DOI: 10.71568/daga2026.117

Unison Notes in Multi-Instrument Polyphonic Music Transcription: Challenges for Evaluation

Hans-Ulrich Berendes, Abhirup Saha, Meinard Müller, Ben Maman

International Audio Laboratories Erlangen, Deutschland, E-Mail: hans-ulrich.berendes@audiolabs-erlangen.de

Introduction

Automatic Music Transcription (AMT) has evolved from controlled settings such as piano transcription [1, 5, 6] to more complex scenarios involving polyphonic multi-instrument and even orchestral music [4, 9, 14]. While this development allows for broader applications, it also exposes fundamental limitations in current evaluation methodologies.

Standard note-level AMT evaluation formulates transcription as a note detection task [1, 13]. Detected notes are matched to annotated reference notes within fixed temporal tolerances, and quality is measured using precision, recall, and F1-score. This framework implicitly assumes that each reference note as notated in the score corresponds to a unique and well-defined transcription target, namely a sound event with a clear onset time, pitch, and, possibly, a duration.

In highly polyphonic orchestral music with many simultaneously sounding instruments, this assumption may no longer hold. In particular, *unison notes*, where multiple players play the same pitch at the same notated time, introduce a fundamental ambiguity in both the number and timing of reference notes. In real performances, notes with the same notated onset time are typically not perfectly synchronized but may exhibit small timing deviations, referred to as *onset asynchrony* [3]. Together, unison notes and onset asynchrony challenge the core assumption that each note event in the score corresponds to a unique and unambiguous event in the audio, and vice versa. Figure 1 exemplifies this.

As a consequence, evaluation metrics may reflect arbitrary annotation choices rather than true transcription quality. In this work, we argue that standard AMT evaluation becomes ill-defined in highly polyphonic and orchestral settings, where the relationship between score annotations and their acoustic realization is inherently ambiguous. We support this claim through a combination of conceptual analysis and dataset-based evidence, showing that these ambiguities are not rare edge cases but a systematic property of complex multi-instrument music. While we focus on instrument-agnostic transcription, we note that the problem is not exclusive to this setting.

Challenges in AMT

The challenges outlined above are closely tied to how reference annotations are obtained. AMT has traditionally been studied in controlled scenarios, most notably for piano recordings [1, 5, 6]. This setting offers several advantages: the instrument is fixed, acoustic conditions are relatively stable, and precise annotations can be obtained using specialized hardware, for example through Disklavier-based datasets [7, 12]. Such systems provide accurate MIDI ground truth, enabling reliable training and well-defined evaluation.

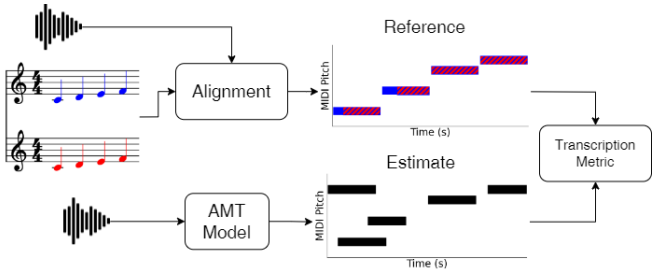

Figure 1: Score-audio alignment enables the transfer of note annotations to the audio domain and often forms the basis for AMT evaluation. When multiple instruments play the same pitch simultaneously, however, the transcription reference becomes ambiguous.

As AMT research has expanded to multi-instrument music, including orchestral works [4, 9, 14], the data annotation methodology has changed. Ensemble recordings involve diverse timbres, complex interactions, and substantially higher levels of polyphony. Moreover, no comparable capture technology exists for orchestras. As a result, reference annotations are typically derived from symbolic scores that are aligned with audio recordings. We can differentiate broadly between two alignment paradigms:

- **Sequence-level alignment:** Computes a global mapping between musical time (score time) and physical time (audio time). All notes that share the same notated musical onset are assigned a single physical onset time. A common method for computing this mapping is Dynamic Time Warping (DTW) [11], which is robust and preserves the overall structure in polyphonic music, but ignores fine-grained temporal variations.
- **Note-level alignment**: Aims at aligning every single note event individually using acoustic cues such as onset activations from transcription models [14]. This allows for resolving small timing differences between notes that are notated as simultaneous in musical time but performed slightly apart in physical time.

Figure 2 illustrates this difference: sequence-level alignment assigns a single physical onset time to all notes at a given score position, whereas note-level alignment can resolve small onset asynchronies between them. While note-level alignment better reflects the physical realiza-

tion of performances, it depends on reliable onset cues and is limited by the temporal resolution of the underlying models. When multitrack recordings are available, note-level alignment becomes considerably easier, as each stem can be aligned individually [10].

