Microtask crowdsourcing for music score Transcriptions: an experiment with error detection

Samiotis, I.P.; Qiu, S.; Mauri, A.; Liem, C.C.S.; Lofi, C.; Bozzon, A.

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the 21st International Society for Music Information Retrieval Conference

Citation (APA)
Samiotis, I. P., Qiu, S., Mauri, A., Liem, C. C. S., Lofi, C., & Bozzon, A. (2020). Microtask crowdsourcing for music score Transcriptions: an experiment with error detection. In *Proceedings of the 21st International Society for Music Information Retrieval Conference* https://program.ismir2020.net/poster_6-17.html

MICROTASK CROWDSOURCING FOR MUSIC SCORE TRANSCRIPTIONS: AN EXPERIMENT WITH ERROR DETECTION

Ioannis Petros Samiotis **Sihang Qiu** **Andrea Mauri** **Cynthia C. S. Liem**
Christoph Lofi **Alessandro Bozzon**
Delft University of Technology, Netherlands
{i.p.samiotis, s.qiu-1, a.mauri, c.c.s.liem, c.lofi, a.bozzon}@tudelft.nl

ABSTRACT

Human annotation is still an essential part of modern transcription workflows for digitizing music scores, either as a standalone approach where a single expert annotator transcribes a complete score, or for supporting an automated Optical Music Recognition (OMR) system. Research on human computation has shown the effectiveness of crowdsourcing for scaling out human work by defining a large number of microtasks which can easily be distributed and executed. However, microtask design for music transcription is a research area that remains unaddressed. This paper focuses on the design of a crowdsourcing task to detect errors in a score transcription which can be deployed in either automated or human-driven transcription workflows. We conduct an experiment where we study two design parameters: 1) the size of the score to be annotated and 2) the modality in which it is presented in the user interface. We analyze the performance and reliability of non-specialised crowdworkers on Amazon Mechanical Turk with respect to these design parameters, differentiated by worker experience and types of transcription errors. Results are encouraging, and pave the way for scalable and efficient crowd-assisted music transcription systems.

1. INTRODUCTION

Written musical resources, such as tablature or musical scores, are increasingly being digitized. The physical form of such resources would typically be a book, hence, the most obvious digitized form is a scan of the book, encoded in the form of images. In fact, numerous PDF files of scanned sheet music are commonly found on popular music websites such as the Petrucci Music Library (IMSLP [1]). One of the main disadvantages of scans though, is that the musical content contained in them cannot be computationally accessed. Having easily machine-readable formats allows for computational analyses, easier enrichment of the digitized musical resources, and, most importantly, to ease the preservation of and the access to our written musical culture.

The majority of transcriptions for professional use involve experts using specialised interfaces, such as Finale [2] and Sibelius [3], to fully transcribe new editions of existing music manuscripts. Optical Music Recognition (OMR) aims at performing the transcription work automatically; state-of-the-art methods show acceptable performance in the case of clean music scores, but their quality quickly degrades in case of hand written notes [1]. In general, they still require substantial human intervention to provide results with consistent quality [1, 2], while interactive systems that could utilize human evaluation in an efficient and scalable way are still an open issue [3].

Microtask crowdsourcing is a popular approach for scaling up digital content annotation tasks. On online microtask crowdsourcing platforms, such as Amazon Mechanical Turk, large groups of individuals - called workers - perform *microtasks* such as image categorization, and audio or text transcription. By splitting a complex and cognitively intensive task into simpler steps, *microtasks* crowdsourcing allows people with little to no expertise, to contribute to knowledge-intensive activities [4].

Explicit control over a crowd's product, is in the heart of microtask crowdsourcing [5, 6]. To that end, microtasks design should allow the measurement of their outcomes in an algorithmic fashion. Few studies addressed the use of microtask crowdsourcing for music scores transcription, and they typically focus on guiding the workers in the transcription of whole scores [7] or by providing support to the experts [8,9]. However, music scores are complex artefacts that need specific domain knowledge to read and understand, making the task of transcribing a score complex and cognitively demanding. To the best of our knowledge, how to address the task of score transcription through microtask crowdsourcing remains an open research question [10].

