AUTOMATIC MUSIC TRANSCRIPTION AND ETHNOMUSICOLOGY: A USER STUDY

Andre Holzapfel
KTH Royal Institute of Technology
holzap@kth.se

Emmanouil Benetos
Queen Mary University of London
emmanouil.benetos@qmul.ac.uk

ABSTRACT

Converting an acoustic music signal into music notation using a computer program has been at the forefront of music information research for several decades, as a task referred to as automatic music transcription (AMT). However, current AMT research is still constrained to system development followed by quantitative evaluations; it is still unclear whether the performance of AMT methods is considered sufficient to be used in the everyday practice of music scholars. In this paper, we propose and carry out a user study on evaluating the usefulness of automatic music transcription in the context of ethnomusicology. As part of the study, we recruited 16 participants who were asked to transcribe short musical excerpts either from scratch or using the output of an AMT system as a basis. We collect and analyze quantitative measures such as transcription time and effort, and a range of qualitative feedback from study participants, which includes user needs, criticisms of AMT technologies, and links between perceptual and quantitative evaluations on AMT outputs. The results show no quantitative advantage of using AMT, but important indications regarding appropriate user groups and evaluation measures are provided.

1. INTRODUCTION

Automatic music transcription (AMT) is the process of transferring an music audio signal to a symbolic representation using computational methods [14, p.30]. Engineering research has been developing AMT methods for a number of decades now (see *e.g.* [17] for an early example), and it represents a recurrent theme in the discourse in the field of music information retrieval (MIR). In the field of comparative musicology – the historical predecessor of ethnomusicology – the idea of using automatic methods to obtain a graphical representation (not necessarily symbolic) from a music recording attracted interest from the early days of audio recording technology [7]. This long history of interest in AMT technology in two rather remote fields motivates us to investigate what the current state of the art in AMT may have to offer for (ethno)musicologists transcribing a piece of music.

In the field of MIR, recent AMT research has mostly focused on automatic transcription of piano recordings in the context of Western/Eurogenetic music (see [3] for a recent overview). The vast majority of proposed methods aim to create systems which can output a MIDI or MIDI-like representation in terms of detected notes with their corresponding onsets/offsets in seconds. Such methods are typically evaluated quantitatively using multi-pitch detection and note tracking metrics also used in the respective MIREX public evaluation tasks [2]. Methods that can automatically convert audio into staff notation include the beat-informed multi-pitch detection system of [8] in the context of folk music (the dataset of this work is also used in the present user study) and the multi-pitch detection and rhythm quantization system of [15], which was applied to Western piano music. Recently, methods inspired by deep learning theory have also attempted to automatically convert audio directly into staff notation [5, 18], although these methods are mostly constrained to synthesized monophonic excerpts using piano soundfonts.

Within ethnomusicology, transcription may take a large variety of forms, depending on analytic goals and the analyzed musical context [20]. An early study of the commonalities and discrepancies between transcriptions of the same piece by several experts was conducted by List in 1974 [12]. The study investigated transcriptions of three pieces by up to eight transcribers, and documented higher consistency in the notation of pitch than in duration. An estimated pitch curve was provided as well, and the study demonstrated that only small corrections were conducted by the transcribers. The value of user studies has been recognized in MIR [9, 10, 19, 21], and MIR user studies have been conducted, for instance, in the context of applying music to achieve certain emotional states [6] and therapeutic applications [11]. However, to the best of our knowledge, since [12] no user studies have been conducted that study a larger group of transcribers in their interaction with the output of an AMT system.

In the context of AMT, the research questions that can be approached by a user study are manifold. In this study, we investigate the relations between the output of a state-of-the-art AMT system and manual transcriptions, the validity of quantitative evaluation metrics when comparing manual transcriptions, and the question of whether using an AMT as a starting point for transcription provides ad-

vantages of any kind. To this end we conducted a study with 16 experienced transcribers, and asked them to transcribe eight excerpts of a particular musical style with a specific analytic goal. Transcriptions were either to be performed completely manually, or using an AMT output as a starting point. Our results document a range of insights into the qualities and problems of AMT and evaluation metrics. Even after decades of development of AMT systems, our study cannot reveal a clear advantage of using AMT to inform manual transcription, but our results indicate promising avenues for future development.

The outline of this paper is as follows. Section 2 presents the method, and Section 3 the results of the study. Sections 4 and 5 provide discussion and conclusions, respectively.

