# TOWARDS MUSICALLY INFORMED EVALUATION OF PIANO TRANSCRIPTION MODELS

**Patricia Hu**[1] **Lukáš Samuel Marták**[1,2] **Carlos Cancino-Chacón**[1] **Gerhard Widmer**[1,2]
[1] Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2] LIT AI Lab, Linz Institute of Technology, Austria
patricia.hu@jku.at

## ABSTRACT

Automatic piano transcription models are typically evaluated using simple frame- or note-wise information retrieval (IR) metrics. Such benchmark metrics do not provide insights into the transcription quality of specific musical aspects such as articulation, dynamics, or rhythmic precision of the output, which are essential in the context of expressive performance analysis. Furthermore, in recent years, MAESTRO has become the de-facto training and evaluation dataset for such models. However, inference performance has been observed to deteriorate substantially when applied on out-of-distribution data, thereby questioning the suitability and reliability of transcribed outputs from such models for specific MIR tasks. In this work, we investigate the performance of three state-of-the-art piano transcription models in two experiments. In the first one, we propose a variety of musically informed evaluation metrics which, in contrast to the IR metrics, offer more detailed insight into the musical quality of the transcriptions. In the second experiment, we compare inference performance on real-world and perturbed audio recordings, and highlight musical dimensions which our metrics can help explain. Our experimental results highlight the weaknesses of existing piano transcription metrics and contribute to a more musically sound error analysis of transcription outputs.

## 1. INTRODUCTION

Automatic Music Transcription (AMT) refers to the task of converting audio signals into symbolic music representations. The target output format can be a full symbolic score including quantized rhythm, time signature and pitch spelling information, or a mid-level physical MIDI(-like) representation, describing notes in terms of their onset and offset times, pitch and velocity [1–3].

AMT methods are typically evaluated using information retrieval (IR) metrics like precision, recall and F1 score [4]. These IR metrics can be computed at the level of frames, by comparing binary piano roll-like matrices, or at the level of note lists, by comparing notes in terms of their onset, offset, pitch and/or velocity attributes. Each error (i.e. misplaced frame or note activity) has equal weight, resulting in limited explanatory power of these metrics with respect to the underlying musical material [1, 5].

As in many other areas in MIR, the current state of the art is defined by deep neural networks [2,3,6,7]. To a large extent, this progress has been enabled by the release of the MAESTRO dataset [8], which made well-aligned audio-MIDI piano performance data available on a large scale. The most up-to-date version of the MAESTRO dataset [1] contains close to 200 hours of performance data from close to 1300 recordings of Western classical piano repertoire. State-of-the-art piano transcription systems achieve beyond 90% frame-level, or 80% note-level F1 scores on its test split [2, 3, 8, 9], and have led to the release of large-scale transcribed solo piano performance datasets [10, 11].

Although these results are impressive, we believe that two important aspects have been largely overlooked: first, the validity and (lack of) explanatory power of the standard evaluation metrics with respect to musically relevant information, and second, the reliability of these transcription models on out-of-distribution data. In this work, we address the first problem by proposing a set of musically informed evaluation metrics that support a more nuanced understanding of piano transcription errors. The metrics are intended to be used in the context of computational performance studies, and therefore focus on musical dimensions that are commonly studied in the context of expressive piano performance analysis and generation. We demonstrate our metrics on a subset of the MAESTRO dataset, which we transcribe using three state-of-the-art transcription models. In particular, we contrast the performance of these models, as evaluated with the standard IR metrics, with their performance on musical dimensions such as timing, articulation and dynamics which we can evaluate using our set of musically informed metrics.

Then, to elucidate the second problem, we re-record a subset of the MAESTRO dataset on a Yamaha Disklavier grand piano and further manipulate the audio recordings by adding different levels of noise and reverberation. An analysis of the outputs of these trained transcription models on these recordings provides some detailed insights into the lack of generalization on out-of-distribution data.

**Attribution:** P. Hu, L. Marták, C. Cancino-Chacón and G. Widmer, “Towards Musically Informed Evaluation of Piano Transcription Models”, in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

[1] https://magenta.tensorflow.org/datasets/maestro

We make our set of metrics, data and all experimental results available at `https://github.com/CPJKU/mpteval`.

## 2. RELATED WORK

This section briefly reviews the standard IR evaluation metrics along with criticism related to these, followed by a description of the benchmark datasets typically used for evaluating transcription methods.

