IMPROVING RHYTHMIC TRANSCRIPTIONS VIA PROBABILITY MODELS APPLIED POST-OMR

Maura Church
Applied Math, Harvard University
and Google Inc.
maura.church@gmail.com

Michael Scott Cuthbert
Music and Theater Arts
M.I.T.
cuthbert@mit.edu

ABSTRACT

Despite many improvements in the recognition of graphical elements, even the best implementations of Optical Music Recognition (OMR) introduce inaccuracies in the resultant score. These errors, particularly rhythmic errors, are time consuming to fix. Most musical compositions repeat rhythms between parts and at various places throughout the score. Information about rhythmic self-similarity, however, has not previously been used in OMR systems.

This paper describes and implements methods for using the prior probabilities for rhythmic similarities in scores produced by a commercial OMR system to correct rhythmic errors which cause a contradiction between the notes of a measure and the underlying time signature. Comparing the OMR output and post-correction results to hand-encoded scores of 37 polyphonic pieces and movements (mostly drawn from the classical repertory), the system reduces incorrect rhythms by an average of 19% (min: 2%, max: 36%).

The paper includes a public release of an implementation of the model in `music21` and also suggests future refinements and applications to pitch correction that could further improve the accuracy of OMR systems.

1. INTRODUCTION

Millions of paper copies of musical scores are found in libraries and archival collections and hundreds of thousands of scores have already been scanned as PDFs in repositories such as IMSLP [5]. A scan of a score cannot, however, be searched or manipulated musically, so Optical Music Recognition (OMR) software is necessary to transform an image of a score into symbolic formats (see [7] for a recent synthesis of relevant work and extensive bibliography; only the most relevant citations from this work are included here). Projects such as Peachnote [10] show both the feasibility of recognizing large bodies of scores and also the limitations that errors introduce, particularly in searches such as chord progressions that rely on accurate recognition of multiple musical staves.

Understandably, the bulk of OMR research has focused on improving the algorithms for recognizing graphical primitives and converting them to musical objects based on their relationships on the staves. Improving score accuracy using musical knowledge (models of tonality, meter, form) has largely been relegated to "future work" sections and when discussed has focused on localized structures such as beams and measures and requires access to the "guts" of a recognition engine (see Section 6.2.2 in [9]). Improvements to score accuracy based on the output of OMR systems using multiple OMR engines have been suggested [2] and when implemented yielded results that were more accurate than individual OMR engines, though the results were not statistically significant compared to the best commercial systems [1]. Improving the accuracy of an OMR score using musical knowledge and a single engine's output alone remains an open field.

This paper proposes using rhythmic repetition and similarity within a score to create a model where measure-level metrical errors can be fixed using correctly recognized (or at least metrically consistent) measures found in other places in the same score, creating a self-healing method for post-OMR processing conditioned on probabilities based on rhythmic similarity and statistics of symbolic misidentification.

2. PRIOR PROBABILITIES OF DISTANCE

Most Western musical scores, excepting those in certain post-common practice styles (e.g., Boulez, Cage), use and gain cohesion through a limited rhythmic vocabulary across measures. Rhythms are often repeated immediately or after a fixed distance (e.g., after a 2, 4, or 8 measure distance). In a multipart score, different instruments often employ the same rhythms in a measure or throughout a passage. From a parsed musical score, it is not difficult to construct a hash of the sequence of durations in each measure of each part (hereafter simply called "measure"; "measure stack" will refer to measures sounding together across all parts); if grace notes are handled separately, and interior voices are flattened (e.g., using the music21 `chordify` method) then hash-key collisions will only occur in the rare cases where two graphically distinct symbols equate to the same length in quarter notes (such as a dotted-triplet eighth note and a normal eighth).

Within each part, the prior probability that a measure m_0 will have the same rhythm as the measure n bars later (or earlier) can be computed (the prior-based-on-distance, or PrD). Similarly, the prior probability that, within a measure stack, part p will have the same rhythm as part q can also be computed (the prior-based-on-part, or PrP).

Figure 1 shows these two priors for the violin I and viola parts of the first movement of Mozart K525 (*Eine kleine Nachtmusik*). Individual parts have their own characteristic shapes; for instance, the melodic violin I (*top left*), shows less rhythmic similarity overall than the viola (*bot. left*). This difference results from the greater rhythmic variety of the violin I part compared to the viola part. Moments of large-scale repetition such as between the exposition and recapitulation, however, are easily visible as spikes in the PrD graph for violin I. (Possible refinements to the model taking into account localized similarities are given at the end of this paper.) The PrP graphs (*right*) show that both parts are more similar to the violoncello part than to any other part. However, the viola is more similar to the cello (and to violin II) that violin I is to any other part.