This paper contributes towards a better understanding of how music transcription could be supported, and potentially scaled up, through microtask crowdsourcing. We focus on a simple yet fundamental problem: the identification of differences (errors) between two music scores segments. Spotting errors is, in itself, a very useful operation

[1] https://imslp.org/wiki/Main_Page

[2] https://www.finalemusic.com/

[3] https://www.avid.com/sibelius

Attribution: Ioannis Petros Samiotis, Sihang Qiu, Andrea Mauri, Cynthia Liem, Christoph Lofi, Alessandro Bozzon, "Microtask Crowdsourcing for Music Score Transcriptions: An Experiment with Error Detection", in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

in music transcription workflows, as it could be of assistance for experts transcribing a score, with the creation of labeled data to train automated OMR systems, or with the identification of errors made by such a system. Workers operating in online microtask crowdsourcing platforms are already accustomed to such type of tasks, but the understanding of music scores is not a common skill.

The main research question addressed in this work is: *To what extent are workers from microtask crowdsourcing platforms able to detect errors in transcribed music scores?*. To answer the question, we setup an experiment where two microtask design factors were adjusted respectively, the score transcription's *modality* (spotting errors on visual vs. audio), and the *size* (in terms of measures) to be analysed. We recruited 144 workers from Mechanical Turk, asked them to check 144 segments for several types of errors, and measured their performance in terms of completion time, accuracy, and sustained cognitive load.

Results show that crowd workers were able to achieve maximum precision of 94% and accuracy of 85% using an interface that combines visual and audio modalities, thus showing that microtask crowdsourcing is useful for error detection, and that workers benefit from the audio extract of the transcribed score, both alone and as a support for the visual comparison.

2. RELATED WORK

The topic of microtask crowdsourcing for music transcription is scarcely addressed in literature, with many relevant research questions left unanswered. In Burghardt et al. [7] the *Allegro* system was developed, a tool to allow the transcription of entire scores by a (single) human worker. However, *Allegro* has only been tested on a limited number of users, and it was not deployed on an online microtask crowdsourcing platform. The same limitation holds for the work in [8], one of the first attempts to study human input and how the task design can affect human input. This study focused on analysing segments which are one measure long, which is the smallest unit of analysis in our study as well. We expand this, by studying also how the size of the segment shown to the crowd affect its performance. An important work to mention is OpenScore [11], up to now the largest scale project to incorporate humans in music score transcription. In terms of user participation though, it was mainly carried out by seven community members with extensive musical background. Moreover they report different issues related to the management of data (done manually by the administrators of the platform) and user engagement (without any control they would focus on their preferred music score) admitting in the end that in their project "OMR (involving humans) is not currently a scalable solution".

So far, there is not any literature that has targeted unknown crowds with varying skills for music transcription tasks, thus research questions on [10] about what type of tasks users can perform and how to evaluate them still remain open. In this work we address this research gap by looking into similar crowdsourcing works in other domains. More specifically, in [12] it was found that for knowledge-intensive tasks involving artworks, a crowd with varying and unknown domain-specific knowledge found on online platforms can produce useful annotations when aided by good task design. Research has shown that UI design is an important part of a microtask design [13]. Research so far has experimented with various designs such as showing spectogram visualisations for audio annotation [14] or the use of chat-bots to assist common types of microtasks [15], all of which have yielded positive results on the performance of the crowdworkers. Inspired by this we make the design of the work interface a central point of our study.

3. EXPERIMENTAL DESIGN

The main focus of this work is to study to what extent a general crowd can identify *errors* in a music score transcription. We therefore designed an experiment aimed at testing the ability of crowd workers to spot errors using interfaces having a combination of visual and audio components.

3.1 Task Design

Our aim is to study how different task design factors can influence the crowdworkers performance, focusing on two aspects:

1. The *modality* (*visual* versus *audio*) used to spot errors: as music scores are complex artefacts, and music is primarily an auditory experience. Therefore, we investigate how the score comparison *modality* affect the error detection performance in workers that are potentially not familiar with musical notation. Intuitively, we want to investigate if "hearing" errors is easier that "seeing" errors.

2. The score *size* offered to crowdworkers for annotation. The goal is to assess how the size (in terms of measures) of the score offered to worker affects their performance.

