A METRIC FOR MUSIC NOTATION TRANSCRIPTION ACCURACY

Andrea Cogliati
University of Rochester
Electrical and Computer Engineering
andrea.cogliati@rochester.edu

Zhiyao Duan
University of Rochester
Electrical and Computer Engineering
zhiyao.duan@rochester.edu

ABSTRACT

Automatic music transcription aims at transcribing musical performances into music notation. However, most existing transcription systems only focus on parametric transcription, i.e., they output a symbolic representation in absolute terms, showing frequency and absolute time (e.g., a pianoroll representation), but not in musical terms, with spelling distinctions (e.g., A♭ versus G♯) and quantized meter. Recent attempts at producing full music notation output have been hindered by the lack of an objective metric to measure the adherence of the results to the ground truth music score, and had to rely on time-consuming human evaluation by music theorists. In this paper, we propose an edit distance, similar to the Levenshtein Distance used for measuring the difference between two sequences, typically strings of characters. The metric treats a music score as a sequence of sets of musical objects, ordered by their onsets. The metric reports the differences between two music scores based on twelve aspects: barlines, clefs, key signatures, time signatures, notes, note spelling, note durations, stem directions, groupings, rests, rest duration, and staff assignment. We also apply a linear regression model to the metric in order to predict human evaluations on a dataset of short music excerpts automatically transcribed into music notation.

1. INTRODUCTION

Automatic Music Transcription (AMT) is the process of inferring a symbolic representation of a musical performance. Despite four decades of active research, AMT is still an open problem, with humans being able to achieve better results than machines [2]. AMT systems can be broadly classified into two categories according to the chosen symbolic representation: parametric transcription and music notation transcription. Parametric transcription systems output a parametric representation of the musical performance, such as an unquantized MIDI pianoroll [14]. This representation is expressed in physical terms, such as seconds for note onset and duration, and hertz or MIDI numbers for pitch [7]. It can faithfully represent the musical performance, but normally it does not explicitly encode high-level musical structures, such as key, meter and voicing [21]. Music notation transcription systems, on the other hand, output a common music notation that human musicians read. This representation is expressed in musically meaningful terms, such as quantized meter for note onset and duration, and spelling distinctions (e.g., A♭ versus G♯) for pitch. Compared to parametric transcription, music notation transcription is generally more desirable for many applications connecting humans and machines, such as computational musicological analysis and music tutoring systems. The vast majority of existing AMT methods, however, are parametric transcription systems.

Researchers have put considerable effort toward building music notation transcription systems by identifying musical structures from unquantized parametric representations, especially MIDI files, from both MIR and cognitive perspectives [20]. Cambouropoulos [3] described the key components necessary to convert a MIDI performance into music notation: identification of elementary musical objects (i.e., chords, arpeggiated chords, and trills), beat identification and tracking, time quantization and pitch spelling. Takeda et al. [18] describe a Hidden Markov Model (HMM) for the automatic transcription of monophonic MIDI performances. Cemgil [4] presents a Bayesian framework for music transcription, identifying some issues related to automatic music typesetting (i.e., the automatic rendering of a musical score from a symbolic representation), in particular tempo quantization, and chord and melody identification. Karydis et al. [12] proposed a perceptually motivated model for voice separation capable of grouping polyphonic groups of notes, such as chords or other forms of accompaniment figures, into a perceptual stream. A more recent paper by Grohganz et al. [11] introduced the concepts of score-informed MIDI file (S-MIDI), in which musical tempo and beats are properly represented, and performed MIDI file (P-MIDI), which records a performance in absolute time. The paper also presented a procedure to approximate an S-MIDI file from a P-MIDI file – that is, to detect the beats and the meter implied in the P-MIDI file, starting from a tempogram then analyzing the beat inconsistency with a salience function based on autocorrelation.

Researchers have also attempted to infer musical structures directly from audio. Ochiai et al. [16] proposed a model for the joint estimation of note pitches, onsets, offsets and beats based on Non-negative Matrix Factorization

(NMF) constrained with a rhythmic structure modeled with a Gaussian mixture model. Collins et al. [8] proposed a model for multiple fundamental frequency estimation, beat tracking, quantization, and pattern discovery. The pitches are estimated with a neural network. An HMM is separately used for beat tracking. The results are then combined to quantize the notes. Note spelling is performed by estimating the key of the piece and assigning to MIDI notes the most probable pitch class given the key.