Precision, recall and F1 score are the standard evaluation metrics used in AMT [1–4]. They can be computed either at the level of frames or at the level of notes. For frame-level evaluation, two binary piano roll matrices $M, \hat{M} \in \{0,1\}^{P \times T}$ are compared, where $p = 1, ..., P$ defines the pitch range and $t = 1, ..., T$ the time step (typically with a resolution of 10ms [4]). Both $M$ and $\hat{M}$ are sparse matrices where 1 at a given index $p, t$ indicates that a note with pitch $p$ is active at time frame $t$.

Note-level metrics are computed by comparing lists of notes, in which each note is described by a tuple describing the onset, offset, pitch, and (where predicted) velocity. Note-based metrics can be based on onset information only, onset and offset information (i.e., note durations), or on predicted onset, offset and velocity. In onset-only note evaluation, a note is considered correct if its onset falls within a ±50 ms threshold of its respective target onset. For onset-and-offset note-level evaluation, the note offset must fall within the greater of either an offset tolerance threshold of ±50 ms, or a duration threshold 20% of the ground truth duration [4]. If velocity is included in the evaluation, an estimated note is considered correct if its velocity (after some normalization and rescaling operations) falls within a 0.1 tolerance threshold of the velocity of the corresponding reference note [1].

The need for better (i.e., musically or perceptually sound) transcription metrics has been expressed by various researchers before. Hawthorne et al. [1] point out that frame- and onset-only note-level evaluation does not sufficiently capture musically relevant information. Similarly, Ycart et al. [5] and Daniel et al. [12] focus on the problem of perceptual saliency of different kinds of transcription errors and each propose a new, perceptually (more) valid transcription metric. Finally, McLeod and Steedman [13] focus on the problem of audio-to-score transcription and propose a new metric that jointly evaluates voice separation, metrical alignment, note value detection and harmonic analysis along with multi-pitch detection.

With respect to training and evaluation data for solo piano transcription, until the introduction of MAESTRO [8], MAPS [14] was used as the standard dataset. Apart from size, the biggest difference between the two is the diversity of the captured recording environments: while MAESTRO exclusively contains Disklavier recordings from the Yamaha International Piano e-Competition [2], MAPS contains Disklavier recordings and synthesized audio simulating various recording environments. The prevailing trend in evaluating current piano transcription models centers around the MAESTRO dataset [2, 3, 6], and most models that do include MAPS in their evaluation [1,3] use the split proposed in [15], which only includes Disklavier recordings in the test split. Both frame- and note-level metrics are usually computed for each piece in a given test set, and their mean is subsequently reported as the inference performance for a given model and dataset/split. Frame-level metrics are typically higher than note-level ones due to common known transcription errors such as merged or segmented notes.

## 3. MUSICALLY INFORMED METRICS

In this section we describe our proposed metrics that are meant to capture different musical dimensions commonly studied in the context of expressive performance. Each metric compares a ground truth to a predicted MIDI performance by measuring the Pearson correlation between a performance parameter computed from the ground truth and from the predicted MIDI, respectively. We choose a correlational measure to ensure all metrics fall into the same range. The goal is to quantify dimensions of musical quality of transcriptions that are otherwise obscured by standard IR metrics. In particular, we wish to capture dimensions that are important for computational performance studies that make use of automatically transcribed piano performances.

### 3.1 Timing

Timing can be described as expressive deviations from the metrical grid. A common measure of expressive timing in computational performance analysis is the inter-onset-interval (IOI), that is, the amount of time passed between two consecutive notes belonging to the same stream. [3]

To evaluate how well a transcription preserves the micro onset deviations, we predict a monophonic melody line and the accompaniment part (i.e., all notes not belonging to the melody line) in a given MIDI using the skyline algorithm for melody identification [17]. [4]

Then we compute the IOIs of these streams both on the ground truth and predicted performance, and measure their correlation. Note that for non-strictly monophonic streams (like the accompaniment part), the IOI between notes that belong to the same onset (i.e., chords) is zero. The result of this process gives us two measures, which we call *Melody IOI* and *Accompaniment IOI*.

### 3.2 Articulation

Articulation in expressive piano performance refers to how (adjacent) notes are played in terms of their duration, intensity, and clarity, resulting in expressive strategies such

[2] http://piano-e-competition.com/

[3] We use the term *stream* as a generalization of the concept of a voice in polyphonic music [16].