To develop a better understanding on the characteristics of the crowd, we open the tasks with questions about their demographic information (occupational status, level of education and age) and their music sophistication. For the latter, we compiled a list of 6 questions from The Goldsmiths Musical Sophistication Index (Gold-MSI) [16].

Crowd workers' performance with error identification is measured using accuracy, precision and recall and time to execute a microtask. In addition, we measure user confidence with their judgements with a seven-value Likert scale.

Finally, we measure the sustained cognitive load when executing the microtask, measured through the NASA Task Load Index (TLX)[4], which ranges from 0 to 100 (higher the TLX is, the heavier cognitive load the worker perceives).

[4] https://humansystems.arc.nasa.gov/groups/TLX/

3.2 Evaluation Dataset

Selecting a suitable music score was our first step preparing our experiments. We use a single classical music score to avoid introducing additional variable in workers' performance. Specifically we use the Urtext of "*32 Variations in C minor*" by Ludwig van Beethoven. It is a piano piece and the music artifacts are all printed typeset forms. This is a slightly easier use case than hand-written scores. The score was retrieved from IMSLP as a PDF[5] .

As a Gold standard transcription of that PDF we used an MEI[6] file that had been transcribed by an expert. This file was accepted as error free, and it allowed us introduce errors in a controlled way for our experiments.

We segmented the music score in varying sizes to investigate how workers cope with shorter or longer tasks. We distinguish 1) *one measure* segments, 2) segments of *two measures* and 3) segments of *three measures*. Both of the two digital versions of the score, the PDF file of the original score and the transcribed MEI file, were segmented using the aforementioned segment sizes. The segmentation of the PDF was performed manually, while for the MEI we used the appropriate identifiers of each measure that was included in the corresponding image segment, to isolate the correct headers in the MEI. Since it's a piano score, each measure contains two staves.

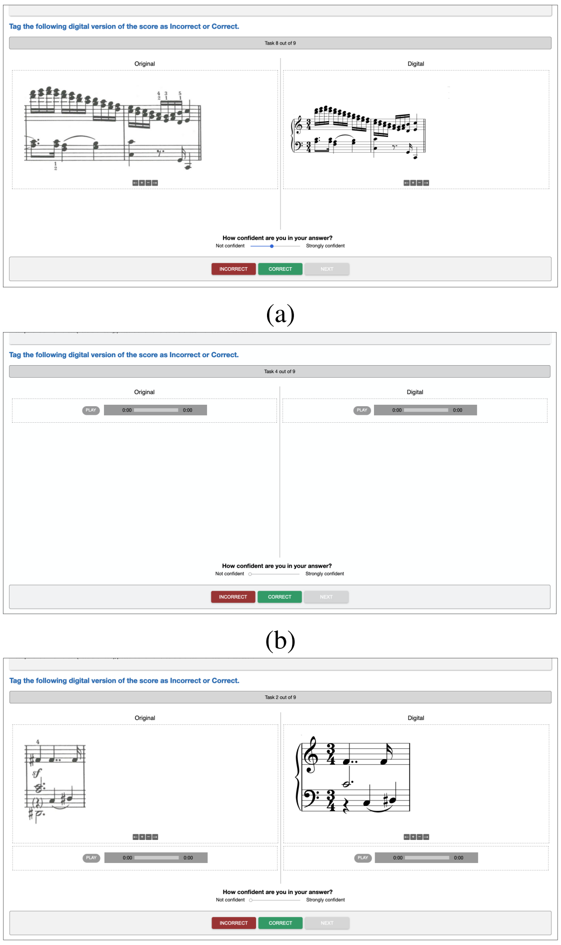
We then introduced errors in the MEI segments derived from common errors that can occur in automatic OMR systems. The type of errors could impact the crowdworkers' ability to spot them and correctly identify them as errors. Some of them can be challenging to notice even to experts of the field. Therefore, we study different types of errors, all focusing on the music notes themselves and their accidentals. Errors on performance annotations, clefs, finger numbers etc, are out of scope in this study. We introduced the following types of error per MEI segment: 1) *Missing notes*; 2) *Wrong vertical position of a note*; 3) *Wrong duration of a note*; 4) *Wrong accidental*.