Figure 3 further demonstrates how the choice of alignment affects transcription evaluation. Note-level alignment yields higher F1-scores, particularly under stricter onset tolerances, highlighting the importance of capturing fine-grained temporal structure for reliable evaluation.

However, even with note-level alignment, a fundamental ambiguity remains when multiple instruments produce the same pitch at the same notated time. In this case, it is unclear how many note events should be represented. In this work, we focus on the following two closely related phenomena:

- **Onset asynchrony:** Notes that are notated as simultaneous are often performed with small timing differences, ranging from unintentional inaccuracies to deliberate expressive timing variations.
- **Unison note ambiguity:** Multiple instruments, often entire sections such as strings, produce the same pitch at the same musical time, frequently resulting in a fused sound that makes it difficult or impossible to distinguish individual contributions in the audio signal.

These effects are particularly common in orchestral music. As a result, the notion of a correct transcription becomes ambiguous: should such notes be represented as a single event or as multiple events? Different choices lead to different evaluation outcomes, even for identical system outputs. Consequently, standard AMT evaluation metrics [13] reflect not only absolute transcription performance, but also the underlying assumptions and choices made in resolving ambiguities.

While we focus on instrument-agnostic transcription, where unison ambiguity is most common, the problem is not exclusive to this setting. In instrument-aware transcription, notes played by two different instruments yield, in principle, two distinct transcription targets, distinguished by their instrument label. However, ambiguity still arises in two common situations: when two voices of the same instrument play the same pitch simultaneously, or when a single voice is performed by multiple players, as is common in the orchestra's string sections.

Dataset-Based Analysis

To assess how often unison notes and the resulting ambiguities arise in practice, we analyze several widely used datasets covering different levels of polyphonic complexity. These include MusicNet [15], which contains recordings of solo instruments and small ensembles with aligned scores and moderate polyphony; URMP [8], which provides multitrack recordings with separately recorded instruments, enabling precise note-level annotations; PHENICX [10], which offers multitrack recordings of orchestral music with corresponding note annotations; and the Beethoven

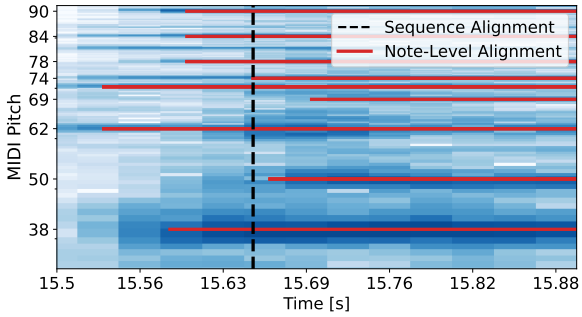

Figure 2: Sequence-level versus note-level alignment for a chord example. All notes share the same musical onset in the score. Sequence-level alignment assigns a single physical onset time (dashed line), whereas note-level alignment resolves small onset differences between individual notes (red boxes), reflecting performance-related asynchrony.

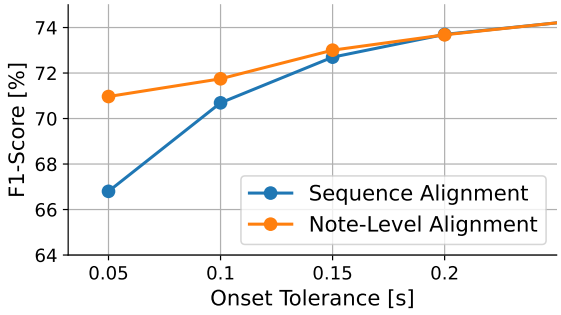

Figure 3: Mean F1-score (note-level, onset-only) as a function of onset tolerance on BSED for sequence-level and note-level alignment.

Symphony Excerpt Dataset (BSED) [2], which consists of short orchestral excerpts with multiple real performances and high-quality score–audio alignments. Together, these datasets allow for a systematic comparison across increasing levels of polyphonic complexity, from small ensembles to full orchestral settings.

We analyze this phenomenon in a stepwise manner. We first examine how frequently unison notes occur (Table 1), then study the temporal dispersion of notes that share the same score onset time, regardless of pitch, which reflects the general asynchrony (Figure 4). We subsequently repeat the same evaluation, focusing this time only on unison notes (Figure 5), and finally assess the impact of these effects on transcription evaluation (Figure 6).

A first important observation concerns the prevalence of unison notes, which we report in Table 1. In MusicNet and URMP unison notes are relatively rare, accounting for 2.7% and 7.6% of notes, respectively. In contrast, orchestral datasets such as PHENICX and BSED exhibit substantially higher rates, with 23.7% and 40.7% of notes involved in unison configurations, respectively. The lower rate in PHENICX compared to BSED is explained by the fact that PHENICX annotations are instrument-wise rather than part-wise: when two parts of the same instrument, e.g., French Horn 1 and 2, play the same pitch simultaneously (in the score), this is treated as a single

MusicNet	URMP	PHENICX	BSED
2.7%	7.6%	23.7%	40.7%

Table 1: Percentage of unison notes, defined as notes that share the same pitch and musical onset time with at least one other note in the score representation.