[4] The skyline algorithm has been shown to be very competitive in identifying melody lines in Western classical piano music, even when compared to more recent machine learning-based algorithms. (e.g., see Figure 5 in [18]).

as *legato*, *staccato* or *marcato*. Computationally, articulation is measured as the ratio between the time interval from the offset of the current note to the onset of the next note, and the time between the onsets of the two notes. [19–21].

We use the skyline algorithm [17] to extract monophonic melody and bass lines within a performance, and compute a sequence of KOR values, for each pair of successive note events, for both the target and the predicted performance MIDI for both streams, and their ratio. We define three metrics for capturing articulation:

1. *Melody KOR*: the correlation between the KOR sequences of the melody lines of the ground truth and the predicted performance MIDI.

2. *Bass KOR*: the correlation between the KOR sequences of the bass lines of the ground truth and the predicted performance MIDI.

3. *Ratio KOR*: for this metric, we consider the ratio of KOR sequences of the melody to the bass line. A ratio KOR greater than 1 indicates that the melody voice is played more legato than the bass voice. The Ratio KOR metric is computed as correlation of this ratio between the ground truth and the predicted performance MIDI.

### 3.3 Harmony

Aspects such as harmonic tension have been shown to be determining factors for various performance decisions (particularly relating to expressive tempo and dynamics [22, 23]). To quantify how well harmonic tension is preserved in a transcription, we use two features proposed by Herremans and Chew [24] based on Chew's spiral array model [25]. This model is a three dimensional representation of pitch classes, chords and keys constructed in such a way that spatial proximity represents close tonal relationships. [5] We use two metrics to capture the preservation of harmonic tension:

1. *Cloud Diameter*: this metric measures the maximal tonal distance as the maximum dispersion between notes in a musical segment

2. *Cloud Momentum*: this metric captures the harmonic movement in a segment as the tonal distance between consecutive sections.

For both metrics, we compute the respective feature on overlapping windows for both the ground truth and transcribed MIDI, and measure their correlation.

### 3.4 Dynamics

For comparing the performance of transcription models regarding expressive dynamics, we use the loudness ratio of the melody and bass lines as a proxy to identify how well a transcription preserves the dynamics of the performance. We estimate the loudness as the "energy" of a stream (i.e., melody or bass line), which is computed using the MIDI velocity following a model proposed by Dannenberg [26]. The loudness ratio is then computed as follows (cf. Equation 8 in [26]):

| **composer** | **pieces** | **performances** | **duration (min)** |
|---|---|---|---|
| Bach | 1 | 7 | 23.36 |
| Beethoven | 5 | 28 | 285.54 |
| Chopin | 4 | 15 | 150.28 |
| Debussy | 2 | 3 | 32.06 |
| Glinka | 1 | 2 | 10.35 |
| Haydn | 3 | 9 | 90.23 |
| Liszt | 3 | 12 | 58.98 |
| Mozart | 2 | 4 | 29.02 |
| Rachmaninoff | 2 | 3 | 11.87 |
| Schubert | 3 | 17 | 107.27 |
| Scriabin | 1 | 5 | 55.05 |
| **Total** | **27** | **105** | **854.01** |

**Table 1**: Overview of chosen composers, pieces, and performances in the MAESTRO subset in our evaluation set.

$$\mathrm{R}(t) = \log\left(\frac{m \cdot \mathrm{vel}_{mel}(t) + b}{m \cdot \mathrm{vel}_{bass}(t) + b}\right) \tag{1}$$

where $\mathrm{vel}_{mel}(t)$ and $\mathrm{vel}_{bass}(t)$ are the MIDI velocities of the melody and bass lines at time $t$, respectively, and $m$ and $b$ are constant parameters that depend on the dynamic range of the audio signal. We compute the loudness ratio for both the ground truth and estimated performance MIDI, and compute the correlation between these ratios as our metric for dynamics. [6]

## 4. DATA

For our experiments, we create an evaluation set with three subsets:

1. *MAESTRO*: We select audio recordings from the MAESTRO dataset, covering a diverse range of musical repertoire, composers, and performers, using all (train, validation, and test) splits as provided by the authors [3]. This choice tests whether the split category affects model generalization. [7] An overview of the selected subset is shown in Table 1.