Each segment that was shown to the user contained only one type of error. The amount of errors per segment was kept constant at 40% of the notes present in the segment. Thus, if a crowdworker is presented with two measures with notes missing, then notes are missing on both measures at a 40% rate of the total notes on both measures combined. No more types of error are present in the segment.

To make the performance comparisons meaningful, we ensured that our dataset is balanced across all error types. In total we used 144 segments derived from the entirety of the selected piano score, with 48 segments per size category, from which 24 were equally distributed to each type of error, while the remaining 24 were kept correct.

3.3 User Interface Design

To test the modalities' effects separately and accurately, we designed three different interfaces: one that would have image to image comparison to test the traditional form of the task, one with only audio to audio comparison, and one with both audio and image comparison. The interfaces are designed to include the following data. 1) **Original Score**: the segment's image from the scanned score. 2) **Correct MEI Render**: a render of the transcribed version of the *Original Score*'s segment; 3) **Incorrect MEI Render**: a render of the MEI transcription containing errors. 4) **Correct MIDI**: the MIDI extract of the correct version of *MEI Transcript*'s segment. 5) **Incorrect MIDI**: the MIDI extract of *MEI Transcript*'s segment containing errors.



Tag the following digital version of the score as Incorrect or Correct.

Task 8 out of 9

How confident are you in your answer?

(a)

Tag the following digital version of the score as Incorrect or Correct.

Task 4 out of 9

How confident are you in your answer?

(b)

Tag the following digital version of the score as Incorrect or Correct.

Task 2 out of 9

How confident are you in your answer?

(c)

Figure 1. Microtask User Interfaces: (a) Visual, (b) Audio and (c) Combination

We refrained from using audio from a recorded performance against a MIDI extract containing errors to avoid confusing the crowd on what constitutes as "different" or an "error" in the audio comparison task. A recorded performance would introduce performance-related artefacts to the audio, which do not exist in a MIDI extract, thus increasing the chance of false negatives in identifying an audio snippet as "incorrect". Finally, for the combined comparison, we used the same elements as with the individual comparisons. For the renders of the MEI transcripts and MIDI extracts we used Verovio[7] on our interfaces.

From a design perspective, the interfaces needed to be

[5] https://imslp.org/wiki/32_Variations_in_C_minor%2C_WoO_80_(Beethoven%2C_Ludwig_van)

[6] https://music-encoding.org/

[7] https://www.verovio.org/index.xhtml

simple and closely resembling each other to minimize their effect on the workers' judgements. They should also be able to facilitate the different segment sizes without changing the layout. Eventually, we also wanted to accommodate error detection in the same manner for both image and audio comparisons, to avoid differences on the annotation tools being another factor to the crowd's performance. Therefore, we designed the error detection task to ask from the users to annotate a given MEI transcript or MIDI extract as "*Incorrect*" if it exhibit errors and as "*Correct*" if they did not.

In all interfaces, the segments to the left are associated to the the original scanned score and the correct MEI transcription of it, while to the right we always place the segments that need to be annotated. The MEI render or the MIDI extract to the right can be either "*Correct*" or "*Incorrect*" and we calculate the performance of the workers based on identifying this correctly. In addition to the two buttons for the labels, we included a slider to indicate the worker's confidence in their label. Later in the analysis of the results, we used this indicator to study how each interface and segment size affected the confidence of the workers' to their judgements. These design considerations resulted in the following three interface designs:

- **Original Score** against **Correct/Incorrect MEI Render (Visual)**: This user interface, depicted in Figure 1(a), shows the segment of the original scanned score to the left, with the corresponding MEI render to the right. The user needs to compare the two images and spot differences related to the types of errors.
- **Correct MIDI** against **Correct/Incorrect MIDI (Audio)**: In this interface, as shown in Figure 1(b), we let the user listen to the correct MIDI extract on the left and the one generated from the MEI transcription to the right.
- **Original Score and Correct MIDI** against **Correct/Incorrect MEI and Correct/Incorrect MIDI (Combination)**: This final user interface, as shown in Figure 1(c), combines elements of the previous two. The user here has the option to either use the visual comparison, the audio comparison or both to realise if there are errors to the segment to the right. The MEI render and MIDI extraction to the right always originate from the same MEI transcription, therefore both will be correct or both will contain errors.