An immediate problem arising when building a music notation transcription system by incorporating the abovementioned musical structure inference methods is to find an appropriate way to evaluate the transcription accuracy of the system. In our prior work [7], we asked music theorists to evaluate music notation transcriptions along three different musical aspects, i.e., the pitch notation, the rhythm notation, and the note positioning. However, subjective evaluation is time consuming and difficult to scale to provide enough feedback to further improve the transcription system. It would be very helpful to have an objective metric for music notation transcription, just like the standard metric F-measure for parametric transcription [1]. Considering the inherent complexity of music notation, such a metric would need to take into account all of the aspects of the high-level musical structures in the notation. To the best of our knowledge, there is no such metric, and the goal of this paper is to propose such a metric.

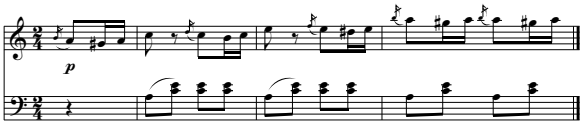
Specifically, in this paper we propose an edit distance, based on similar metrics used in bioinformatics and linguistics, to compare a music transcription with the ground-truth score. The design of the metric was guided by a data-driven approach, and by simplicity. The metric is calculated in two stages. In the first stage, the two scores are aligned based on the pitch content; in the second stage, the differences between the two scores are accumulated, taking into account twelve different aspects of music notation: barlines, clefs, key signatures, time signatures, notes, note spelling, note durations, stem directions, groupings, rests, rest duration, and staff assignment. This will serve the same purpose as F-measure in evaluating parametric transcription. To validate the saliency and the usefulness of this metric we also apply a linear regression model to the errors measured by the metric to predict human evaluations of transcriptions.

2. BACKGROUND

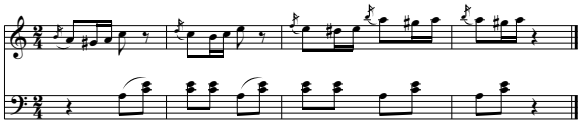
Approximate sequence comparison is a typical problem in bioinformatics [13], linguistics, information retrieval, and computational biology [15]. Its purpose is to find similarities and differences between two or more sequences of elements or characters. The sequences are assumed sufficiently similar but potentially corrupted by errors. Possible differences include the presence of different elements, missing elements or extra elements. Several metrics have been proposed to measure the distance between two sequences, including the family of edit metrics [15], and gap-penalizing alignment techniques [13].

A music score in traditional Western notation can be viewed as a sequence of musical characters, such as clefs, time and key signatures, notes and rests, possibly occurring concurrently, such as in simultaneous notes or chords. Transcription errors include alignment errors due to wrong meter estimation or quantization, extra or missing notes and rests, note and rest duration errors, wrong note spelling, wrong staff assignment, wrong note grouping and beaming, and wrong stem direction. All of these errors contribute to a various degree to the quality of the resulting transcription. However, the impact of each error and error category has not, to the best of our knowledge, been researched.

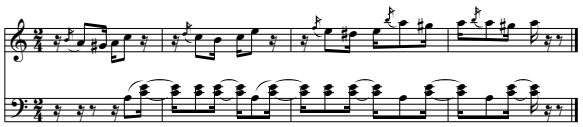
As an example, Fig. 1 shows two transcriptions of the same piece. Both transcriptions contain similar errors, i.e., wrong meter detection, but the transcription in Fig. 1c is arguably worse than that in Fig. 1b. A similar problem can be observed with the standard F-measure typically used to evaluate parametric transcriptions [1]; while the metric is objective and widely used, the impact of different errors on the perceptual quality of a transcription has not been researched. Intuitively, certain errors, such as extra notes outside of the harmony, should be perceptually more objectionable than others, such as octave errors. This is the reason for both proposing an objective metric and correlating the metric with human evaluations of transcriptions.