2. *Disklavier*: We re-record our MAESTRO subset on a Yamaha Disklavier Enspire ST C1X using the Focusrite Scarlett 18i8 and a pair of AKG P420 microphones in a moderately bright, fully carpeted room with asymmetric geometry and low background noise level.

3. *revnoise*: To simulate more challenging real-world environments, we further add perturbations using

[5] We chose this model for its simplicity and music-theoretical grounding. Note that these features were designed for Western tonal music and may be less effective in capturing tension in other types of music.

[6] Dannenberg's model was chosen for its simplicity, relying only on MIDI velocity and dynamic range (note that parameters $m$ and $b$ cancel each other out when computing the correlation of the loudness ratio).

[7] The official MAESTRO splits [8] ensure a unique piece-to-split mapping.

different levels of reverberation and noise (see Section 6) on selected recordings from both the *MAESTRO* and *Disklavier* subsets.

We compare three state-of-the-art piano transcription models: Onsets and Frames [8] and the Transformer transcription model [3] by Google/Magenta (which we will refer to as OaF and T5 respectively, in the following), and the high-resolution onset and offset regression model by Bytedance [2] (referred to as Kong).

We transcribe all the recordings in our MAESTRO, *Disklavier*, and *revnoise* subsets using the (officially provided) trained models, and these transcriptions then form our evaluation set. Note that all audio recordings from the *Disklavier* and *revnoise* subsets are only used for testing; we use the MAESTRO-trained models as they are provided by the respective authors via their repositories.

## 5. DEMONSTRATION OF MUSICALLY INFORMED METRICS

We now discuss the experimental results obtained with the three systems and explain the relation between our metrics and the standard IR metrics as computed on transcriptions of the MAESTRO subset of our evaluation set. We focus in particular on the musical dimensions that can be better understood through our metrics. For the standard evaluation, we include the frame-level score, and all note-level F1 scores other than the onset-only one as it does not capture offset and velocity information. We compute the note-level metrics with the official `mir_eval` python implementation [27].

We start our discussion with Table 2, which summarizes the evaluation results per model and metric. For comparative reasons, we also include a perceptually informed piano transcription metric, PEAMT [5].

Generally, it can be observed that the Kong model performs the best across most metrics. This implies that most of our metrics, overall, correlate with the performance ranking as measured on the standard metrics. Furthermore, it can be seen that the two Magenta models perform considerably different when measured against frame-level F1 score, yet this difference becomes less pronounced when evaluated on note-level metrics, which would suggest superior performance of the T5 model. Comparing these results to the model performance as evaluated on our set of metrics, however, reveals that while both models perform similarly on onset time prediction, the T5 model is worse at adequately capturing note durations, particularly in lower voices/frequency ranges, but superior in estimating MIDI velocity and the overall loudness ratio between voices than the OaF model. Lastly, we can observe that the perceptually informed PEAMT metric correlates most with the frame-level and harmony metric *Cloud Momentum*, which might suggest (if PEAMT is indeed a veridical listening model) that listeners place relatively high importance on harmonic context.

| **Metric** | **OaF** | **Kong** | **T5** |
|---|---|---|---|
| Frame F1 | 0.8710 | **0.9138** | 0.7048 |
| Note Offset F1 | 0.6167 | **0.8736** | 0.6358 |
| Note Offset Velocity F1 | 0.5917 | **0.8587** | 0.6309 |
| Melody IOI | 0.2377 | **0.5481** | 0.2217 |
| Accompaniment IOI | 0.2168 | 0.3679 | **0.4329** |
| Melody KOR | 0.4057 | **0.7415** | 0.2825 |
| Bass KOR | 0.2638 | **0.6967** | 0.2672 |
| Ratio KOR | 0.4247 | **0.6938** | 0.3094 |
| Cloud Diameter | 0.7240 | **0.8301** | 0.7472 |
| Cloud Momentum | **0.2461** | 0.2250 | 0.1671 |
| Dynamics | 0.5501 | **0.6503** | 0.6355 |
| PEAMT [5] | **0.6570** | 0.6241 | 0.5789 |

**Table 2**: Model performance measured by standard metrics, our musically informed metrics, and PEAMT [5] on the MAESTRO subset of our evaluation set

We continue our discussion in Figure 1, which compares the note-offset F1 score per composer (averaged over pieces and performers) and model to our musically in-

