EVALUATING AUTOMATIC POLYPHONIC MUSIC TRANSCRIPTION

Andrew McLeod
University of Edinburgh
A.McLeod-5@sms.ed.ac.uk

Mark Steedman
University of Edinburgh
steedman@inf.ed.ac.uk

ABSTRACT

Automatic Music Transcription (AMT) is an important task in music information retrieval. Prior work has focused on multiple fundamental frequency estimation (multi-pitch detection), the conversion of an audio signal into a time-frequency representation such as a MIDI file. It is less common to annotate this output with musical features such as voicing information, metrical structure, and harmonic information, though these are important aspects of a complete transcription. Evaluation of these features is most often performed separately and independent of multi-pitch detection; however, these features are non-independent. We therefore introduce $MV2H$, a quantitative, automatic, joint evaluation metric based on musicological principles, and show its effectiveness through the use of specific examples. The metric is modularised in such a way that it can still be used with partially performed annotation—for example, when the transcription process has been applied to some transduced format such as MIDI (which may itself be the result of multi-pitch detection). The code for the evaluation metric described here is available at https://www.github.com/apmcleod/MV2H. [1]

1. INTRODUCTION

Automatic Music Transcription (AMT) involves converting an acoustic musical signal into some form of music notation. The process has generally been divided into two steps: first, multi-pitch detection, which is the conversion of the signal into a piano-roll notation (such as a MIDI file) by detecting which pitches are present at each time; and second, the conversion of that piano-roll notation into a musical score by annotating it with further musical information. Readers can refer to [2] for an overview of AMT.

The first step, multi-pitch detection, has been the focus of a great amount of research in AMT. The second step involves many subtasks of musical analysis, including voice separation, metrical alignment, note value detection, and harmonic analysis. Each of these has been the subject of research both directly from the acoustic signal and from other input formats such as MIDI. They are usually performed separately, though some recent work has attempted to analyse subsets of them jointly. For example, [28] estimates both chords and downbeats directly from acoustic input. [34] performs voice streaming, metrical alignment, and harmonic analysis jointly from symbolic input. However, even in the case of these joint models, evaluation is performed separately for each subtask. Rather than simply taking an average of a model's score on each subtask, there is a need for a standardised way to compute the joint score in a way that reflects overall AMT performance.

In this paper, we introduce $MV2H$ (from **M**ulti-pitch detection, **V**oice separation, **M**etrical alignment, note **V**alue detection, and **H**armonic analysis), a metric to quantitatively evaluate AMT systems that perform both multi-pitch detection and musical analysis. The metric can be used for AMT systems that do not perform all aspects of a full musical analysis—for example, those that perform multi-pitch detection and meter detection, but nothing else. One of the main principles of the new metric is that of disjoint penalties: that mistakes should only be penalised once. That is, if an error in one part of the transcription causes a mistake in another part, that error should not be counted twice. For example, if a pitch is missed during multi-pitch detection, the metric should not further penalise missing that note from the voice separation results.

Based on this principle, we do not include errors related to the proper typesetting of a transcription in our metric, and we do not even require a typeset musical score to perform our evaluation. Most typesetting decisions come down to the proper analysis of the underlying piece. For example, if metrical alignment is performed properly, beaming comes naturally. Likewise, stem directions can follow from voice separation and pitch spelling is a consequence of a proper harmonic analysis. For details related to the proper typesetting of music and its relation to the underlying music analysis, see [14].

2. EXISTING METRICS

Each of the separate tasks involved in the full AMT process has been the subject of much prior research, and there are existing metrics for each of them. This section gives a brief overview of the most widely used metrics for each subtask.

[1] Since the publication of this article, we have updated MV2H to handle non-time-aligned transcriptions (e.g., systems which output a musical score directly). The updated code is on github, and a technical report detailing the changes is available on arXiv [19].

Attribution: Andrew McLeod, Mark Steedman. "Evaluating Automatic Polyphonic Music Transcription", 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

2.1 Multi-pitch Detection

Multi-pitch detection is evaluated both at the frame level and at the note level depending whether a given model includes some form of note tracking or not. As the goal of this paper is to define a metric which is useful for a complete AMT system, the note-level evaluation metrics are most applicable here, and readers interested in the frame-based evaluation, an accuracy metric, should refer to [29].