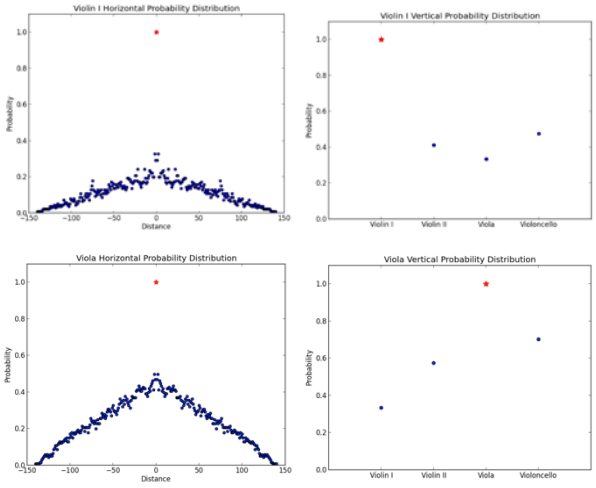

Figure 1. Priors based on distance (*l.* in measure separation) and part (*r.*) for the violin I (*top*) and viola (*bot.*) parts in Mozart, K525.

3. PRIOR PROBABILITIES OF CHANGE

3.1 Individual Change Probabilities

The probability that any given musical glyph will be read correctly or incorrectly is dependent on the quality of scan, the quality of original print, the OMR engine used, and the type of repertory. One possible generalization used in the literature [8] is to classify errors as class confusion (e.g., rest for note, with probability of occurring c), omissions (e.g., of whole symbols or of dots, tuplet marks: probability o), additions (a), and general value confusion (e.g., quarter for eighth: v). Other errors, such as sharp for natural or tie for slur, do not affect rhythmic accuracy. Although accuracy would be improved by computing these values independently for each OMR system and quality of scan, such work is beyond the scope of the current paper. Therefore, we use Rossant and Bloch's recognition rates, adjusting them for the differences between working with individual symbols (such as dots and note stems) and symbolic objects (such as dotted-eighth and quarter notes). The values used in this model are thus: $c = .003$, $o = .009$, $a = .004$, $v = .016$.[1] As will become clear, more accurate measures would only improve the results given below. Subtracting these probabilities from 1.0, the rate of equality, e, is .968.

3.2 Aggregate Change Distances

The similarity of two measures can be calculated in a number of different ways, including the earth mover distance, the Hamming distance, and the minimum Levenshtein or edit distance. The nature of the change probabilities obtained from Rossant and Bloch along with the inherent difficulties of finding the one-to-one correspondence of input and output objects required for other methods, made Levenshtein distance the most feasible method. The probability that certain changes would occur in a given originally scanned measure (source, S) to transform it into the OMR output measure (destination, D) is determined by finding, through an implementation of edit distance, values for $i, j, k, l,$ and m (for number of class changes, omissions, additions, value changes, and unchanged elements) that maximize:

$$p_{S,D} = c^i \cdot o^j \cdot a^k \cdot v^l \cdot e^m \qquad (1)$$

Equation (1), the prior-based-on-changes or PrC, can be used to derive a probability of rhythmic change due to OMR errors between any two arbitrary measures, but the model employed here concerns itself with measures with incorrect rhythms, or flagged measures.

3.3 Flagged Measures

Let F_{Pi} be the set of flagged measures for part P_i, that is, measures whose total durations do not correspond to the total duration implied by the currently active time signature, and $F = \{F_{P1}, \ldots, F_{Pj}\}$ for a score with j parts. (Measure stacks where each measure number is in F can be removed as probable pickup or otherwise intended incomplete measures, and long stretches of measures in F in all parts can be attributed to incorrectly identified time signatures and reevaluated, though neither of these refinements is used in this model). It is possible for rhythms within a measure to be incorrectly recognized without the entire measure being in F; though this problem only arises in the rare case where two rhythmic errors cancel out each other (as in a dotted quarter read as a quarter with an eighth read as a quarter in the same measure).

[1] Rossant and Bloch give probabilities of change given that an error has occurred. The numbers given here are renormalizations of those error rates after removing the prior probability that an error has taken place.