(a) Ground truth



(b) Transcription with a wrong pickup measure



(c) Transcription off by a 16th note

Figure 1: Comparison of two transcriptions of the same piece containing similar errors but with different readability.

3. PROPOSED METHOD

The proposed metric is calculated in two stages: in the first stage, the transcription is aligned with the ground-truth music notation based on its pitch content only, i.e., all of the other objects, such as rests, barlines, and time and key signatures are ignored; in the second stage, all of the objects occurring at the aligned portions of the scores

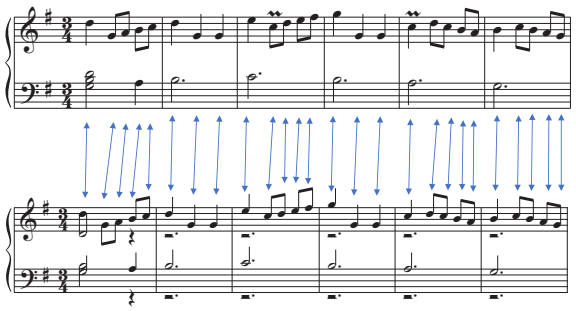


Figure 2: Alignment between the ground-truth (top) and a transcription (bottom) of Bach's Minuet in G. Arrows indicate aligned beats.

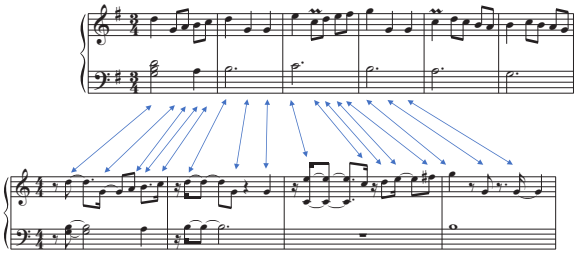


Figure 3: Alignment between the ground-truth (top) and another transcription (bottom) of Bach's Minuet in G. Arrows indicate aligned beats.

are grouped together and compared. The metric reports the differences in aligned portions in terms of twelve aspects: barlines, clefs, key signatures, time signatures, notes, note spelling, note durations, stem directions, groupings, rests, rest duration, and staff assignment.

Some algorithms to efficiently calculate certain edit distances, e.g., the Wagner-Fischer algorithm to calculate the Levenshtein distance between two strings, are able to align two sequences and calculate the edit costs in a single stage. We initially tried to apply the same strategy to our problem, but we discovered that the algorithm was not sufficiently robust, especially with transcriptions highly corrupted by wrong meter estimation. Intuitively, notes are the most salient aspects of music, so it is arguable that the alignment of two transcriptions should be based primarily on that aspect, while the overall quality of the transcription should be judged on a variety of other aspects.

The ground truth and the transcription are both encoded in MusicXML, a standard format to share sheet music files between applications [10]. The two scores are aligned using Dynamic Time Warping [17]. The local distance is simply the number of mismatching pitches, regardless of duration, spelling and staff positioning.

To illustrate the purpose of the initial alignment, we show two examples in Fig. 2 and Fig. 3. The alignment stage outputs a list of pairs of aligned beats. Fig. 2 shows the alignment of a fairly good transcription of Bach's Minuet in G from the Notebook for Anna Magdalena Bach, with the ground truth, which corresponds to the following sequence, expressed in beats, numbered as quarter notes starting from 0 (GT is ground truth, T is transcription):

GT	0.0	1.0	1.5	2.0	2.5	3.0	4.0
T	0.0	1.0	1.5	2.0	2.5	3.0	4.0
4.0	5.0	6.0	7.0	7.5	8.0	8.5	9.0
5.0	5.0	6.0	7.0	7.5	8.0	8.5	9.0
10.0	10.0	11.0	12.0	13.0	13.5	14.0	14.5
10.0	11.0	11.0	12.0	13.0	13.5	14.0	14.5
15.0	16.0	16.5	17.0	17.5			
15.0	16.0	16.5	17.0	17.5			