For the note-level metric, a note is defined by its pitch, onset time, and offset time. [1] defines two different precision, recall, and F-measures for note-level multi-pitch detection. For the first, they define true positives as those notes detected whose pitch lies within a quartertone of that of a ground truth note, and whose onset time is within 50 ms of the same ground truth note's onset time, regardless of offset time. Spuriously detected notes are each assigned a false positive, and ground truth notes which are not matched by a detected note are each assigned a false negative. The second metric they propose is identical, with the additional constraint that a detected note's offset time must be accurate to within a certain threshold for it to be considered a true positive. Both of these metrics are used by both the Music Information Retrieval Evaluation Exchange (MIREX) [26] and the mir_eval package [30].

For our purposes, we care mostly about onset time and pitch (to the nearest semitone) as these aspects are most directly relevant to the underlying musical score. Offset time, on the other hand, is applicable as far as it relates to note value, and is discussed in Section 2.4. Our multi-pitch detection metric will therefore be based most closely on the first multi-pitch F-measure, which doesn't account for offset time.

2.2 Voice Separation

Voice separation refers to the separation of the notes of a piece of music into perceptual streams called voices. There is no standardised definition of what constitutes a voice, and a full discussion can be found in [3]. In this work, we restrict each voice to be monophonic. This aligns with the majority of work on voice separation, and is beneficial in AMT in that it allows simpler processing of monophonic data to occur in the later musical analysis steps.

There is no standardised metric for evaluating voice separation performance. [5] defines *Average Voice Consistency* (AVC), which returns an average of the percentage of notes in each voice which have been assigned to the correct voice. (A note is said to be assigned to the correct voice if its ground truth voice is the most common one for notes assigned to its voice.) This metric has a problem in that if a model assigns each note to a distinct voice, it achieves a perfect AVC of 100%. For acoustic input, [20, 31] use a similar metric, with the addition that spuriously detected notes automatically count as incorrect.

[17] defines two metrics: *soundness*, which measures the percentage of consecutive notes in an assigned voice which belong to the same ground truth voice; and *completeness*, which measures the percentage of consecutive notes in a ground truth voice which were assigned to the same voice. Finally, [12] defines a precision, recall, and F-measure evaluation, in which the problem of voice assignment is treated as a graph problem where each note is represented by a node, and two nodes are connected by an edge if and only if they are consecutive notes in an assigned voice. The values are calculated by counting the number of correct edges (true positives), spurious edges (false positives), and omitted edges (false negatives).

Each of these metrics would penalise an AMT system for any spurious notes, so for our proposed metric, we will need to use a modified version of one of them (or design a new metric) in order to enforce the principle of disjoint penalties.

2.3 Metrical Alignment

Metrical alignment is most often approached as one of three different tasks: downbeat tracking, beat tracking, or metrical structure detection. Downbeat tracking and beat tracking each involve identifying points in time, and thus can theoretically be evaluated using the same metrics, which are summarised by [8, 9]. F-measure [11] (which downbeat tracking work uses almost exclusively), is calculated by counting the number of (down)beats within some window length (usually 70 ms) of an annotated (down)beat. [4] proposes a similar metric, where accuracy is calculated instead using a Gaussian window around each annotated beat. [13] proposes a binary metric which is 1 if the beats are correctly tracked for at least 25% of the piece, and 0 otherwise. *P-score* [18], is the proportion of tracked beats which correctly match an annotated beat, normalised by either the number of tracked beats or the number of annotated beats (whichever is greater). Finally, [15] proposes metrics based on the longest continuously tracked section of music. All of the above are used to some extent in beat-tracking, and all are used by both mir_eval [30] and MIREX. [22, 24] In addition, evaluation is also often presented at twice and half the annotated beat length, to handle models which may track a beat at the wrong metrical level.

By comparison, the evaluation of metrical structure detection is far less sophisticated. Meter detection is the organisation of the beats of a given musical performance into a sequence of trees at the bar level, in which each node represents a single note value. The structure of each of these trees is directly related to the music's time signature, where the head of each tree splits into a number of nodes equal to the number of beats per bar, and each of these beat nodes splits into a number of nodes equal to the number of sub-beats per beat. Thus, each time signature uniquely describes a single metrical tree structure as defined by the number of beats per bar and sub-beats per beat in that time signature. The most basic evaluation is to simply report the proportion of musical excerpts for which the model guesses the correct metrical structure and phase (such that each tree aligns correctly with a single bar). Another approach is to simply report the proportion of musical excerpts for which the model correctly classifies the meter as duple or triple [10]. Both of these metrics are simplistic, and fail to take into account some idea of partially correct metrical structure trees.