Each combination of interface with a segment size consists of a microtask. To efficiently and effectively gather performance data, we wanted the same worker to be "exposed" to all nine possible combinations. Therefore, in its final form, a worker would have to execute a task that would begin with a set of demographic and music sophistication questions, followed by the nine microtasks and end with the cognitive load questionnaire. To minimise the impact of issues related to the familiarity of workers with the interface, the task also includes an introductory explanation of the work interface, with examples of errors and expected responses. The results of our experiment are analysed based on the overall, but also on error type, performance of workers.

4. RESULTS

As discussed in previous sections, we published our tasks on Amazon Mechanical Turk (MTurk). The platform offers several configurations for each batch of tasks submitted. We published them as public so they can be accessed by all the users of the platform and we did not require any Mechanical Turk Master (expert workers). Only to avoid malicious workers, we filter them by their previous HIT Approval Rate, and we set it to 95%.

In total, 144 workers executed our tasks on MTurk and we paid them per task execution according to the average US minimal hourly wage [8] . In order to minimize the effect of any biases or learning effect we randomized the order of the presentation of the different task designs (UI-segment size combination). One worker eluded the quality verification on task interface, which results in 143 unique workers.

4.1 Worker Demographics

From a demographic aspect, most of the workers (84.6%) reported that they had a full time occupation. Also, 67.8% of total workers reported their education level was Bachelor's degree, while 14.9% of them had Master's degree. Only 8.3% of the workers were above 45 years old.

Answers to the Gold-MSI questions indicate that the majority of workers seem to be familiar with listening to music, as 56% of them listen to music for at least 1 hour a day and 65% say they can hit the proper notes while listening to a record. Also, the majority of them (75%) state they can properly compare and discuss different performances of the same music piece. On the other hand, 52.4% of the workers reported having up to one year formal training in music theory, where the 26.6% has no prior education on the subject. Also, 41.95% of the workers have trained for maximum one year on a music instrument, while 22.4% of never had. Their answers here suggest that the majority of the crowd has little to no music theory background, and a considerable amount of them also no formal studies on an instrument. The results also suggest that the crowd executing our tasks was mainly composed of workers with little expertise with music theory.

4.2 Accuracy

The results per target segment were aggregated from three different individual workers. Table 1 shows that tasks performed with the *Audio* interface consistently achieved higher accuracy than the *Visual* one. The *Combined* interface achieved good accuracy figures with all segments sizes, but best accuracy with the 3-measure-long segments. The *Visual* interface yielded consistently the lowest recall and accuracy results, for all segment sizes. Interestingly,

[8] We estimated an average task completion time of 15'; each crowdworker was awarded 2.5$

Interface	**segment size**		
	1	2	3
Visual	**P=66.22** R=59.76 A=60.27	**P=65.28** R=61.84 A=62.24	P=59.72 R=74.14 A=69.23
Audio	P=60.27 **R=81.48 A=72.92**	P=62.50 **R=81.82 A=74.13**	P=64.79 R=80.70 A=74.83
Combined	P=59.70 R=68.97 A=67.63	**P=65.28** R=74.60 A=71.33	**P=68.06 R=83.05 A=76.92**

Table 1. Precision, Recall and Accuracy of *individual* answers, by segment sizes and interfaces.

Interface	**segment size**		
	1	2	3
Visual	P=60.71 **R=70.83** A=62.50	P=67.86 **R=79.17** A=70.83	P=88.89 R=66.67 A=79.17
Audio	**P=100.0** R=62.50 **A=81.25**	**P=88.24** R=62.50 A=77.08	P=90.00 R=75.00 A=83.33
Combined	P=76.47 R=56.52 A=70.21	P=85.00 R=70.83 **A=79.17**	**P=94.74 R=75.00 A=85.42**

Table 2. Overall Precision, Recall and Accuracy of *aggregated* answers by segment sizes and interfaces. In bold you find the highest precision, recall and accuracy by segment size, while underlined you find the highest overall performance

the precision for this interface on segment size one, was the highest compared to *Audio* and *Combined* for the same size.