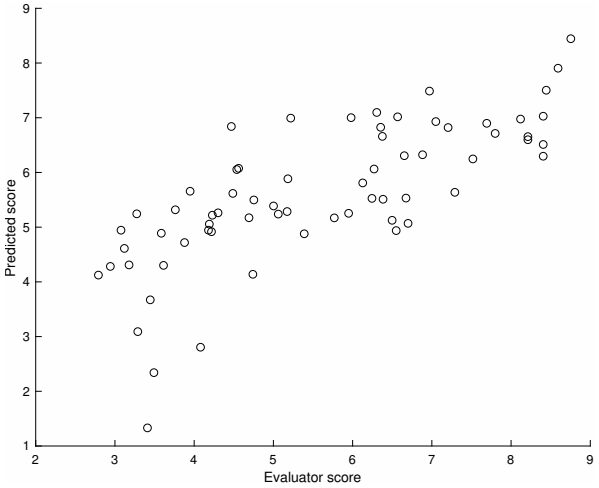
In this case, since the transcription is properly aligned with the ground truth, the sequence is just a list of all equal numbers, one for each onset of the notes in the score. However, beat 4.0 in the ground truth is matched with beats 4.0 and 5.0 in the transcription; the same happens for beats 10.0 and 11.0, so DTW cannot properly distinguish repeated pitches. Only one alignment is shown in the figure for clarity.

Fig. 3 shows an example of an alignment for a badly aligned transcription of the same piece. The corresponding sequence is the following:

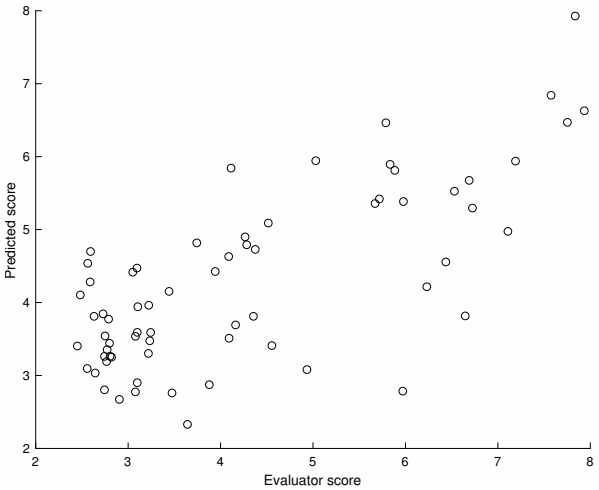
GT	0.0	0.0	0.0	1.0	1.0	1.5
T	0.0	0.5	1.0	1.75	2.0	2.5
2.0	2.5	3.0	3.0	3.0	4.0	4.0
3.0	3.75	4.25	4.5	5.0	5.5	7.0
5.0	6.0	6.0	6.0	7.0	7.5	8.0
7.0	8.25	8.5	9.0	9.75	10.25	10.75
8.0	8.5	9.0	10.0	10.0	10.0	11.0
11.0	11.5	12.0	13.5	14.75	15.0	15.0

In this case, multiple beats in the transcription correspond to the same beat in the ground truth, e.g., beat 1.0 in the ground truth corresponds to beats 1.75 and 2.0 in the transcription, because a single note in the ground truth has been transcribed as two tied notes. Only one alignment is shown in the figure for clarity.

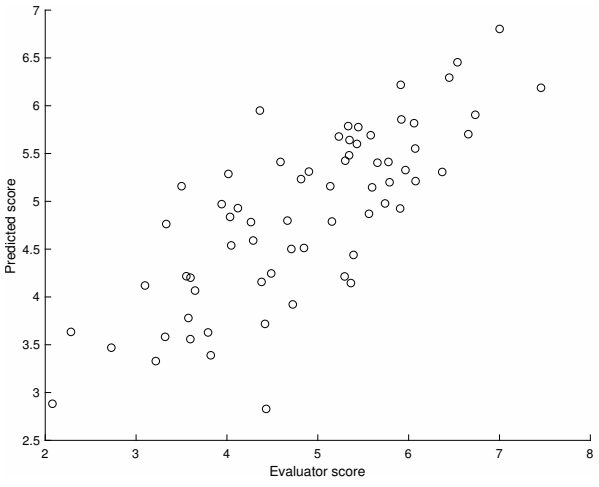
To calculate the distance between the two aligned scores, we proceed by first grouping all of the musical objects occurring inside aligned portions of the two scores into sets, thus losing the relative location of the objects within each set but preserving all of the other aspects, including staff assignment. Then the aligned sets are compared, and the differences between the two sets are reported separately. The following aspects only allow binary matching: barlines, clefs, key signatures, and time signatures. Rests are matched for duration and staff assignment, i.e., a rest with the correct duration but on the wrong staff will be considered a staff assignment error, a rest with the correct staff assignment but wrong duration will be considered a rest duration error. A missing or an extra rest will be considered a rest error. Notes are matched for spelling, duration, stem direction, staff assignment, and grouping into chords. For groupings, we only report the absolute value of the difference between the number of chords present in the two sets. The metric does not distinguish missing or