Two metrics have been used for metrical structure detection evaluation which contain within them an evaluation of beat tracking and downbeat tracking, making them ideal for an evaluation of a joint model. [32] proposes a metric which takes into account the level on the metrical tree at which each note lies in order to capture some idea of partial correctness. However, since it is based on detected notes, it is not robust to errors in multi-pitch detection. [21] introduces an F-measure metric where each level of the detected metrical structure is assigned a true positive if it matches any level of the ground truth metrical structure (even if it is not the same level). A false positive is given for any level of the detected metrical structure which clashes with a metrical grouping in the ground truth, and a false negative for any metrical level in the ground truth which remains unmatched by a level of the detected metrical structure. As it is based solely on metrical groupings, rather than notes, it is robust to multi-pitch detection errors, and would not violate our principle of disjoint penalties. However, it was designed for use with metronomic input, and would therefore need to be adapted for our purposes of evaluating a complete AMT system on live performance data.

2.4 Note Value Detection

Note value detection (identifying a note as a quarter note, eighth note, dotted note, tied note, etc.) is not a widely researched problem, related to a combination of note offset time and metrical alignment. [27] describes two metrics for the task. One, error rate, is simply the percentage of notes whose transcribed value is incorrect. The other, scale error, takes into account the magnitude of the error as well (relative to the metrical grid), in log space such that errors from long notes do not dominate the calculation.

However, since the measured note values are reported relative to the underlying meter, they violate our property of disjoint penalties and we must design a new measure of note value detection accuracy for our metric.

2.5 Harmonic Analysis

Harmonic analysis involves both key detection, a classification problem of identifying one of twelve tonic notes, each with two possible modes (major or minor—alternate mode detection has not been widely researched); and chord tracking, identifying a sequence of chords and times given an audio recording. The possible chords to identify range from simply identifying the correct root note, to determining major or minor, identifying seventh chords, and even identifying different chord inversions.

The standard key detection evaluation, used by both mir_eval [30] and MIREX [25], is to assign a score of 1.0 to the correct key, 0.5 to a key which is a perfect fifth too high, 0.3 to the relative major or minor of the correct key, 0.2 to the parallel major or minor of the correct key, and 0.0 otherwise.

The standard chord tracking evaluation is *chord symbol recall* (CSR)—described by [16], and used by both MIREX [23], and mir_eval [30]—defined as the proportion of the input for which the annotated chord matches the ground truth chord. There can be varying levels of specificity for what exactly constitutes a match, since different sets of possible chords can be used as described above.

2.6 Joint Metric

For the joint evaluation of AMT performance, [7] presents a system to transcribe MIDI input into a musical score (thus including errors from typesetting), and evaluate it using five human evaluators. The evaluators were asked to: "1) Rate the pitch notation with regard to the key signature and the spelling of notes. 2) Rate the rhythmic notation with regard to the time signature, bar lines, and rhythmic values. 3) Rate the notation with regard to stems, voicing, and placement of notes on staves," each on a scale of 1 to 10. The three questions roughly correspond with four of our sections above: 1) harmonic analysis; 2) metrical alignment, note value detection; and 3) voice separation.

[6] describes an automatic metric for the same task, similar to string edit distance, taking into account the ordering of 12 different aspects of a musical score: barlines, clefs, key signatures, time signatures, notes, note spelling, note durations, stem directions, groupings, rests, rest duration, and staff assignment.

While this metric is a great step towards an automatic evaluation of AMT performance, it violates our principle of disjoint penalties. A single mistake in metrical alignment can manifest itself in the time signature, rest durations, note durations, and even additional notes (tied notes are counted as separate objects in the metric).

It appears that both of the above metrics measure something slightly different from what we want. They measure the readability of a score produced by an AMT system, while we really want a metric which measures the accuracy of the analysis performed by the AMT system, a slightly different task. To our knowledge, no metric exists which measures the accuracy of the analysis performed by a complete AMT system in the way we desire.

3. NEW METRIC

Our proposed metric, $MV2H$, draws from existing metrics where possible, though we take care to ensure that our principle of disjoint penalties is not violated. Essentially, we calculate a single score for each aspect of the transcription, and then combine them all into the final joint metric.

