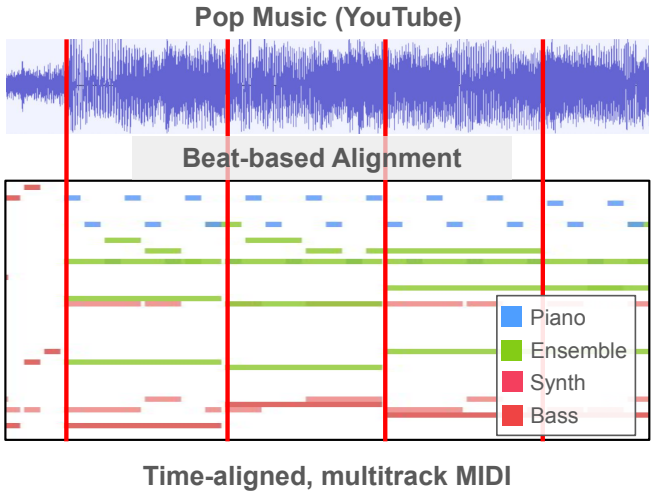

Fig. 1: MulTTiPop contains segments of YouTube audio beat-aligned with multitrack MIDI transcriptions.

To address this gap, **we introduce MulTTiPop, a benchmark of multitrack MIDI transcriptions aligned to segments of commercial pop music** (Fig. 1). We compile multitrack MIDI from the Lakh MIDI Dataset's LMD-matched subset [10], and pair MIDI files with audio segments from YouTube by performing metadata matching with the TheoryTab [4] dataset. TheoryTab is sourced from user-provided audio and chord transcriptions, thus indicating segments of audio that users would be interested in transcribing. MulTTiPop has the following key properties:

- **Diverse, Popular Music:** By sourcing audio from TheoryTab, we provide labels for segments of commercial recordings of popular music from the past decades. This enables MulTTiPop to be a representative sample of use cases for multitrack AMT.
- **Multitrack MIDI:** For each audio segment, we include a time-aligned MIDI track comprised of many instrument parts present in the original recording.

Dataset	Music Genre	Annotation Type	Audio Type	Size (Hours)	Total Samples	MIDI Programs
MusicNet [1]	Classical	**Multitrack MIDI**	**Commercial**	34	330	11
Slakh2100 [2]	**Popular**	**Multitrack MIDI**	Synthesized	145	2100	128
Pop909 [3]	**Popular**	Single track MIDI	None	60	909	1
TheoryTab [4]	**Popular**	Melody and Chords	**Commercial**	50	22000	2
RWC-Pop [5]	**Popular**	**Multitrack MIDI**	Original	7	100	107
MulTTiPop	**Popular**	**Multitrack MIDI**	**Commercial**	3.5	572	123

Table 1: Comparison of existing multi-instrument AMT and pop music datasets, and MulTTiPop. MulTTiPop is the only AMT benchmark containing multitrack time-aligned MIDI annotations for commercial pop music.

- **Commercial Music:** We list links and timestamps for YouTube videos corresponding to all audio segments.

We release MulTTiPop with a CC BY 4.0 license. We provide pairs of aligned multitrack MIDI files and associated metadata files with audio information, including YouTube video IDs and timestamps for the audio segment that corresponds to the MIDI labels. **We do not redistribute the copyrighted music audio**. Instead, we make two key recommendations for the use of MulTTiPop: (1) researchers should only obtain the relevant segments of the original audio, and (2) researchers should only use this dataset for evaluation, not training. We compare MulTTiPop and related AMT and pop music datasets in Table 1.

2. METHODOLOGY

We outline our approach to aligning audio segments sourced from TheoryTab to multitrack MIDI in the Lakh MIDI Dataset (Fig. 2). We first perform Metadata Matching between audio segments and MIDI files. Then, we perform Beat-based Alignment between the MIDI and audio by warping the MIDI beat timings to match detected beats in the audio. Finally, we perform Human Anchor Beat Selection by having annotators select which of a few candidate anchor beats properly align the audio segment to the corresponding portion of the MIDI file.