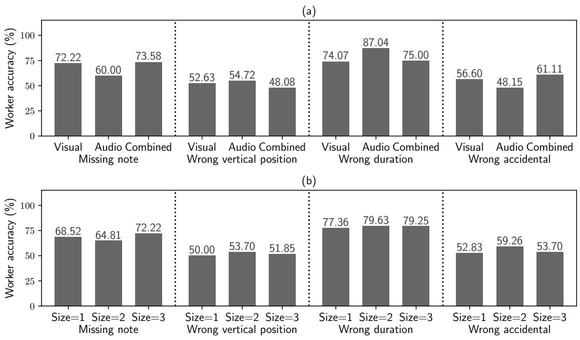


Figure 2. Workers error detection accuracy (unit:%) (a) per user interface and (b) per segment size.

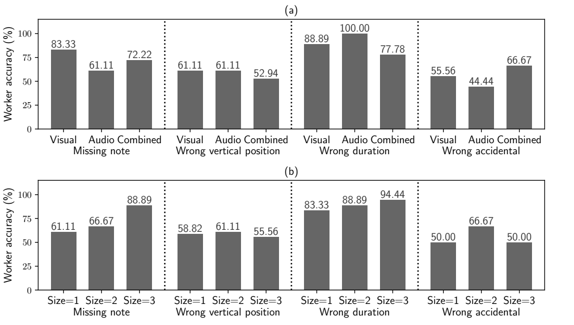


Figure 3. Aggregated error detection accuracy (unit:%) (a) per user interface and (b) per segment size.

Figure 2 shows the accuracy of the workers in detecting specific type of errors. *Wrong duration* error seems to be accurately spotted in any user interface and segment size, with the *Audio* interface resulting in the highest accuracy (87.04%). Workers perform better detecting the *Missing Note* error using the *Combined* interface and the 3-measure segments. The accuracy obtained with the *Visual* interface though, suggests that workers might rely more on the image of the score rather than the audio for this type of error. The *Wrong Vertical position* error was more difficult to detect in general; the highest accuracy was obtained with the *Audio* interface (54.72%) and with the segment size of 2 measures (53.70%). Finally, the *Wrong accidental* type was the second most demanding error to be detected with the highest accuracy achieved using *Combined* interface (61.11%), with a slight improvement in segments containing 2 measures.

In microtask crowdsourcing it is common to aggregate individual results to improve overall quality. Table 2 shows the performance achieved using a simple *majority voting* aggregation scheme. The *Combined* interface with 3-measure-long segments still achieves best performance with a remarkable 94% in precision, and 85% in accuracy. The *Audio* interface achieves best precision performance for both 1-measure-long and 2-measure-long segments, while the *Visual* interface achieves best recall.

Figure 3 shows the aggregated accuracy in detecting specific type of errors. In terms of *Wrong Duration* error, the accuracy remains the highest after the aggregation. The *Audio* interface and the 3-measure-long segments achieve 100% and 94% in accuracy respectively. *Visual* interface and the 3-measure-long segments obtain the highest accuracy (82% and 88% respectively) in detecting *Missing note* error. The *Wrong Vertical position* error and the *Wrong accidental* error still have relatively low accuracy.

4.3 Execution Time

Figure 4 shows that, as expected, execution time generally increases as the segment size increase. We performed statistical tests (independent t-tests, $\alpha = 0.05$) to find significant differences between interfaces and segment sizes. In the case of *Wrong vertical position* error though, the *Audio* interface allowed the worker to spot the errors significantly faster compared to *Combined* interface (p = 3.5e-3). For the *Wrong duration* error the addition of audio and the increased segment size can lead to a significantly longer average execution time (for both *Audio* and *Combined* interfaces compared to *Visual*, p = 1.3e-4 and p = 1.9e-5 respectively; and for 3-measure-long segment vs 1-measure-long segment, p = 2.5e-3).

For the *Wrong accidental* case, we see that worker spent less execution time on the *Visual* interface (no significance). However, comparing it with the results, the worker most probably dismissed the segment as "Correct", rather spend more time in case they had missed the error.