4. INTEGRATING THE PRIORS

For each $m \in F_{Pi}$, the measure n in part P_i with the highest likelihood of representing the prototype source rhythm before OMR errors were introduced is the source measure S_D that maximizes the product of the prior-based-on-distance, that is, the horizontal model, and the prior-based-on-changes:

$$S_D = argmax(\mathrm{PrD}_n \bullet \mathrm{PrC}_n) \; \forall \; n \notin \mathrm{F}. \quad (2)$$

(In the highly unlikely case of equal probabilities, a single measure is chosen arbitrarily) Similarly, for each m in F_P the measure t in the measure stack corresponding to m, with the highest likelihood of being the source rhythm for m, is the source measure S_P that maximizes the product of the prior-based-on-part, that is, the vertical model, and the prior-based-on-changes:

$$S_P = argmax(\mathrm{PrP}_t \bullet \mathrm{PrC}_t) \; \forall \; t \notin \mathrm{F}. \quad (3)$$

Since the two priors PrD and PrP have not been normalized in any way, the best match from S_D and S_P can be obtained by simply taking the maximum of the two:

$$S = argmax(P(m)) \; \forall \; m \text{ in } [S_D, S_P] \quad (4)$$

Given the assumption that the time signature and barlines have accurately been obtained and that each measure originally contained notes and rests whose total durations matched the underlying meter, we do not need to be concerned with whether S is a "better" solution for correcting m than the rhythms currently in m, since the probability of a flagged measure being correct is zero. Thus any solution has a higher likelihood of being correct than what was already there. (Real-world implementations, however, may wish to place a lower bound on $P(S)$ to avoid substitutions that are below a minimum threshold to prevent errors being added that would be harder to fix than the original.)

5. EXAMPLE

In this example from Mozart K525, mvmt. 1, measure stack 17, measures in both Violin I and Violin II have been flagged as containing rhythmic errors (marked in purple in Figure 2).

Both the OMR software and our implementation of the method, described below, can identify the violin lines as containing rhythmic errors, but neither can know that an added dot in each part has caused the error. The vertical model (PrP * PrC) will look to the viola and cello parts for corrections to the violin parts. Violin II and viola share five rhythms (e^5) and only one omission of a dot is required to transform the viola rhythm into violin II (o^1), for a PrC of 0.0076. The prior on similarities between violin II and viola (PrP) is 0.57, so the complete probability of this transformation is 0.0043. The prior on similarities between violin II and cello is slightly higher, 0.64, but the prior based on changes is much smaller ($4 \cdot 10^{-9}$). Violin I is not considered as a source since its measure has also been flagged as incorrect. Therefore the viola's measure is used for S_P.

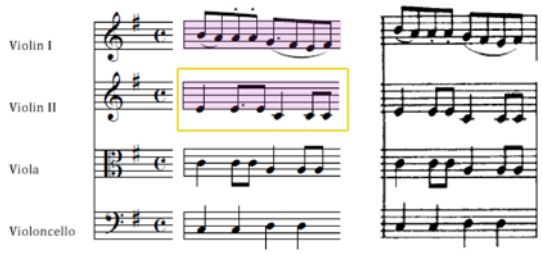

Figure 2. Mozart, K525 I, in OMR (*l.*) and scanned (*r.*) versions.

A similar search is done for the other (unflagged) measures in the rest of the violin II part in order to find S_D. In this case, the probability of S_P exceeds that of S_D, so the viola measure's rhythm is, correctly, used for violin II.

6. IMPLEMENTATION

The model developed above was implemented using conversion and score manipulation routines from the open-source Python-based toolkit, `music21` [4] and has been contributed back to the toolkit as the `omr.correctors` module in v.1.9 and above. Example 1 demonstrates a round-trip in MusicXML of a raw OMR score to a post-processed score.

```
from music21 import *
s = converter.parse('/tmp/k525omrIn.xml')
sc = omr.correctors.ScoreCorrector(s)
s2 = sc.run()
s2.write('xml', fp='/tmp/k525post.xml')
```

Example 1. Python/`music21` code for correcting OMR errors in Mozart K525, I.

Figure 3, below, shows the types of errors that the model is able, and in some cases unable, to correct.