2.1. Metadata Matching

We gather audio segments for MulTTiPop from TheoryTab [4], which contains audio through YouTube video IDs with start and end times, user-created melody and chord annotations, and metadata information from the HookTheory platform [1]. We begin with 25,947 target segments from TheoryTab, all with valid start and end timestamps, user annotations that are convertible to MIDI, and available audio on YouTube at the time of collection. We source multitrack MIDI from the full LMD-matched segment (31,305 songs) of the Lakh MIDI Dataset [10] [2]. This subset has multitrack MIDI transcriptions and metadata from the Million Song Dataset [11], including song title and artist information. We calculate the Levenshtein distance between the TheoryTab and LMD-matched title and artist metadata using RapidFuzz [12] [3] and filter only to near-identical title and artist fields, yielding 1,164 potential matches between audio segments and multitrack MIDI.

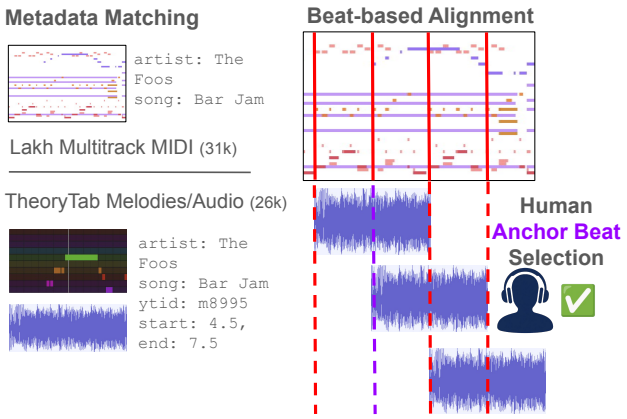

Fig. 2: The methodology of creating MulTTiPop. We metadata match the Lakh MIDI Dataset for multitrack MIDI and TheoryTab for YouTube audio information, align MIDI and audio through beat matching, then use human review to identify an anchor beat in the MIDI matching the audio.

2.2. Beat-based Alignment

While the MIDI transcriptions we gather are defined with strict tempos and beat grids, actual recordings may mismatch in tempo with these transcriptions and may have small changes in tempo throughout. Because of this, we perform fine-grained alignment between the multitrack MIDI and audio via beat matching in a similar style to the preprocessing used in the TheoryTab dataset [4].

[1] https://www.hooktheory.com/theorytab

[2] https://colinraffel.com/projects/lmd/

[3] https://github.com/rapidfuzz/RapidFuzz

Given an audio segment $\mathbf{a}$ with YouTube start and end timestamps from the TheoryTab database $[s, e]$ (in seconds) and its corresponding multitrack MIDI file $\mathbf{m}$, we extract their respective beats $\mathbf{b_a} = [b_a^1, \ldots, b_a^N]$ and $\mathbf{b_m} = [b_m^1, \ldots, b_m^M]$ The multitrack MIDI file contains a full song, while the audio only contains one song segment, so $M \gg N$ in most cases.

The MIDI beats $\mathbf{b_m}$ are defined by the MIDI file's beat grid. To obtain $\mathbf{b_a}$, we perform RNN-based beat tracking [13] on the audio from $[s - 0.5, e + 0.5]$ using the `madmom` [14] package. We include 0.5 seconds of additional context on either side of s and e to mitigate edge issues for the beat tracker, and only preserve beats within $[s, e]$ in $\mathbf{b_a}$. If necessary, we modify the tempo of $\mathbf{b_a}$ by inserting (doubling) or removing (halving) beats, until the tempo is close to that of $\mathbf{b_m}$.