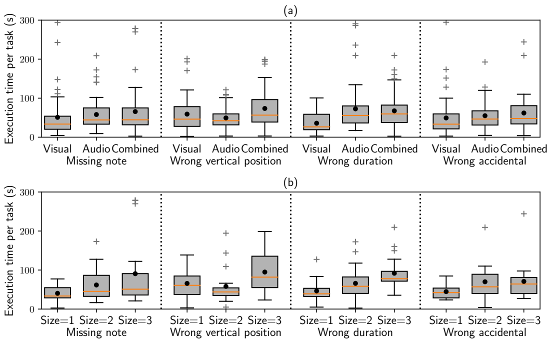


Figure 4. Worker execution time (unit: seconds) of each microtask by (a) user interface and (b) segment size.

4.4 Music Sophistication and Cognitive Load

The average score of cognitive task load (NASA-TLX) is 47.7%, a typical value for classification and similar cognitive tasks [17]. To investigate how music sophistication relates to worker performance and cognitive load on spotting music errors, we select and analyze a corresponding question from Gold-MSI questionnaire – "I find it difficult to spot mistakes in a performance of a song even if I know the tune.", where workers need to select an option from `Completely Disagree` to `Completely Agree`, before they execute the music transcription tasks.

Results show that 47% workers agreed with the statement that it was difficult to spot mistakes in a performance of a song; 33% of them disagreed with the statement, and the rest of them (20%) were unsure. We calculated the worker accuracy and the cognitive task load score, and performed significant testings (independent t-test, $\alpha = 0.05$). Workers who reported lower difficulty with spotting music errors (accuracy = $81 \pm 15\%$, TLX score = 44.36 ± 13.05) outperformed workers who had higher difficulty (accuracy = $63 \pm 16\%$, TLX score = 50.86 ± 13.61) in terms of both worker accuracy and perceived cognitive load (p = 3.3e-8 and 0.013 respectively). The workers who reported lower difficulty also had significantly higher accuracy ($p = 0.011$) compared to unsure workers (accuracy = $0.70 \pm 0.20\%$, TLX score = 46.01 ± 14.27). Results suggest that the self-reported music sophistication in some specific aspects strongly relates to actual worker performance in error identification and cognitive load. Nonetheless, workers with lower sophistication still achieved good performance, with a small additional cost in terms of cognitive load.

5. DISCUSSION

As expected, people with some formal knowledge in music, which could be useful to comprehend music scores, are very rare "in the wild". To enable the use of microtask crowdsourcing for music score transcription, good task design is therefore of essence. Results show that error detection is a task that could be successfully performed in a microtask crowdsourcing setting. Offering audio extracts of a target music score can positively affect the performance of the crowdworkers, especially for short segments of one or two measures. With larger segments, even though audio extracts are still yielding better results against to the textual measures of the score, a combination of the two modalities is more preferable. This result gives important indications for task splitting and scheduling purposes, as it suggests that it is possible to evaluate larger portions of scores without incurring accuracy penalties. This has obvious implications in terms of overall transcription costs.

In terms of types of detected errors, results suggest that the *Missing Note* and *Wrong Duration* errors are the easiest to be found, while the crowd had relatively more difficulty detecting *Wrong Accidentals* and *Wrong vertical position* ones. Furthermore, we see the clear effect of user interface and segment sizes in identifying correctly specific errors. Specifically, the *Audio* interface helps in finding *Wrong duration* errors, while the *Combined* one increases the accuracy in finding *Wrong accidental* mistakes. Showing segments longer than two measures seems to slightly hinder the ability of the crowd to detect any errors besides *Missing notes*.

Limitations. Correct MEI render and correct MIDI files of scores to be transcribed are typically not available in the real world. The distribution of errors in the evaluation dataset might not reflect the actual distribution of errors produced, for instance, by OMR systems. Given these limitations, the results of our experiment are probably to be interpreted as an "upper bound" in terms of achievable performance; nonetheless, they clearly indicate that the detection of errors in transcribed music score is an activity that can be successfully performed by crowdworkers.

6. CONCLUSION

Music score transcription is an important activity for written music preservation. Through this work, we show that microtask crowdsourcing can be used to scale up specific transcription activities. Worker interfaces that combine visual and audio modalities allow the evaluation of longer score segments. Focusing on the error detection task, results show that crowd workers can achieve high precision and recall, especially with *Missing Note* and *Wrong Duration errors*. In future work, we plan to expand the evaluation dataset, perform experiments where workers are asked to compare recorded performance, and address a broader set of transcription errors. Finally, we will investigate other types of microtasks, and study to what extent crowd workers could also be employed to *transcribe* scores.