3.1 Multi-pitch Detection

For multi-pitch detection, we use an F-measure very similar to the one by [1] described above, counting a detected note as a true positive if its detected pitch (in semitones) is correct and its onset lies within 50 ms of the ground truth onset time. All other detected notes are false positives, and any unmatched ground truth notes are false negatives. Note offset time does not factor into our evaluation; rather, see Section 3.4 for a discussion on the related problem of note value detection.

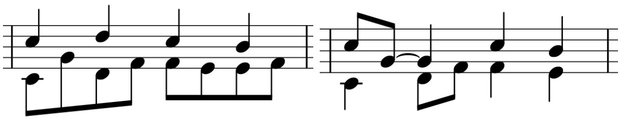


Figure 1: An example transcription of the ground truth bar (left) is shown (right). The voice connection between the last two notes in the lower voice counts as a true positive, even though they are not consecutive in the ground truth.

3.2 Voice Separation

For voice separation, we use an F-measure very similar to [12], taking care not to violate our principle of disjoint penalties. Specifically, we don't want to penalise any model in voice separation for multi-pitch detection errors.

Recall that the F-measure is calculated as a binary classification problem where for each ordered pair of notes, we must decide if they occur consecutively in the same voice or not. To address the disjoint penalties violation, we alter this slightly. We first remove from both the ground truth voices and the detected voices any notes which have not been matched as a true positive. Then, we perform the same F-measure calculation with the new voices.

As an illustration of this, see Figure 1. In the transcribed music, the last two notes in the lower voice are both matched with a ground truth note (in pitch and onset time), but are not immediately sequential in the ground truth voice. However, because the intervening note was not correctly transcribed, the link between these two notes counts as a true positive. (The second note in the transcribed lower voice does indeed count as an error.) This new F-measure calculation is equivalent to the standard voice separation F-measure when multi-pitch detection is performed perfectly.

3.3 Metrical Alignment

For metrical alignment, we would like to use a metric similar to that from [21] which has some idea of the partial correctness of a metrical alignment. However, as it is designed for use mainly on metronomic data where a metrical hypothesis cannot move in and out of phase throughout a piece, a few adjustments must be made to adapt it for use on live performance data. We call our newly designed evaluation metric the metrical F-measure. It takes into account every grouping at three levels of the metrical hierarchy throughout an entire piece: the sub beat level, the beat level, and the bar level.

For each hypothesised grouping at these metrical levels, we check if it matches a ground truth grouping at any level. A hypothesised grouping is said to match a ground truth grouping if its beginning and ending times are each within 50 ms of the beginning and ending times of that particular ground truth grouping, regardless of the metrical level of either grouping.[2] Each matched pair of groupings within a piece count as a true positive, while any unmatched hypothesis groupings count as false positives, and any unmatched ground truth groupings count as false negatives. The metrical F-measure of a piece is then calculated as the harmonic mean of precision and recall as usual.

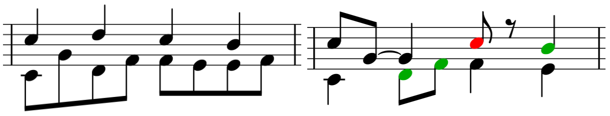


Figure 2: An example transcription of the ground truth bar (left) is shown (right). Those notes which are assigned a note value score are coloured. Of those, the C (assuming treble clef) is assigned a score of 0.5, while the others are assigned a score of 1.

3.4 Note Value Detection

It is difficult to disentangle note value detection from multi-pitch detection, voice separation, and metrical alignment in order to include it in our evaluation without violating our principle of disjoint penalties. Clearly, note value should only be regarded if the note has been counted as a true positive in the multi-pitch detection evaluation. Less obviously, we also disregard any detected note which is not followed in its transcribed voice by the correct note. Additionally, note value depends directly on meter such that any note value accuracy metric must measure note value relative to time rather than the underlying metrical grid.