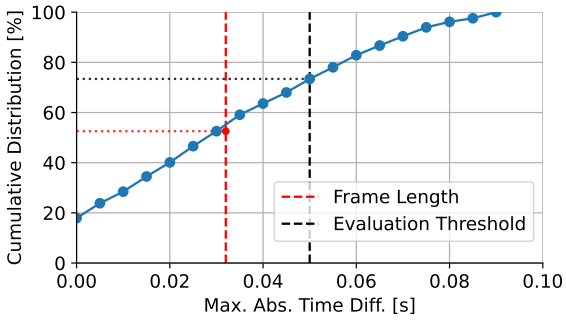

Figure 4: Temporal dispersion of notes sharing the same score onset. Cumulative distribution of maximum absolute onset time differences among notes sharing the same score onset time.

note event, whereas in BSED it counts as two separate notes.

Despite differences in annotation conventions, the results consistently reveal a clear trend: the prevalence of unison notes increases markedly with ensemble size and polyphonic complexity. This helps explain why the problem has received little attention in earlier AMT research. In simpler settings, unison notes are too infrequent to significantly affect evaluation, whereas in orchestral music they form a common structural element, for example through instrument doublings or instrument-section-wise unison passages.

A second important aspect is the temporal realization of simultaneous notes. Even when notes share the same score onset time, their physical onset times are typically not identical. Instead, small timing differences arise due to performance asynchrony, often on the order of tens of milliseconds. Figure 4 illustrates this effect by showing the distribution of onset differences within groups of notes that share the same musical onset on the PHENICX dataset. A substantial proportion exceeds typical frame resolutions (approximately 30 ms) and commonly used evaluation tolerances (approximately 50 ms), illustrating the prevalence of onset asynchrony in orchestral performances.

Focusing on the subset of unison notes, Figure 5 shows that similar temporal dispersion is present when notes share both the same pitch and score onset. For example, at a frame length of 32 ms, approximately 22% of unison notes would fall into different frames and could thus be counted as distinct onsets, confirming that onset asynchrony further compounds the note count ambiguity.

Finally, we examine the impact of these effects on evaluation using the PHENICX data and a given instrument-agnostic transcriber from [2]. There is no single correct

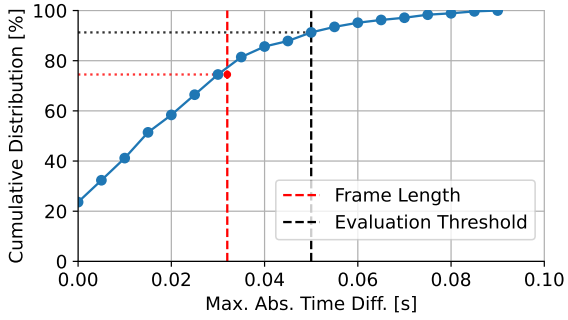

Figure 5: Temporal dispersion of unison notes. The plot shows the cumulative distribution of maximum absolute onset time differences within groups of notes that share both pitch and score onset.

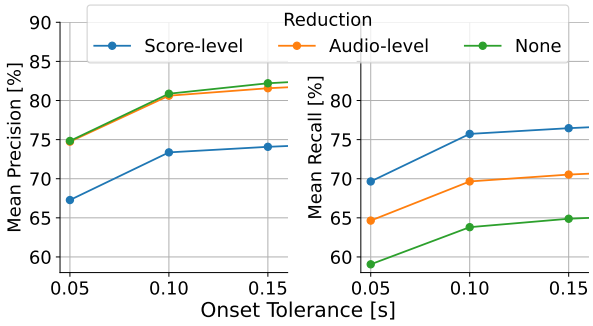

Figure 6: Impact of unison note handling on transcription evaluation. Mean F1-scores (note-level, onset only) on PHENICX obtained under different reduction strategies.

way to handle unison notes in the reference annotations and we explore three strategies: *Score-level reduction*, where all note events sharing the same pitch and score onset time are reduced to a single note event; *audio-level reduction*, where note events sharing the same pitch whose physical onsets fall within the same frame (of 32 ms length) are reduced to a single event; and *no reduction*, where the PHENICX annotations are used as-is, such that notes with even just slightly differing onset times are counted as separate events. As illustrated in Figure 6, these choices lead to systematically different precision and recall values for the same transcription output. Note that because of restrictions due to the annotation conventions of PHENICX (see Table 1), these differences likely represent a lower bound for the true impact. In summary, AMT evaluation on orchestral music depends on how unison notes are handled. We have shown that specific choices can affect precision and recall scores by more than 10% on PHENICX, and likely even more on other orchestral datasets.