Given $\mathbf{b_m}$ and $\mathbf{b_a}$, we can generate a time alignment by defining an *anchor beat* $k \in 1, 2, \ldots, M - N$ that links the first audio beat (b_a^1) to the k-th beat of the MIDI (b_m^k). We can then define an aligned MIDI transcription by performing linear interpolation on the timing of notes within $\mathbf{m}$ with $|\mathbf{b_a}|$ control points, defined by $[(b_a^1, b_m^k), (b_a^2, b_m^{k+1}), \ldots, (b_a^N, b_m^{k+N-1})]$. We export the aligned MIDI to a new file where each note is warped to the corrected timing determined by the above interpolation, creating time-aligned MIDI labels compatible with standard transcription evaluation.

2.3. Identifying Candidate Anchor Beats

In most cases, only one anchor beat k will yield an alignment that matches the original audio content. In preliminary testing, we find that automatic algorithms for identifying the anchor beat are unreliable. Thus, we select up to 4 candidate anchor beats using the following similarity metrics, then perform manual human annotation on the resulting candidates.

As a baseline metric, we use chroma similarity and onset correlation between the time-warped MIDI and audio. We equally weight the cosine similarity between the MIDI and audio chroma and the Pearson correlation between the onset envelopes. The highest score with this alignment is identified as our `base` candidate anchor beat.

By using melody annotations and YouTube audio data from the TheoryTab dataset, we define 3 additional candidate anchor beats. `melody` is defined by the chroma cosine similarity between TheoryTab's melody annotations and the most similar instrument in the LMD MIDI, weighted at 75%, and `base` weighted at 25%. In `yt`, we find the start time of the YouTube video as a fraction of the video's duration, and only use our baseline metric on beats at the corresponding fraction of the MIDI's duration, ± 10 seconds. For `yt_melody`, we use our melody-detection similarity metric only on beats corresponding with the appropriate time in the YouTube video.

2.4. Human Anchor Beat Selection

To select the correct beat alignment out of the generated options, we present all anchor beat candidates to a set of annotators. Annotators are given the original TheoryTab audio, a piano roll visualization of the MIDI for each candidate starting beat, a synthesized version of the MIDI for each candidate, and an overlay of the synthesized MIDI and original audio. We ask annotators to select an option that transcribes the original audio, or to reject all transcriptions, emphasizing that we expect the synthesized MIDI to be a perfect multitrack transcription. We paid six annotators a rate of $20 USD per hour and verify annotation quality using a hidden control set labeled by an author. Overall, annotators found a matching alignment for 49.1% of the total segments. The success rate of each anchor beat identification method is shown in Table 2.

Method	Selected Alignments	Selection Rate
`base`	25	2.3%
`yt`	58	5.3%
`melody`	346	31.7%
`yt_melody`	437	40.0%

Table 2: Total correct alignments identified by a certain method, with selection rate as a percentage of all generated alignments.

3. THE MULTTIPOP BENCHMARK

The resulting MulTTiPop benchmark contains 572 segments, totaling approximately 3.5 hours of audio. Segments come from 374 unique songs and 263 unique artists. Each segment is 22 seconds on average, and contains an average of 583 MIDI notes. A preview of MulTTiPop including synchronized YouTube video and MIDI playback is available at `https://gclef-cmu.org/multtipop`.

We partition MulTTiPop into a dev split for development and a test split for model evaluation. The dev and test splits are at approximately a 3:7 ratio, and are split so that no artist appears in both dev and test. The dev split contains 169 segments and 61 minutes of audio, while the test split contains 403 segments and 152 minutes. We do not design MulTTiPop for training, so we do not include a train split.

3.1. Benchmark Diversity

MulTTiPop contains a wide range of songs within the overarching pop genre, ranging from rock to New Wave to Motown. MulTTiPop's component songs have release years ranging from 1939 to 2009, with the greatest number of songs being published in the 2000s compared to other decades.

The diversity of MulTTiPop exceeds that of RWC-Pop [5], despite its reduced size in hours. We compare RWC-Pop, the RWC Database [18], and MulTTiPop in Table 4. MulTTiPop contains songs from over twice as many genres and artists as the whole RWC database for modern genres.