Therefore, we define a note value score which measures only a subset of the detected notes: those which both (1) correspond with a true positive multi-pitch detection; and (2) correspond with a true positive ground truth voice segment as described in the previous paragraph. Each note which matches those two criteria is assigned a score according to the accuracy of its normalised duration (that is, the duration corresponding to its note value rather than its performed duration). Specifically, each note is counted as correct and assigned a score of 1 if its normalised duration is within 100 ms of the normalised duration of the corresponding ground truth note.[3] Otherwise, its score is calculated as in Equation 1, where dur_{gt} is the ground truth note's normalised duration and dur_{det} is the detected note's normalised duration. This score is 1 when the durations match exactly and scales linearly on both sides to a score of 0 for a note with 0 duration or a note with twice the ground truth note's duration. The overall note value score is calculated as the arithmetic mean of the scores of those notes which are assigned a score.

$$score = max\Big(0, 1 - \frac{|dur_{gt} - dur_{det}|}{dur_{gt}}\Big) \qquad (1)$$

Figure 2 illustrates this note value score. Only those notes which are coloured are considered for the note value

[2] We use a 50 ms threshold, rather than the more common 70 ms, because it was shown by [8] that 50 ms corresponds more exactly with human judgement for beat tracking. However, this threshold may need to be tuned for different genres as regular syncopation can tend to misalign notes with the metrical grid in certain genres more than others [33].

[3] We use 100 ms here to allow for a 50 ms error in both onset and offset time, although this value again may need to be tuned for different genres.

score. Notice that the two C's (assuming treble clef) on the downbeat are not considered due to errors in voice separation. Likewise, the last two notes in the lower voice are also not counted against note value score due to note detection errors, even though they count as true positives for the voice separation F-measure. Of the coloured notes, the C would be assigned a score of around 0.5 (depending on exact timing), since its value duration is off by exactly half of the ground truth note's value duration. The others would receive scores of 1. Thus, the final note value score would be the average of 1, 1, 1, and 0.5, or about 0.875.

3.5 Harmonic Analysis

For harmonic analysis, we use the standard key detection and CSR metrics described above, as neither one violates our principle of disjoint penalties since they are based on time rather than notes or the metrical alignment. For now, we take the set of possible chords to include a major and minor version for each root note, but not sevenths or inversions, although the full collection of chords should be used for the final version of our metric.

To combine the two into a single harmonic analysis score, we take the arithmetic mean of the two values, since they are both on the range [0–1]. Models which only perform one of the above tasks may simply use that task's score as their harmonic analysis score.

3.6 Joint Metric

We now have five values to combine into a single number: the multi-pitch detection F-measure, the voice separation F-measure, the metrical F-measure, the note value detection accuracy score, and the harmonic analysis mean. All of these values are on the range [0–1] such that a value of 1 results from a perfect transcription in that aspect. We consider three different approaches for their combination: harmonic mean, geometric mean, and arithmetic mean.

Harmonic mean is most useful when there is potential for one of the values involved to be significantly larger than the others, and thus dominate the overall result. F-measure, for example, is the harmonic mean between precision and recall, and is used so that models cannot receive a high F-measure by simply tuning their model to have a very high recall or precision; rather, both recall and precision must be relatively high in order for their harmonic mean to also be high. This is not relevant in our case as there is no way for a model to tune itself towards one very high score at the expense of the others as is the case with some binary classification problems.

Geometric mean is most useful when the values involved are on different scales. Then, a given percent change in one of the values will result in the same change in mean as the same percent change to another of the values. This property is not necessary for us because all of our values lie on the same range.

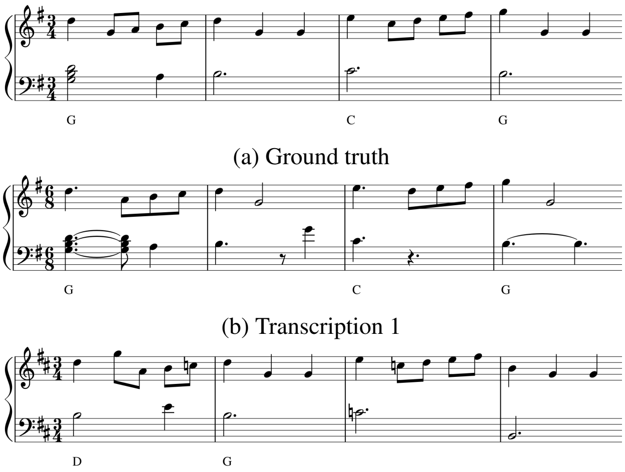
Arithmetic mean is a simple calculation that weights each of the values involved equally. This property is desirable for us because, for a complete transcription, all five aspects of an analysis must be correct. Furthermore, due to our property of disjoint penalties, we have kept the five values involved disjoint, and a model must fairly perform well on all aspects in order for its overall score to be high.