7. ACKNOWLEDGEMENTS

This work is partially supported by the European Commission under the TROMPA project (H2020 770376) and was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. We thank Dr Werner Goebl for providing us the correct MEI transcription of 32 Variations in C minor, WoO 80 by Ludwig van Beethoven.

8. REFERENCES

[1] B. Almeida and S. Spanner, “Allegro: User-centered Design of a Tool for the Crowdsourced Transcription of Handwritten Music Scores,” in *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, vol. 25, no. 23. New York, New York, USA: ACM Press, 2017, pp. 15–20. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3078081.3078101

[2] P. Bellini, I. Bruno, and P. Nesi, “Assessing Optical Music Recognition Tools,” *Computer Music Journal*, vol. 31, no. 1, pp. 68–93, mar 2007. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/comj.2007.31.1.68

[3] J. Calvo-Zaragoza, J. H. Jr., and A. Pacha, “Understanding optical music recognition,” *ACM Comput. Surv.*, vol. 53, no. 4, Jul. 2020. [Online]. Available: https://doi.org/10.1145/3397499

[4] J. Oosterman, J. Yang, A. Bozzon, L. Aroyo, and G.-J. Houben, “On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks,” *Computer Networks*, vol. 90, pp. 133 – 149, 2015, crowdsourcing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128615002315

[5] E. Law and L. v. Ahn, “Human computation,” *Synthesis lectures on artificial intelligence and machine learning*, vol. 5, no. 3, pp. 1–121, 2011.

[6] A. Bozzon, M. Brambilla, S. Ceri, and A. Mauri, “Reactive crowdsourcing,” in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 153–164.

[7] M. Burghardt and S. Spanner, “Allegro: User-centered design of a tool for the crowdsourced transcription of handwritten music scores,” in *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage*, ser. DATeCH2017. New York, NY, USA: ACM, 2017, pp. 15–20. [Online]. Available: http://doi.acm.org/10.1145/3078081.3078101

[8] L. Chen and C. Raphael, “Human-Directed Optical Music Recognition,” *Electronic Imaging*, vol. 2016, no. 17, pp. 1–9, feb 2017. [Online]. Available: http://www.ingentaconnect.com/content/10.2352/ISSN.2470-1173.2016.17.DRR-053

[9] L. Chen, R. Jin, and C. Raphael, “Human-Guided Recognition of Music Score Images,” in *Proceedings of the 4th International Workshop on Digital Libraries for Musicology - DLfM ’17*. New York, New York, USA: ACM Press, 2017, pp. 9–12. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3144749.3144752

[10] C. Saitis, A. Hankinson, and I. Fujinaga, “Correcting large-scale omr data with crowdsourcing,” in *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, 2014, pp. 1–3.

[11] M. Gotham, P. Jonas, B. Bower, W. Bosworth, D. Rootham, and L. VanHandel, “Scores of scores: an openscore project to encode and share sheet music,” in *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, 2018, pp. 87–95.

[12] J. Oosterman, A. Bozzon, G.-J. Houben, A. Nottamkandath, C. Dijkshoorn, L. Aroyo, M. H. Leyssen, and M. C. Traub, “Crowd vs. experts: nichesourcing for knowledge intensive tasks in cultural heritage,” in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 567–568.

[13] U. Gadiraju, A. Checco, N. Gupta, and G. Demartini, “Modus operandi of crowd workers: The invisible role of microtask work environments,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–29, 2017.

[14] M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J. P. Bello, and O. Nov, “Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–21, 2017.

[15] P. Mavridis, O. Huang, S. Qiu, U. Gadiraju, and A. Bozzon, “Chatterbox: Conversational interfaces for microtask crowdsourcing,” in *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 2019, pp. 243–251.

[16] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, “The musicality of non-musicians: an index for assessing musical sophistication in the general population.” *PloS one*, 2014.

[17] R. A. Grier, “How high is high? a meta-analysis of nasa-tlx global workload scores,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1, pp. 1727–1731, 2015. [Online]. Available: https://doi.org/10.1177/1541931215591373