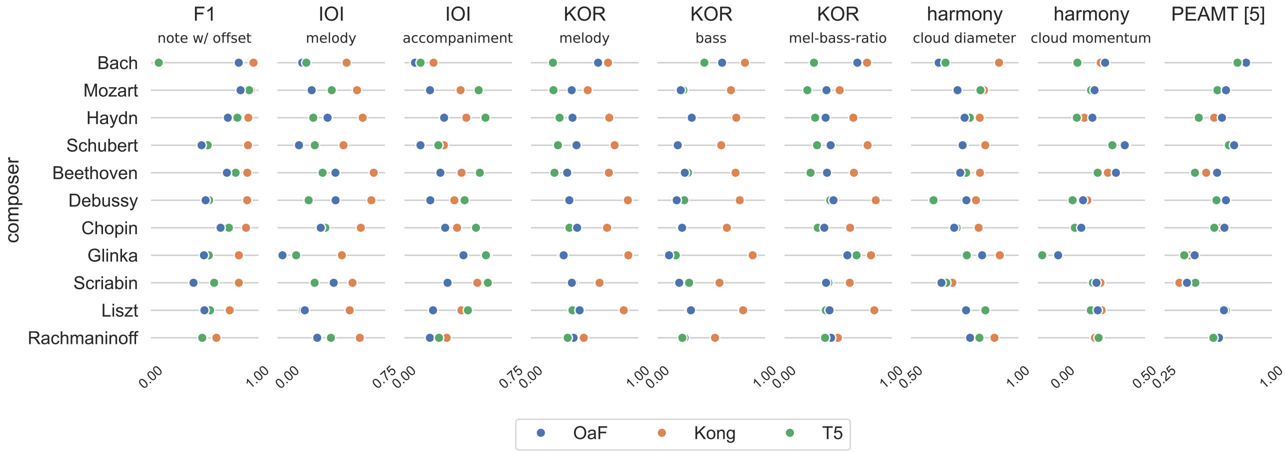

**Figure 1**: Model performance comparison as evaluated on note-offset F1 score and our proposed musical metrics, by composer.

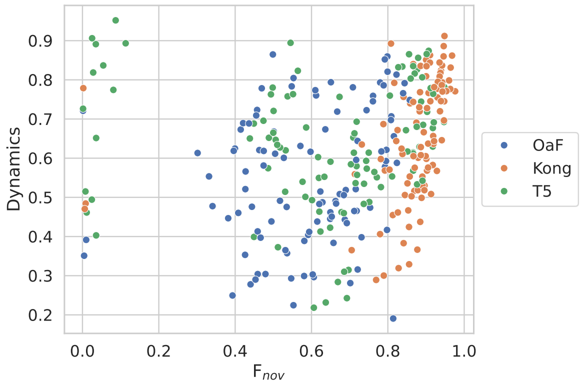

**Figure 2**: Relationship between model performance evaluated on note-offset-velocity F1 score and our proposed dynamics measure.

formed timing, articulation and harmony metrics. We can see again that the Kong model performs best across all composers and metrics except for the *Accompaniment IOI* timing and the *Cloud Momentum* harmony metrics. The fact that it performs better on the *Bass KOR* articulation metric, but poorly on those timing accompaniment and harmony metrics might suggest that this model detects many out-of-key extra notes that both erroneously influence the IOI sequence on the accompaniment part, and the estimation of the tonal context. Interestingly, we can can also observe, as a general trend, that the F1 score somewhat deteriorates with increasing virtuoso and challenging musical repertoire. This general deterioration with increasing musical difficulty is not reflected correspondingly by our metrics, which show more variation with respect to different composers and aspects of the underlying music: While the F1 score for Bach, Mozart, Haydn and Schubert all suggest a near-perfect transcription, our metrics indicate more diverse results, and e.g., suggest poor(er) accuracy (and therefore reliability in a performance study context) in the *Melody IOI* and *Melody KOR* metrics. Another illustrative example can be found in the case of Chopin: here again the F1 score (particularly of the Kong model) would suggest a highly accurate transcription output, while our metrics reveal that the expressive dimension of articulation is not well captured. Lastly, we can observe again that the PEAMT and harmony metric *Cloud Momentum* show a similar trend for most composers, suggesting a greater weight of the harmonic context in that trained metric.