Therefore, for the final joint metric, $MV2H$ (for **M**ulti-pitch detection, **V**oice separation, **M**etrical alignment, note **V**alue detection, and **H**armonic analysis), we take the arithmetic mean of the five previously calculated values. We also want the metric to be usable no matter what subset of analyses is performed, for example, for models which run on MIDI input and therefore do not perform multi-pitch detection. In these cases, we advise using our metric and simply taking the arithmetic mean of only those scores which correspond with analyses performed. In future work, we will investigate whether a linear combination of the five values involved, perhaps weighting some more strongly than others, aligns more exactly with human judgements than the current arithmetic mean.

4. EXAMPLES

To illustrate the effectiveness and appropriateness of our metric, we present in Figure 3 two example transcriptions of the first four bars of Bach's Minuet in G, each exhibiting different errors. Figure 3a shows the ground truth transcription (where the chord progression is shown beneath the staff), and the example transcriptions are shown below. We make two assumptions: (1) ground truth voices are separated by clef (plus the bottom two notes in the initial chord, which each belong to their own voice); and (2) The sub beats of each transcription align in time with the sub beats of the ground truth.

Figure 3b shows an example transcription which is good in general, with just a few mistakes, mostly related to the metrical alignment. First, for the multi-pitch detection F-measure, we can see that the transcription has 20 true positives, 3 false negatives (a G on the second beat in the first bar, a C on the second beat of the third bar, and the final G in the fourth bar), and 0 false positives, resulting in an F-measure of 0.93. For voice separation, this transcription is generally good, making a single bad assignment in the second bar, resulting in 3 false positives (the connections to and from the incorrect assignment, as well as the incorrect connection in the treble clef), 3 false negatives (the correct connections to and from the misclassified note, as well as the correct connection in the bass clef), and a voice separation F-measure of 0.83. Notice that the missed G in the upper voice in the treble clef of the first bar does not result in a penalty for voice assignment due to our principle of disjoint penalties. For metrical alignment, we can see that this transcription is notated in $\frac{6}{8}$ time, correctly grouping all sub beats (eighth notes) and bars, yielding 28 true positives, but incorrectly grouping three sub beats together into dotted quarter note beats, yielding 8 false positives and 12 false negatives. This results in a metrical F-measure of 0.74. For note value detection, 14 notes are counted: all of the bass clef notes and all of the eighth notes in the first bar, only the high D in the second bar, the low C and all of the eighth notes in the third bar, and the high G and the low B in the fourth bar. Notice that the initial high D isn't



(a) Ground truth

(b) Transcription 1

(c) Transcription 2

Figure 3: Two different example transcriptions of the first four bars of Bach's Minuet in G.

Transcription	1	2
Multi-pitch	**0.93**	0.77
Voice	0.83	**1.0**
Meter	0.74	**1.0**
Note Value	0.96	**1.0**
Harmonic	**1.0**	0.5
$MV2H$	**0.89**	0.85

Table 1: The resulting evaluation scores from each of the example transcriptions from Figure 3.

counted because the next note in its voice has not been detected. Similarly, neither the G on the second beat of the second bar nor any of the bass clef notes in the second bar are counted due to voice separation errors. Of the 14 notes, 13 of them are assigned the correct note value (even the first bass chord, since its incorrect typesetting and the ties are related to the incorrect metrical alignment—the note value still ends at the correct point in time). One note (the C in the bass clef on the downbeat of the third bar) is assigned a value score of 0.5 (since its value duration is half of the correct value duration). This results in a note value detection score of 0.96. The harmonic analysis in this transcription is entirely correct, resulting in a harmonic score of 1.0. Thus, the $MV2H$ of the first transcription is 0.89. This makes sense because the transcription is quite good in general, but a few mistakes are made, the most glaring of which is the metrical alignment (its lowest score).