(a) Pitch Notation



(b) Rhythm Notation



(c) Note Positioning

Figure 4: Correlation between the predicted ratings and the average human evaluator ratings of all of the transcriptions in the dataset.

extra elements. These choices were dictated by simplicity of design and implementation.

All of the errors are cumulated for all of the matching sets. The errors for barlines, notes, note spelling, note durations, stem directions, groupings, rests, rest duration, and staff assignment are then normalized by dividing the total number of errors for each aspect by the total number of musical objects taken into account in the score. This step is necessary to normalize the number of errors for pieces of different lengths. The errors for clefs, key signatures, and time signatures are not normalized, as they are typically global aspects of the scores, and not influenced by the length of the piece. This might be a limitation for pieces with frequent changes in key signature or time signature.

As an example, the set of objects at the first beat of the first measure of Fig. 2 include the initial barlines, clefs, time signature, key signature, and notes starting on the downbeat of the measure. Barlines, clefs, time signature, and key signature are all correctly matched. All of the notes are correct in pitch, spelling and duration, however there are two errors in stem direction, one error in grouping, and one error in staff assignment. All of the rests are considered rest errors at each respective onsets.

For the first beat of the first measure of Fig. 3, all of the elements of the transcription till the first transcribed notes (the three notes pointed by the first arrow) and the notes tied to them will be considered as part of the same set. The wrong key signature and time signature will be reported as errors. The two eight rests will be reported as rest errors. The three notes in the transcription are properly spelled, but their duration is wrong, so that will be counted as three note duration errors. The missing D from the chord will be reported as a note error. The extra tied notes will be reported as note errors as well.

In summary, the following twelve normalized error counts are calculated by the metric: barlines, clefs, key signatures, time signatures, notes, note spelling, note durations, stem directions, groupings, rests, rest duration, and staff assignment. In order to translate these error counts into a musically relevant evaluation, we propose to use linear regression of the twelve error counts to fit human ratings of three musical aspects of automatic transcriptions, i.e., the pitch notation, the rhythm notation, and the note positioning. For each aspect, the linear regression learns twelve weights, one for each of the normalized error counts, to fit the human ratings. These weights can then be used to predict the human ratings of other music notation transcriptions.

4. EXPERIMENTAL RESULTS

To evaluate the proposed approach, we calculate the normalized error count and run linear regression to fit human ratings of 19 short music excerpts collected in our prior work [7]. These music excerpts were from the Kostka-Payne music theory book, all of them piano pieces by well-known composers, and were performed on a MIDI keyboard by a semi-professional piano player. These excerpts were then transcribed into music notation using four differ-

ent methods: a novel method proposed in the paper (which will be referred to as CDT), MuseScore, GarageBand and Finale. For each transcription, the human evaluators were asked to assign a numerical rating between 1 and 10 for three musical aspects, i.e., the pitch notation, the rhythm notation, and the note positioning.