We conclude our discussion by examining the dynamics aspect. Figure 2 illustrates the relationship between the note-offset-velocity F1 score and our proposed *Dynamics* metric. While both metrics show a weak correlation (Pearson $r = 0.21$), the figure also indicates that our metric evaluates dynamics in a more differentiated way and leads to a wider range of evaluation results than the standard metric. Note that our metric only evaluates the dynamics aspect, in particular how well the overall balance in loudness between different voice streams is preserved in a transcription. It does not account for onset, offset and pitch information, which also explains the results in the very left part of the figure that score low on F1 score but high on our metric.

## 6. OUT-OF-DISTRIBUTION INFERENCE

In this section we illustrate the problem of out-of-distribution performance of the models analysed. We believe that this is an important aspect to emphasize, as transcription models are ultimately intended to be (and have been) used on real-world audio performances [10, 11].

We approach the problem in two stages: First, we elucidate the problem by performing a short evaluation of the three analysed models on real-world recordings using only the standard IR metrics. Second, we simulate more challenging real-world environments using different levels of noise and reverberation, and evaluate the analysed models again using the standard and our proposed metrics, where we highlight how our musically informed metrics can reveal aspects that the standard metrics would otherwise have missed.

### 6.1 Generalization on real-world recordings

Table 3 shows the mean frame- and note-level F1 scores per model and per piano/acoustic recording environment on the three splits of the MAESTRO dataset (as they are officially defined). [8]

| | | **frame** | | | **note$_{\text{off}}$** | | | **note$_{\text{off-vel}}$** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **split** | **model/audio** | **OaF** | **Kong** | **T5** | **OaF** | **Kong** | **T5** | **OaF** | **Kong** | **T5** |
| train | MAESTRO | 0.8807 | 0.9207 | 0.7262 | 0.6183 | 0.8899 | 0.6350 | 0.5929 | 0.8756 | 0.6308 |
| | Disklavier | 0.8185 | 0.8508 | 0.6157 | 0.5269 | 0.7384 | 0.5132 | 0.4853 | 0.6660 | 0.4663 |
| validation | MAESTRO | 0.8404 | 0.8936 | 0.6546 | 0.6492 | 0.8617 | 0.7117 | 0.6236 | 0.8471 | 0.7063 |
| | Disklavier | 0.7696 | 0.8678 | 0.6093 | 0.5539 | 0.8142 | 0.6570 | 0.5161 | 0.7480 | 0.6031 |
| test | MAESTRO | 0.8527 | 0.9002 | 0.6552 | 0.5931 | 0.8215 | 0.5968 | 0.5695 | 0.8041 | 0.5896 |
| | Disklavier | 0.8049 | 0.8530 | 0.6048 | 0.4989 | 0.6979 | 0.5135 | 0.4607 | 0.6304 | 0.4624 |

**Table 3**: Frame-, note-offset, and note-offset-velocity F1 score results computed on our evaluation set, grouped per data set split, evaluated model and piano / audio environment.

[8] We note that for the two Magenta models, OaF and T5, our evaluation results do not come close to the reported ones in [8] and [3]. The

We group the results per split to test whether the analysed models would perform worse on out-of-distribution recordings of performances (pieces) from the test set compared to those from the train/validation sets. Next we conduct a Kruskal Wallis ANOVA [28] to test for differences between frame-, note-offset and note-offset-velocity F1 scores, grouping the evaluation scores each by *split* and by *audio environment* and comparing each model separately. For each ANOVA we use a significance threshold of $\alpha = 0.05$. The ANOVA on the *audio environment* dimension show a statistically significant difference ($p < 0.05$) between the MAESTRO and Disklavier audio recordings for all three analysed models.

The ANOVA on the *split* dimension yields more differentiated results: For all models, the frame-level F1 scores are significantly different, and there are no statistically significant differences in the note-offset-velocity F1 scores. For the model by Kong, the note-offset-score is significantly different depending on the split, whereas the two Magenta models show no significant differences. This suggests that the most musically meaningful metric from the current set of standard metrics [1] does not sufficiently capture overfitting tendencies.