Figure 3c shows another example transcription which is again good in general, this time with a few more errors in multi-pitch detection, as well as a poor harmonic analysis. For multi-pitch detection, it contains 17 true positives, 4 false positives, and 6 false negatives, resulting in an F-measure of 0.77. This number is 0.16 lower than that the previous transcription's corresponding F-measure, and this makes sense intuitively: the first transcription does seem to have resulted from a more accurate multi-pitch detection than the second. For voice separation, this second transcription contains no errors. Some erroneous notes are placed into one voice or the other, but all of the correctly detected notes are also correctly separated into voices, resulting in a perfect voice separation F-measure of 1.0. Likewise the metrical alignment is performed perfectly, resulting in a metrical F-measure of 1.0. For note value detection, we look at all of the true positive note detections except (1) the initial D on the downbeat of the first bar, (2) the B in the bass clef of the first bar, (3) the C in the bass clef of the third bar, and (4) the high F at the end of the third bar. (All of these exceptions are due to missed note detections of the following note in each voice.) All of the remaining notes have been assigned the correct value, resulting in a note value detection score of 1.0. For the harmonic analysis, the model has incorrectly transcribed the excerpt in D major, resulting in a key score of 0.5. Likewise, the model has incorrectly labelled the chord progression as D-G-G-G, rather than G-G-C-G. Thus, it has transcribed the correct chord for half of the transcription, resulting in a CSR of 0.5, and a harmonic score of 0.5. The $MV2H$ of the second transcription is therefore 0.85: slightly worse than the first transcription, but still good.

The scores of both transcriptions are summarised in Table 1, and intuitively, they make sense. Both seem good overall, though they both contain errors. The first transcription has an incorrectly notated meter (although its bars and sub beats still align correctly) and a few other smaller mistakes related to multi-pitch detection, voice separation, and note value detection. The second transcription, on the other hand, correctly aligns the meter, and makes its only errors in its harmonic analysis (which is quite poor), and in multi-pitch detection (it is worse than the first model in this regard). Given these examples, for applications which need a good all-around transcription, we would recommend the system which produced the first transcription. However, applications which emphasise metrical structure detection or voice separation should consider using the system which produced the second transcription instead.

5. CONCLUSION

As research moves towards the goal of a complete AMT system, an automatic, standardised, quantitative metric for the task will become a necessity. To that end, we have proposed a joint metric, $MV2H$, which measures multi-pitch detection, voice separation, metrical alignment, note value detection, and harmonic analysis and summarises them in a single number. Our metric is based on the property of disjoint penalties: that a model should not be penalised twice for errors which come from a single mistake or misinterpretation. While our metric may not be the final standardised metric used for the task, we believe that it should become part of the discussion, and that the principles that guided us through its creation should continue to be addressed by any future proposed metrics.

Future work will evaluate our metric on a wider corpus of realistic transcriptions. In particular, we will investigate how well our metric aligns with human judgements, testing a weighted average of the five values involved, rather than using the arithmetic mean. A more advanced multi-pitch detection metric, for example one which weights errors according to their perceptual salience, could be another avenue for improvements.

6. ACKNOWLEDGEMENTS

This work was partially funded by EU ERC H2020 Advanced Fellowship GA 742137 SEMANTAX and a Google faculty award.

7. REFERENCES

[1] Mert Bay, Andreas F. Ehmann, and J. Stephen Downie. Evaluation of Multiple-f0 Estimation and Tracking Systems. In *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, July 2013.

[3] Emilios Cambouropoulos. Voice And Stream: Perceptual And Computational Modeling Of Voice Separation. *Music Perception*, 26(1):75–94, September 2008.

[4] A. Taylan Cemgil, Bert Kappen, Peter Desain, and Henkjan Honing. On tempo tracking: Tempogram Representation and Kalman filtering. *Journal of New Music Research*, 29(4):259–273, December 2000.

[5] Elaine Chew and Xiaodan Wu. Separating Voices in Polyphonic Music: A Contig Mapping Approach. In *Computer Music Modeling and Retrieval*, pages 1–20, 2004.

[6] Andrea Cogliati and Zhiyao Duan. A metric for music notation transcription accuracy. In *ISMIR*, pages 407–413, 2017.

[7] Andrea Cogliati, David Temperley, and Zhiyao Duan. Transcribing Human Piano Performances into Music Notation. In *ISMIR*, pages 758–764, 2016.

[8] Matthew E. P. Davies and Sebastian Böck. Evaluating the Evaluation Measures for Beat Tracking. In *ISMIR*, pages 637–642, 2014.

[9] Matthew E. P. Davies, Norberto Degara, and Mark D. Plumbley. Evaluation Methods for Musical Audio Beat Tracking Algorithms. *Queen Mary University of London, Centre for Digital Music, Technical Report C4DM-TR-09-06*, 2009.

[10] W. Bas De Haas and Anja Volk. Meter Detection in Symbolic Music Using Inner Metric Analysis. In *ISMIR*, pages 441–447, 2016.