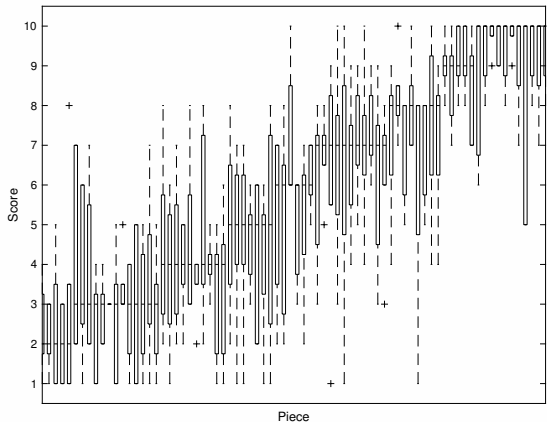
The proposed method of calculating the error counts uses MusicXML [10], the de facto standard for sharing sheet music files between applications, as the format of music notation. Two of the methods evaluated in the paper (Finale and MuseScore) can output the scores into MusicXML. For GarageBand, CDT and the ground truth, however, MusicXML was not available or was difficult to output automatically. We had to manually convert the scores into MusicXML. The transcribed scores are named with the initial of the transcription method and a number indicating the excerpt. So, `M-8.mxl` represents the eight excerpt transcribed with MuseScore. The letter K, for Kostka-Payne, indicates the ground truth scores. This dataset and a Python implementation of the proposed approach are available at `http://www.ece.rochester.edu/~acogliat/repository.html`. The implementation uses the music21 toolkit [9] for parsing the MusicXML files and processing the imported scores. The implementation has been tested with music21 V3.1.0.

In order to validate the quality of the prediction we calculated the coefficient of determination R^2, which is the square of the Pearson correlation coefficient. The R^2 was 0.558 for the pitch notation correlation, 0.534 for the rhythm notation, and 0.601 for note positioning. These results are reflected in Fig. 4; the proposed metric fits the data adequately, in general, even though the correlation is not perfect. It can also be noted that the prediction of the score for note positioning is the best, while the prediction of the score for rhythm notation is the worst.

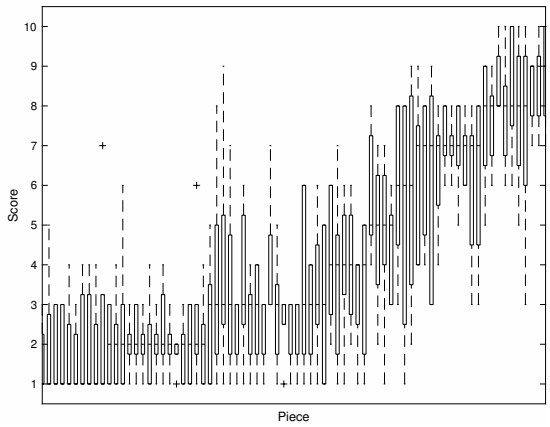
To understand the underlying causes of the covariance we firstly analyzed the ratings given by the human evaluators. As we can see from Fig. 5, the human evaluators were oftentimes in disagreement among themselves. It must also be noted that in our prior work [7], the human annotators were not given exact instructions on what features to consider for the evaluation, so a considerable amount of subjectivity and judgment calls were likely to be present in the ratings.

We also analyzed two transcriptions with the largest deviation from the predicted ratings, i.e., one transcription with a high predicted rating and a low human rating, and one transcription with a low predicted rating and a high human rating. The largest positive deviation occurred for the rhythm notation of transcription M-1, for which the proposed metric predicted a rating of 2.78, while the average human rating was 5.98. If we compare the transcription with the ground truth in Fig. 6 we can see that MuseScore misinterpreted the meter, causing the proposed metric to report a large number of note duration errors and barline errors, which resulted in a low rating. Human annotators, on the other side, likely penalized the meter error only once globally, but still considered the transcription acceptable overall.