7. RESULTS

Nine scores of four-movement quartets by Mozart (5),[1] Haydn (1), and Beethoven (4) were used for the primary evaluation. (Mozart K525, mvmt. 1 was used as a test score for development and testing but not for evaluation.) Scanned scores came from out-of-copyright editions (mainly Breitkopf & Härtel) via IMSLP and were converted to MusicXML using SmartScore X2 Pro (v.10.5.5). Ground truth encodings in MuseData and MusicXML formats came via the music21 corpus originally from the Stanford's CCARH repertories [6] and Project Gutenberg.

[1] Mozart K156 is a three-movement quartet, however, both the ground truth and the OMR versions include the abandoned first version of the Adagio as a fourth movement.

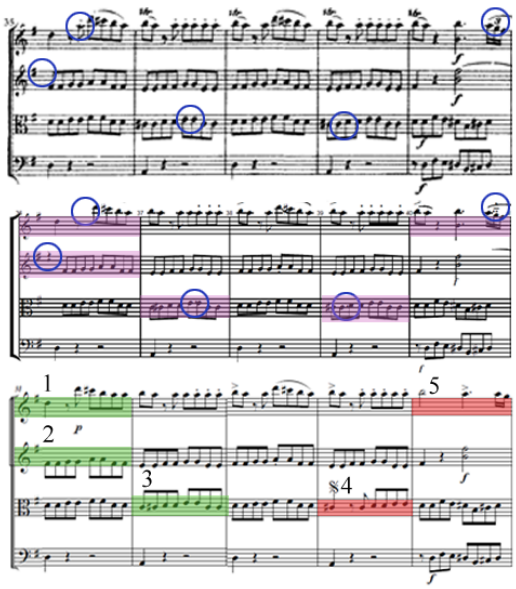

Figure 3: Comparison of Mozart K525 I, mm. 35–39 in the original scan *(top)*, SmartScore OMR output (*middle*), and after post-OMR processing (*bot.*). Flags 1–3 were corrected successfully; Flags 4 and 5 result in metrically plausible but incorrect emendations. The model was able to preserve the correct pitches for Flags 2 (added quarter rest) and Flag 3 (added augmentation dot). Flag 1 (omitted eighth note) is considered correct in this evaluation, based solely on rhythm, even though the pitch of the reconstructed eighth note is not correct.

The pre-processed OMR movement was aligned with the ground truth by finding the minimum edit distance between measure hashes. This step was necessary for the many cases where the OMR version contained a different number of measures than the ground truth. The number of differences between the two versions of the same movement was recorded. A total of 29,728 measures with 7,196 flagged measures were examined. Flag rates ranged from 0.6% to 79.2% with a weighed mean of 24.2% and median of 21.7%.

The model was then run on each OMR movement and the number of differences with the ground truth was recorded again. (In order to make the outputted score useful for performers and researchers, we added a simple algorithm to preserve as much pitch information as possible from the original measure.) From 2.1% to 36.1% of flagged measures were successfully corrected, with a weighed mean of 18.8% and median of 18.0%: a substantial improvement over the original OMR output.

Manually checking the pre- and post-processed OMR scores against the ground truth showed that the highest rates of differences came from scores where single-pitch repetitions (tremolos) were spelled out in one source and written in abbreviated form in another; such differences could be corrected for in future versions. There was no significant correlation between the percentage of measures originally flagged and the correction rate (r = .17, p > .31).

The model was also run on two scores outside the classical string quartet repertory to test its further relevance. On a fourteenth-century vocal work (transcribed into modern notation), *Gloria: Clemens Deus artifex* and the first movement of Schubert's "Unfinished" symphony, the results were similar to the previous findings (16.8% and 18.7% error reduction, respectively).

The proportion of suggestions taken from the horizontal (PrD) and vertical models (PrP) depended significantly on the number of parts in the piece. In Mozart K525 quartet, 72% of the suggestions came from the horizontal model while for the Schubert symphony (fourteen parts), only 39% came from the horizontal model.

8. APPLICATIONS

The model has broad applications for improving the accuracy of scores already converted via OMR, but it would have greater impact as an element of an improved user experience within existing software. Used to its full potential, the model could help systems provide suggestions as users examine flagged measures. Even a small scale implementation could greatly improve the lengthy error-correcting process that currently must take place before a score is useable. See Figure 4 for an example interface.