### 6.2 Evaluation on perturbed audio recordings

Similar as in Section 5, we again compare our musically informed metrics to the standard IR and the PEAMT metrics, however, this time on a set of more challenging audio recordings. To this end, we choose six (MAESTRO and *Disklavier*) audio recordings which we artificially perturb by introducing reverberation and synthetic noise. We use three Impulse-Response filters, modelling short, medium and long reverberation times (RT60@1kHz $\in \{0.19, 1.85, 10.5\}$ seconds) and sourced from the OpenAIR [9] database. We further add white noise into the recordings at three different Signal-to-Noise Ratio levels (SNR$_{dB} \in \{24, 12, 6\}$). Following a factorial design with these two independent variables, each with four levels, we first perturb the audio recordings on all conditions, and transcribe these recordings using all three analysed models. Following this procedure, we obtain 284 transcribed MIDI performances. [10]

Each grid cell in Figure 3 compares the mean note-offset F1 scores per model to the *Melody IOI* timing metric and *Cloud Momentum* harmony metric, where grid rows represent increasing reverb levels, and grid columns represent increasing noise levels. Generally, it can be observed that the performance range of models as measured by the F1 score is notably reduced compared to our metrics, indicating that our metrics possess higher discriminative capacity than the standard ones.

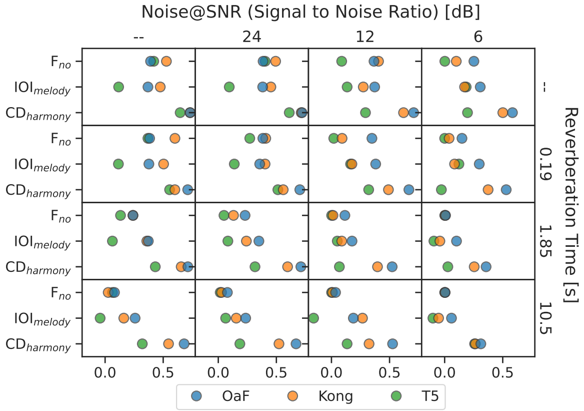

**Figure 3**: Performance degradation measured by note-offset F1, *Melody IOI* and *Cloud Momentum* metrics.

As expected, the inference performance of all three analysed models deteriorates with increasing noise and reverberation levels, though the deterioration is less pronounced on the noise than on the reverberation axis. Furthermore, analysing the results on the timing metric *Melody IOI* suggests that the model by Kong predicts onset times worse with increasing noise levels, while the onset times prediction by OaF seems to be more resistant to this form of perturbation. Finally, the results measured on the harmony metric *Cloud Momentum* suggest that the overall harmonic context is relatively well preserved at higher perturbation levels by the OaF and Kong models, and less so by the T5 model.

## 7. CONCLUSION

In this study, we investigated two aspects that are commonly neglected in the evaluation of transcription models: (i) limited explanatory power of the standard IR evaluation metrics with respect to the underlying musical material, and (ii) poor inference on out-of-distribution data. We study both problems in the context of solo piano transcription, and, in addressing the first aspect, propose a set of musically informed metrics designed to capture more musically relevant information, particularly for the context of computational studies of expressive performance.

We demonstrated our metrics on transcriptions obtained by three state-of-the-art piano transcription models on a subset of the MAESTRO dataset, the de-facto standard train and test set for current transcription models, and highlighted musical dimensions for which they provide more informative value than the standard information retrieval metrics. We have further illustrated the lack of generalization with respect to the acoustic environment, both on real-world and perturbed audio recordings.

Future work in this direction may include an extension and further validation of our new musically informed metrics, in order to capture additional qualities of expressive performance, potentially by making use of score alignment information. Additionally, a listening study with human experts could help further investigate the perceptual validity of our proposed metrics.

differences are particularly pronounced in the note-level F1 scores.

[9] `https://www.openair.hosted.york.ac.uk`

[10] Note that 6 pieces x 4 noise levels x 4 reverberation levels x 3 models yield 288 transcriptions, but 4 recordings (each at the two higher most of either reverberation and/or noise levels) resulted in empty transcriptions (zero predicted note events) by the T5 model, and are therefore excluded from the evaluation.

## 8. ACKNOWLEDGMENTS

This work receives funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 101019375 (*Whither Music?*). The LIT AI Lab is supported by the Federal State of Upper Austria.

## 9. REFERENCES

[1] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, 2018, pp. 50–57.

[2] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[3] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, "Sequence-to-sequence piano transcription with transformers," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, 2021, pp. 246–253.