Model	Exact Instrument Program			Instrument Category			Harmonic/Percussive		
	Prec	Recall	Onset F1	Prec	Recall	Onset F1	Prec	Recall	Onset F1
MT3 [7]	11.58	10.97	10.87	15.24	14.84	14.51	29.59	28.23	27.82
YourMT3+ [8]	12.88	11.25	11.51	18.17	16.36	16.54	32.17	28.11	28.74
MuScriptor [15]	**20.80**	**16.64**	**17.89**	**30.98**	**25.57**	**27.12**	**49.02**	**38.77**	**41.66**
Gemini Flash 3.8 [16]	6.08	2.68	3.51	6.48	2.92	3.79	8.51	3.79	4.90
Muse Spark 1.3 [17]	4.09	0.90	1.39	4.58	1.00	1.50	7.34	1.34	2.10

Table 3: Evaluation of models on MulTTiPop using exact instrumentation, instrument category, and harmonic-percussive instrument reduction. These results suggest that models have substantial headroom for improvement.

Dataset	Songs	Artists	Genres
RWC-Pop	100	34	2
RWC (All Modern)	250	83	43
MulTTiPop	374	263	101

Table 4: Comparison of RWC and MulTTiPop. RWC (All Modern) contains RWC-Pop, RWC-Jazz, and RWC-Genre.

3.2. Recommended Evaluation Metrics

MulTTiPop is compatible with standard evaluation methodologies for AMT systems that depend on reference transcriptions with fine-grained alignments. We recommend evaluation with Onset F1 using a standard 50ms tolerance [19, 20]. We opt to only consider Onset F1, rather than Onset and Offset F1, since: (i) offsets are not consistently defined in the upstream Lakh MIDI dataset, and (ii) Onset F1 is sufficiently aligned with human perception of transcription quality [21].

MulTTiPop contains diverse instrumentation covering 123 unique instrument labels. Accordingly, we recommend three different evaluation protocols for instrument labels with decreasing strictness. First, “Exact Instrument Program” requires the estimated instrument MIDI program code to be identical to the labeled code. Second, "Instrument Category", accepts any estimated code in the same category as the labeled code, using the "MT3_FULL_PLUS" categories defined in YourMT3+ [8]. Finally, “Harmonic/Percussive” only requires the estimated instruments to correctly distinguish between harmonic (pitched) and percussive (unpitched).

4. EVALUATION

We evaluate both AMT models and leading industry Large Audio Language Models (LALMs) on MulTTiPop (Table 3).

4.1. AMT Models

We evaluate three high-performing open weights general AMT models: MT3 [7], YourMT3+ [8], and MuScriptor [15]. All AMT model transcriptions are available to view on our web preview (see Section 3). We find that YourMT3+, which builds on MT3 primarily architecturally, does not yield a noticeable improvement in quantitative performance. MuScriptor’s expanded dataset does lead to improvements across all metrics, although not to the levels of accuracy (between 80-95% Onset F1 [8]) that AMT models have achieved on piano or classical music. Note that since MuScriptor’s training set is undisclosed, we cannot rule out potential test set contamination with songs present in MulTTiPop.

Qualitatively, while AMT models’ transcriptions generally match the samples’ tonality and chord progressions, they regularly fail to be coherent especially for texturally dense music. We observe disjointed instrument parts and fully missing notes with MT3, and a lack of instrument diversity and rhythmic nuance with YourMT3+. While MuScriptor does give more consistent musical lines and full harmonies, we observe missing instruments and irregular instrumentation at times, as well as rhythmic errors, especially added notes.

4.2. Large Audio Language Models

We prompt two commercial multimodal large language models, Google Gemini Flash 3.8 [16] and Meta Muse Spark 1.3 [17], with MulTTiPop audio and a system prompt to generate MIDI-like transcriptions. We find that these models perform significantly worse quantitatively than AMT models on this task. The output of these models is inconsistent in formatting and quality, often do not transcribe the full segment, and sometimes produce empty transcriptions.