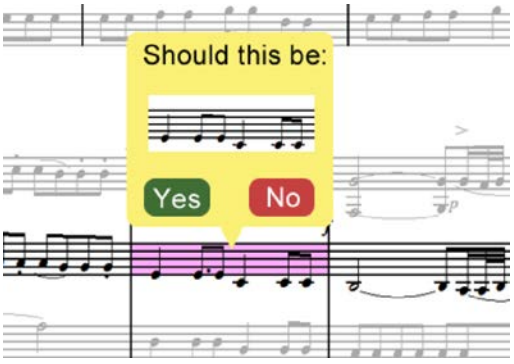

Figure 4. A sample interface improvement using the model described.

A similar model to the one proposed here could also be integrated into OMR software to offer suggestions for pitch corrections if the user selects a measure that was not flagged for rhythmic errors. Integration within OMR software would also potentially give the model access to

rejected interpretations for measures that may become more plausible when rhythmic similarity within a piece is taken into account.

The model could be expanded to take into account spatial separation between glyphs as part of the probabilities. Simple extensions such as ignoring measures that are likely pickups or correcting wrong time signatures and missed barlines (resulting in double-length measures) have already been mentioned. Autocorrelation matrices, which would identify repeating sections such as recapitulations and rondo returns, would improve the prior-based-on-distance metric. Although the model runs quickly on small scores (in far less than the time to run OMR despite the implementation being written in an interpreted language), on larger scores the O(len(F) · len(Part)) complexity of the horizontal model could become a problem (though correction of the lengthy Schubert score took less than ten minutes on an i7 MacBook Air). Because the prior-based-on-distance tends to fall off quickly, examining only a fixed-sized window worth of measures around each flagged measure would offer substantial speed-ups.

Longer scores and scores with more parts offered more possibilities for high-probability correcting measures. Thus we encourage the creators of OMR competitions and standard OMR test examples [3] to include entire scores taken from standard repertories in their evaluation sets.

The potential of post-OMR processing based on musical knowledge is still largely untapped. Models of tonal behavior could identify transposing instruments and thus create better linkages between staves across systems that vary in the number of parts displayed. Misidentifications of time signatures, clefs, ties, and dynamics could also be reduced through comparison across parts and with similar sections in scores. While more powerful algorithms for graphical recognition will always be necessary, substantial improvements can be made quickly with the selective deployment of musical knowledge.

9. ACKNOWLEDGEMENTS

The authors thank the Radcliffe Institute of Harvard University, the National Endowment for the Humanities/Digging into Data Challenge, the Thomas Temple Hoopes Prize at Harvard, and the School of Humanities, Arts, and Social Sciences, MIT, for research support, four anonymous readers for suggestions, and Margo Levine, Beth Chen, and Suzie Clark of Harvard's Applied Math and Music departments for advice and encouragement.

10. REFERENCES

[1] E. P. Bugge, et al.: "Using sequence alignment and voting to improve optical music recognition from multiple recognizers," *Proc. ISMIR,* Vol. 12, pp. 405–410, 2011.

[2] D. Byrd, M. Schindele: "Prospects for improving OMR with multiple recognizers," *Proc. ISMIR,* Vol. 7, pp. 41–47, 2006.

[3] D. Byrd, J. G. Simonsen, "Towards a Standard Testbed for Optical Music Recognition: Definitions, Metrics, and Page Images," http://www.informatics.indiana.edu/donbyrd/Papers/OMRStandardTestbed_Final.pdf, in progress.

[4] M. Cuthbert and C. Ariza: "`music21`: A Toolkit for Computer-Aided Musicology and Symbolic Music Data," *Proc. ISMIR*, Vol. 11, pp. 637–42, 2010.

[5] E. Guo et al.: *Petrucci Music Library,* imslp.org, 2006–.

[6] W. Hewlett, et al.: *MuseData: an Electronic Library of Classical Music Scores,* musedata.org, 1994, 2000.

[7] A. Rebelo, et al.: "Optical music recognition: State-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, Vol. 1, No. 3, pp. 173–190, 2012.

[8] F. Rossant and I. Bloch, "A fuzzy model for optical recognition of musical scores," *Fuzzy sets and systems,* Vol. 141, No. 2, pp. 165–201, 2004.

[9] F. Rossant, I. Bloch: "Robust and adaptive OMR system including fuzzy modeling, fusion of musical rules, and possible error detection," *EURASIP Journal on Advances in Signal Processing*, 2007.

[10] V. Viro: "Peachnote: Music score search and analysis platform," *Proc. ISMIR,* Vol. 12, pp. 359–362, 2011.

This Page Intentionally Left Blank