Conclusions

Our observations establish a clear picture: unison notes occur frequently in complex musical settings, exhibit non-negligible temporal dispersion, and lead to systematically different evaluation outcomes depending on how they are represented.

We have argued that standard evaluation methodologies for automatic music transcription become fundamentally ill-defined in highly polyphonic, multi-instrument settings. The underlying issue is not model performance, but an intrinsic ambiguity of the task itself: unison notes and onset asynchrony break the assumption of a one-to-one correspondence between annotated and transcribed note events. Our analysis shows that these ambiguities are pervasive in orchestral music and directly affect evaluation outcomes, therefore reflecting annotation conventions and not only actual transcription quality.

More importantly, this challenges the task definition itself. In highly polyphonic settings, it is often unclear what the "correct" transcription should be. Consider, a musical part such as Violin I performed by multiple violins in unison: should the transcription contain one note per instrument, or a single note representing the entire section? In practice, the individual instruments often perceptually fuse into a single sound, making it neither well-defined nor musically meaningful to decompose this mixture into separate note events. More generally, does it make sense to represent every sounding event as an individual note?

These questions indicate that the problem goes beyond evaluation and call for a reconsideration of how music is represented in AMT. Future work should therefore not only refine evaluation metrics, but also explore alternative representations that explicitly account for ambiguity, for example by incorporating note multiplicities, grouping structures, or perceptually grounded descriptions. Unison notes expose a fundamental limitation of current AMT formulations, and addressing this limitation is essential for developing meaningful transcription methods in complex musical settings.

Acknowledgments: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 500643750 (MU 2686/15-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

References

[1] Benetos, E., Dixon, S., Duan, Z. & Ewert, S.: Automatic Music Transcription: An Overview. In: IEEE Signal Processing Magazine (2019), **36**, 1: 20–30.

[2] Berendes, H.-U., Saha, A., Maman, B. & Müller, M. (2026): Beethoven Symphony Excerpt Dataset (BSED): An Evaluation Dataset for Orchestral Music Transcription. Submitted for publication.

[3] Devaney, J.: Estimating onset and offset asynchronies in polyphonic score-audio alignment. In: Journal of New Music Research (2014), **43**, 3: 266–275.

[4] Gardner, J., Simon, I., Manilow, E., Hawthorne, C. & Engel, J. H.: MT3: Multi-Task Multitrack Music Transcription. In: Proceedings of the International Conference on Learning Representations (ICLR). Virtual (2022).

[5] Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J. H., Oore, S. & Eck, D.: Onsets and Frames: Dual-Objective Piano Transcription. In: Proceedings of the International Society for Music Information Retrieval Conference, (ISMIR), 50–57 (2018).

[6] Hawthorne, C., Simon, I., Swavely, R., Manilow, E. & Engel, J. H.: Sequence-to-Sequence Piano Transcription with Transformers. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 246–253. Online (2021).

[7] Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. A., Dieleman, S., Elsen, E., Engel, J. H. & Eck, D.: Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In: Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, Louisiana, USA (2019).

[8] Li, B., Liu, X., Dinesh, K., Duan, Z. & Sharma, G.: Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. In: IEEE Transactions on Multimedia (2019), **21**, 2: 522–535.

[9] Maman, B. & Bermano, A. H.: Unaligned Supervision for Automatic Music Transcription in The Wild. In: Proceedings of the International Conference on Machine Learning (ICML), 14918–14934. Baltimore, Maryland, USA (2022).

[10] Miron, M., Carabias-Orti, J. J., Bosch, J. J., Gómez, E. & Janer, J.: Score-Informed Source Separation for Multichannel Orchestral Recordings. In: Journal of Electrical and Computer Engineering (2016), **2016**, 8363507.

[11] Müller, M.: Fundamentals of Music Processing – Using Python and Jupyter Notebooks. 2nd Aufl. Springer Verlag (2021).

[12] Müller, M., Konz, V., Bogler, W. & Arifi-Müller, V.: Saarland Music Data (SMD). In: Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR). Miami, Florida, USA (2011).

[13] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D. & Ellis, D. P. W.: MIR_EVAL: A Transparent Implementation of Common MIR Metrics. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 367–372. Taipei, Taiwan (2014).

[14] Saha, A., Berendes, H.-U., Müller, M. & Maman, B.: Snapping Matters: Context-Aware Onset Refinement for Automatic Music Transcription. In: Proceedings of the International Computer Music Conference (ICMC). Hamburg, Germany (2026).

[15] Thickstun, J., Harchaoui, Z. & Kakade, S. M.: Learning Features of Music from Scratch. In: Proceedings of the International Conference on Learning Representations (ICLR). Toulon, France (2017).