[11] Simon Dixon. Evaluation of the Audio Beat Tracking System BeatRoot. *Journal of New Music Research*, 36(1):39–50, March 2007.

[12] Ben Duane and Bryan Pardo. Streaming from MIDI using constraint satisfaction optimization and sequence alignment. In *Proceedings of the International Computer Music Conference*, pages 1–8, 2009.

[13] Masataka Goto and Yoichi Muraoka. Issues in Evaluating Beat Tracking Systems. In *Workshop on Issues in AI and Music*, pages 9–16, 1997.

[14] Elaine Gould. *Behind bars : the definitive guide to music notation*. Faber Music, 2011.

[15] Stephen W. Hainsworth. *Techniques for the Automated Analysis of Musical Audio*. PhD thesis, University of Cambridge, 2003.

[16] Christopher Harte. *Towards automatic extraction of harmony information from music signals*. PhD thesis, Queen Mary University of London, 2010.

[17] Phillip Kirlin and Paul Utgoff. VOISE: Learning to Segregate Voices in Explicit and Implicit Polyphony. In *ISMIR*, pages 552–557, 2005.

[18] M. F. McKinney, D. Moelants, Matthew E. P. Davies, and Anssi Klapuri. Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms. *Journal of New Music Research*, 36(1):1–16, March 2007.

[19] Andrew McLeod. Evaluating non-aligned musical score transcriptions with MV2H. In *arXiv:1906.00566*, June 2019.

[20] Andrew McLeod, Rodrigo Schramm, Mark Steedman, and Emmanouil Benetos. Automatic transcription of polyphonic vocal music. *Applied Sciences*, 7(12), 2017.

[21] Andrew McLeod and Mark Steedman. Meter detection in symbolic music using a lexicalized pcfg. In *Proceedings of the 14th Sound and Music Computing Conference*, pages 373–379, 2017.

[22] MIREX. Audio beat tracking. `http://www.music-ir.org/mirex/wiki/2017:Audio\_Beat\_Tracking`, 2017. Accessed: 2017-07-18.

[23] MIREX. Audio chord estimation. `http://www.music-ir.org/mirex/wiki/2017:Audio\_Chord\_Estimation`, 2017. Accessed: 2017-07-18.

[24] MIREX. Audio downbeat estimation. `http://www.music-ir.org/mirex/wiki/2017:Audio\_Downbeat\_Estimation`, 2017. Accessed: 2017-07-18.

[25] MIREX. Audio key detection. `http://www.music-ir.org/mirex/wiki/2017:Audio\_Key\_Detection`, 2017. Accessed: 2017-07-18.

[26] MIREX. Multiple fundamental frequency estimation & tracking. `http://www.music-ir.org/mirex/wiki/2017:Multiple\_Fundamental\_Frequency\_Estimation\_\%26\_Tracking`, 2017. Accessed: 2017-07-18.

[27] Eita Nakamura, Kazuyoshi Yoshii, and Simon Dixon. Note value recognition for piano transcription using markov random fields. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9):1846–1858, sep 2017.

[28] Hélène Papadopoulos and Geoffroy Peeters. Joint Estimation of Chords and Downbeats From an Audio Signal. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):138–152, January 2011.

[29] Graham E Poliner and Daniel P. W. Ellis. A Discriminative Model for Polyphonic Piano Transcription. *EURASIP Journal on Advances in Signal Processing*, 2007(1):154–162, 2007.

[30] Colin Raffel, Brian Mcfee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel P. W. Ellis, C Colin Raffel, Brian Mcfee, and Eric J. Humphrey. mir_eval: A Transparent Implementation of Common MIR Metrics. In *ISMIR*, 2014.

[31] Rodrigo Schramm, Andrew McLeod, Mark Steedman, and Emmanouil Benetos. Multi-pitch detection and voice assignment for a cappella recordings of multiple singers. In *ISMIR*, pages 552–559, Suzhou, October 2017.

[32] David Temperley. An Evaluation System for Metrical Models. *Computer Music Journal*, 28(3):28–44, September 2004.

[33] David Temperley. Communicative pressure and the evolution of musical styles. *Music Perception*, 21:313–337, 2004.

[34] David Temperley. A Unified Probabilistic Model for Polyphonic Music Analysis. *Journal of New Music Research*, 38(1):3–18, March 2009.