[4] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 315–320.

[5] A. Ycart, L. Liu, E. Benetos, and M. Pearce, "Investigating the perceptual validity of evaluation metrics for automatic piano music transcription," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 3, no. 1, pp. 68–81, 2020.

[6] W.-T. Lu, J.-C. Wang, and Y.-N. Hung, "Multitrack Music Transcription with a Time-Frequency Perceiver," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] B. Maman and A. H. Bermano, "Unaligned Supervision for Automatic Music Transcription in The Wild," in *International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 14 918–14 934.

[8] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *International Conference on Learning Representations, (ICLR)*, 2019.

[9] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-task multitrack music transcription," in *International Conference on Learning Representations (ICLR)*, 2022.

[10] Q. Kong, B. Li, J. Chen, and Y. Wang, "GiantMIDI-piano: A large-scale MIDI dataset for classical piano music," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 5, no. 1, pp. 87–98, 2022.

[11] H. Zhang, J. Tang, S. Rafee, S. Dixon, and G. Fazekas, "ATEPP: A dataset of automatically transcribed expressive piano performance," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023, pp. 446–453.

[12] A. Daniel, V. Emiya, and B. David, "Perceptually-based evaluation of the errors usually made when automatically transcribing music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2008, pp. 550–556.

[13] A. McLeod and M. Steedman, "Evaluating Automatic Polyphonic Music Transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 42–49.

[14] V. Emiya, N. Bertin, B. David, and R. Badeau, "MAPS - A piano database for multipitch estimation and automatic transcription of music," INRIA, Research Report, 2010. [Online]. Available: https://inria.hal.science/inria-00544155

[15] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 927–939, 2016.

[16] D. Temperley, "A Unified Probabilistic Model for Polyphonic Music Analysis," *Journal of New Music Research*, vol. 38, no. 1, pp. 3–18, 2009. [Online]. Available: https://doi.org/10.1080/09298210902928495

[17] A. Uitdenbogerd and J. Zobel, "Melodic matching techniques for large music databases," in *Proceedings of the ACM international conference on Multimedia*, 1999, pp. 57–66. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/319463.319470

[18] F. Simonetta, C. Cancino-Chacón, S. Ntalampiras, and G. Widmer, "A convolutional approach to melody line identification in symbolic scores," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 924–931.

[19] R. Bresin and G. Umberto Battel, "Articulation strategies in expressive piano performance analysis of legato, staccato, and repeated notes in performances of

the Andante movement of Mozart's Sonata in G Major (K 545)," *Journal of New Music Research*, vol. 29, no. 3, pp. 211–224, 2000.

[20] B. H. Repp, "Acoustics, perception, and production of legato articulation on a computer-controlled grand piano," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1878–1890, 1995.

[21] C. Drake and C. Palmer, "Accent structures in music performance," *Music Perception*, vol. 10, no. 3, pp. 343–378, 1993.

[22] C. Cancino-Chacón and M. Grachten, "A computational study of the role of tonal tension in expressive piano performance," 2018. [Online]. Available: https://arxiv.org/pdf/1807.01080

[23] D. Herremans and E. Chew, "Towards emotion based music generation: A tonal tension model based on the spiral array," in *The Annual Meeting of the Cognitive Science Society*, 2019, pp. 52–53. [Online]. Available: https://hal.science/hal-03277753/document

[24] D. Herremans, E. Chew *et al.*, "Tension ribbons: Quantifying and visualising tonal tension." in *Second International Conference on Technologies for Music Notation and Representation (TENOR)*, 2016. [Online]. Available: https://hal.science/hal-03165896/document

[25] E. Chew, "Playing with the edge: Tipping points and the role of tonality," *Music Perception: An Interdisciplinary Journal*, vol. 33, no. 3, pp. 344–366, 2016. [Online]. Available: https://hal.science/hal-03787367/document

[26] R. B. Dannenberg, "The Interpretation of MIDI Velocity," in *International Computer Music Conference (ICMC)*, 2006, pp. 193–196.

[27] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. Ellis, "MIR_EVAL: A Transparent Implementation of Common MIR Metrics." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[28] A. Vargha and H. D. Delaney, "The Kruskal-Wallis Test and Stochastic Homogeneity," *Journal of Educational and behavioral Statistics*, vol. 23, no. 2, pp. 170–192, 1998.