5. CONCLUSION

We present *MulTTiPop*, a benchmark of diverse commercial, popular music and associated multitrack MIDI transcriptions. Through our evaluation of current AMT and LALM systems on MulTTiPop, we find that there is still substantial room for improvement, as evidenced by relatively low Onset F1 scores, and inconsistency in transcriptions’ rhythm and instrumentation. Through MulTTiPop, we provide researchers with greater visibility into AMT systems’ performance on popular music, relevant to many downstream uses of these systems.

6. ACKNOWLEDGEMENTS

This work was supported by funding from Sony AI, and we thank our Sony AI collaborators for helpful conversations. We are also grateful for the contributions of our data annotators at Carnegie Mellon University: Alex Cheng, Woody Li, Arisa Okamura, Sean Xue, Lynn Ye, and Ryan Zhang.

7. REFERENCES

[1] John Thickstun, Zaid Harchaoui, and Sham Kakade, "Learning features of music from scratch," *arXiv preprint arXiv:1611.09827*, 2016.

[2] Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux, "Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 45–49.

[3] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Xianbin Gu, and Gus Xia, "POP909: A pop-song dataset for music arrangement generation," *arXiv preprint arXiv:2008.07142*, 2020.

[4] Chris Donahue, John Thickstun, and Percy Liang, "Melody transcription via generative pre-training," in *ISMIR*, 2022.

[5] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC music database: Popular, classical and jazz music databases.," in *ISMIR*, 2002, vol. 2, pp. 287–288.

[6] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck, "Onsets and frames: Dual-objective piano transcription," *arXiv preprint arXiv:1710.11153*, 2017.

[7] Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel, "MT3: Multi-task multitrack music transcription," *arXiv preprint arXiv:2111.03017*, 2021.

[8] Sungkyun Chang, Emmanouil Benetos, Holger Kirchhoff, and Simon Dixon, "YourMT3+: Multi-instrument music transcription with enhanced transformer architectures and cross-dataset stem augmentation," in *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2024, pp. 1–6.

[9] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019.

[10] Colin Raffel, *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*, Columbia University, 2016.

[11] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere, "The million song dataset," 2011.

[12] Max Bachmann, "rapidfuzz/rapidfuzz: Release 3.13.0," Apr. 2025.

[13] Sebastian Böck and Markus Schedl, "Enhanced beat tracking with context-aware neural networks," in *Proc. Int. Conf. Digital Audio Effects*, 2011, pp. 135–139.

[14] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer, "madmom: a new Python audio and music signal processing library," in *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 10 2016, pp. 1174–1178.

[15] Simon Rouard, Michael Krause, Axel Roebel, Carl-Johann Simon-Gabriel, and Alexandre Défossez, "MuScriptor: An open model for multi-instrument music transcription," *arXiv preprint arXiv:2607.08168*, 2026.

[16] Google DeepMind, "Gemini 3.8 flash model card," Tech. Rep., Google DeepMind, 2026.

[17] Cristina Menghini, Peter Ney, Hamza Kwisaba, Miles Turpin, Felix Binder, Jean-Christophe Testud, Aidan Boyd, Nathaniel Li, Ivan Evtimov, Klaudia Krawiecka, et al., "Muse spark safety & preparedness report," *arXiv preprint arXiv:2606.12429*, 2026.

[18] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC music database: Music genre database and musical instrument sound database," 2003.

[19] Mert Bay, Andreas F Ehmann, and J Stephen Downie, "Evaluation of multiple-f0 estimation and tracking systems.," in *ISMIR*, 2009, pp. 315–320.

[20] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel, "MIR_EVAL: A transparent implementation of common MIR metrics.," in *ISMIR*, 2014, vol. 10, p. 2014.

[21] Adrien Ycart, Lele Liu, Emmanouil Benetos, and Marcus T Pearce, "Investigating the perceptual validity of evaluation metrics for automatic piano music transcription," *Transactions of the international society for music information retrieval*, vol. 3, no. 1, pp. 68–81, 2020.