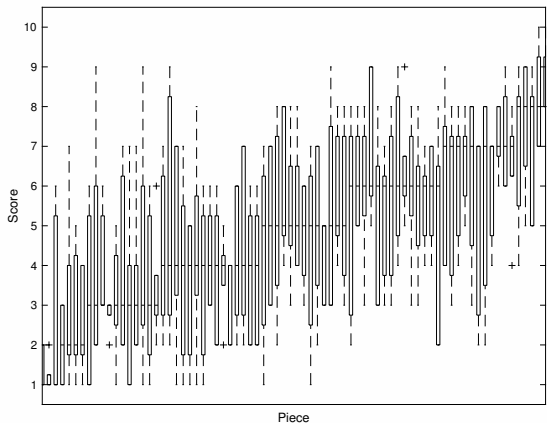
The largest negative deviation occurred for the pitch no-



(a) Pitch Notation

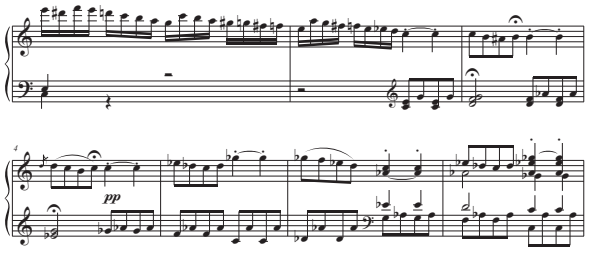


(b) Rhythm Notation

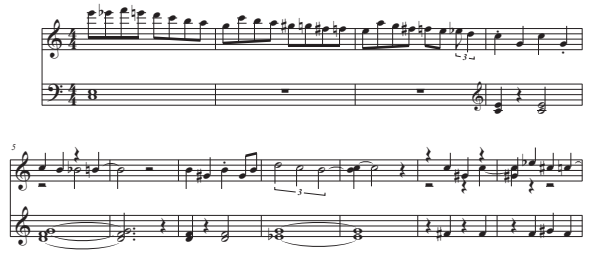


(c) Note Positioning

Figure 5: Distributions of the human ratings of the 76 transcriptions contained in the dataset. Each boxplot represents the ratings from 5 human evaluators.

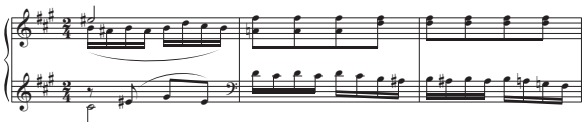


(a) Ground Truth

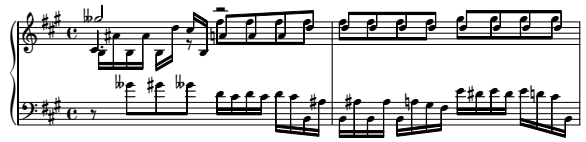


(b) M-1

Figure 6: Transcription of the first excerpt in the dataset by MuseScore, which shows the largest positive difference between the average human rating and the predicted rating, that is a high predicted rating and a low human rating. This evaluation difference occurs on the rhythm notation.



(a) Ground Truth



(b) C-13

Figure 7: Transcription of the thirteenth excerpt in the dataset by CDT, which shows the largest negative deviation between the average human rating and the predicted rating on rhythm notation, that is a low predicted rating and a high human rating. This evaluation difference occurs on the pitch notation.

tation of transcription C-13, for which the proposed metric predicted a rating of 6.83, while the annotators assigned an average score of of 4.48. If we compare the transcription with the ground truth in Fig. 7, we can notice that CDT makes a single mistake in notating the pitches, i.e., G♭♭ instead of E♯. It also makes a systematic error notating all Bs one octave lower. Finally, not grouping the eight notes in the treble staff makes the transcription hard to read. Possibly, the human annotators penalized the transcription because of its poor readability.

5. CONCLUSION AND FUTURE WORK

In this paper we proposed an objective metric to measure the differences between music notation transcriptions and the ground truth score. The metric is calculated by first aligning the pitch content of the transcription and the ground-truth music notation, and then counting the differences in twelve key musical aspects: barlines, clefs, key signatures, time signatures, notes, note spelling, note durations, stem directions, groupings, rests, rest duration, and staff assignment. We then used linear regression to predict human evaluator ratings along three aspects of music notation, namely, pitch notation, rhythm notation, and note positioning, from the error counts. Experiments show a clear correlation between the predicted ratings and the average human ratings, even though the correlation is not perfect.

One issue with the prediction is the high variance of the evaluator ratings, which likely originates from the inherent subjectivity of the tasks. Another issue of the proposed metric is that it does not incorporate music theory knowledge, such as the method proposed by Temperley to evaluate metrical models [19].

The current experiments were conducted on music notation transcriptions of human performances recorded on a MIDI keyboard; as a consequence, the transcriptions do not contain the errors commonly observed in audio-to-MIDI conversion processes, such as octave errors and extra or missing notes [5, 6]. More research is necessary to evaluate the performance of the proposed method in the presence of such errors. In addition, the excerpts in the dataset were very short, compared to real piano pieces, so additional research is necessary to assess the robustness of the metric, and its computational complexity on longer pieces.

A Python implementation of the proposed approach, along with the dataset, is available at `http://www.ece.rochester.edu/~acogliat/repository.html`. This implementation can be used to calculate the twelve error counts as well as to predict human ratings on the three musical aspects of a music notation transcription.

6. REFERENCES

[1] Mert Bay, Andreas F Ehmann, and J Stephen Downie. Evaluation of Multiple-F0 Estimation and Tracking Systems. In *Proc. of International Society for Music Information Retrieval (ISMIR)*, pages 315–320, 2009.

[2] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013.

[3] Emilios Cambouropoulos. From MIDI to traditional musical notation. In *Proc. of the AAAI Workshop on Artificial Intelligence and Music: Towards Formal Mod-*

els for Composition, Performance and Analysis, volume 30, 2000.

[4] Ali Taylan Cemgil. *Bayesian music transcription*. PhD thesis, Radboud University Nijmegen, 2004.

[5] Andrea Cogliati and Zhiyao Duan. Piano Music Transcription Modeling Note Temporal Evolution. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 429–433, Brisbane, Australia, 4 2015. IEEE.

[6] Andrea Cogliati, Zhiyao Duan, and Brendt Wohlberg. Piano Transcription with Convolutional Sparse Lateral Inhibition. *IEEE Signal Processing Letters*, 24(4):392–396, 2017.

[7] Andrea Cogliati, David Temperley, and Zhiyao Duan. Transcribing human piano performances into music notation. In *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2016.

[8] Tom Collins, Sebastian Böck, Florian Krebs, and Gerhard Widmer. Bridging the audio-symbolic gap: The discovery of repeated note content directly from polyphonic music audio. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, 2014.

[9] Michael Scott Cuthbert and Christopher Ariza. music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In *Proc. of International Society for Music Information Retrieval (ISMIR)*, 2010.

[10] Michael Good. MusicXML for notation and analysis. *The virtual score: representation, retrieval, restoration*, 12:113–124, 2001.

[11] Harald Grohganz, Michael Clausen, and Meinard Müller. Estimating Musical Time Information from Performed MIDI Files. In *Proc. of International Society for Music Information Retrieval (ISMIR)*, pages 35–40, 2014.

[12] Ioannis Karydis, Alexandros Nanopoulos, Apostolos Papadopoulos, Emilios Cambouropoulos, and Yannis Manolopoulos. Horizontal and vertical integration/segregation in auditory streaming: a voice separation algorithm for symbolic musical data. In *Proc. 4th Sound and Music Computing Conference (SMC2007)*, 2007.

[13] Jonathan M. Keith, editor. *Bioinformatics*, volume 1525 of *Methods in Molecular Biology*. Springer New York, New York, NY, 2017.

[14] Meinard Müller. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[15] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 3 2001.

[16] Kazuki Ochiai, Hirokazu Kameoka, and Shigeki Sagayama. Explicit beat structure modeling for non-negative matrix factorization-based multipitch analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 133–136, 2012.

[17] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 2 1978.

[18] Haruto Takeda, Naoki Saito, Tomoshi Otsuki, Mitsuru Nakai, Hiroshi Shimodaira, and Shigeki Sagayama. Hidden Markov model for automatic transcription of MIDI signals. In *Multimedia Signal Processing, 2002 IEEE Workshop on*, pages 428–431, 2002.

[19] David Temperley. An Evaluation System for Metrical Models. *Computer Music Journal*, 2004.

[20] David Temperley. *Music and probability*. The MIT Press, 2007.

[21] David Temperley. A unified probabilistic model for polyphonic music analysis. *Journal of New Music Research*, 38(1):3–18, 2009.