2. PROPOSED STUDY

2.1 Subjects

Participants for the proposed study were recruited from the Institute of Musicology in Vienna, Austria, from SOAS University of London, UK, and from City, University of London, UK. In total, 16 subjects participated in our study, nine male and seven female. The criteria for the participation in the study were being an advanced student or recent graduate in a musicology or ethnomusicology program, having attended training on music transcription / musical dictation and being recommended by a member of faculty as being good transcribers. Apart from these students, two musicology lecturers also participated as subjects.

The participants had 16 years of music training on average, with a standard deviation of 9 years. In terms of their interests, 6 participants closely identified with Western classical music, and 10 participants identified with world/folk/traditional music. In terms of their professional practice, 9 participants engaged with Western classical music, and 8 with world/folk/traditional music. In terms of software for music notation and transcription, 7 participants were familiar with MuseScore, 6 with Transcribe!, 5 with Sibelius, and 2 with Sonic Visualiser.

2.2 Material

For this study, we use audio recordings and corresponding transcriptions collected as part of the Crinnos project [1], which were also used as part of the *Sousta Corpus* for AMT research in [8]. All recordings used in this study were recorded in 2004 in Crete, Greece, and all regard a specific dance called *sousta*. Recordings selected for this study were transcribed by ethnomusicologists in Western staff notation as part of the Crinnos project.

These recordings were chosen for the present study for several reasons. They provide a dataset that is highly consistent in terms of musical style, thus appropriate for an AMT user study consisting of multiple excerpts. The sousta dance is usually notated in 2/4 meter and has a relatively stable tempo, again providing consistency for human transcribers. The instrumental timbres are likewise highly consistent, with one Cretan lyra (a pear-shaped fiddle) playing the main melody, and usually two Cretan lutes playing the accompaniment.

Eight audio excerpts from the *Sousta Corpus* were selected for the present study. The length of each excerpt was set to 4 bars, which results in a duration of 7-8 seconds per excerpt. The number of excerpts and their duration were determined through pilot studies, with the goal to constrain the duration of the proposed study for each participant to 2 hours. The position of the 4 bars within each piece was chosen as such to provide study participants with a complete musical phrase, in order to aid transcription. The corpus of [1] also contains corresponding reference transcriptions in musicXML format, which were used for quantitative evaluations. [1]

We did not assume that participants are familiar with the music culture used in this study. Therefore, one complete recording from the corpus of [8] was also selected in order to familiarize participants with the music culture prior to the start of the study.

2.3 AMT methods

For the purposes of this study, an AMT method is needed that can convert an audio recording into machine-readable Western staff notation, suitable for audio recordings from the particular music culture employed for this study. In terms of academic research, the pool of candidate AMT methods for audio-to-staff music transcription is limited to the beat-informed matrix factorization-based system used for the same corpus in [8], the two-stage piano-specific polyphonic transcription system from [15], plus preliminary works for end-to-end piano-only transcription using synthesized audio [5, 18]. In terms of commercial AMT software, a partial list is included in [3], out of which only a small subset (ScoreCloud, Sibelius' AudioScore plugin) produces transcriptions in staff notation.

We selected the beat-informed matrix factorization-based AMT method [8] from the above list of candidate AMT methods, because of its suitability for the present corpus. In terms of commercial tools, we selected ScoreCloud [2], given its competitive performance in monophonic transcription of violin recordings, an instrument with timbre characteristics similar to the Cretan lyra. Based on quantitative AMT evaluations of both systems (shown in Section 3.1), it was decided to use the ScoreCloud AMT system for the present user study.

Since the objective of this study is for participants to transcribe the main melody and not to focus on accompaniment and ornamentations, the automatic transcriptions produced by the systems of [8] and ScoreCloud were modified as to remove the bass staff (if it exists) along with all transcribed notes in that staff; all ornamentations (e.g. trills, grace notes) and note groupings were deleted. We also changed sharp or flat symbols in the automatic transcriptions as to have consistent accidentals for each excerpt.

[1] The excerpts and reference transcriptions can be obtained at `https://bit.ly/2ZKhmnY`

[2] `https://scorecloud.com/`

2.4 Procedure

Experiments took place in quiet rooms; participants were provided with a laptop (if they did not have their own), headphones, printed or digital automatic transcriptions (as desired by participant), manuscript paper, and the study questionnaire. Participants were video recorded in order to assist with the subsequent annotation process.

Participants were asked to transcribe the main melody for each excerpt and to not transcribe the accompaniment or ornamentations. The purpose of this specification was to clarify the analytic goal of the transcription. Participants were free to use the music notation software of their preference or to transcribe on manuscript paper. The study consisted of 8 excerpts per participant, with 4 excerpts to be manually transcribed, and 4 excerpts to be accompanied with AMT outputs in printed and machine-readable format, to be used as a starting point for transcriptions. The order of manual and edited transcriptions was interleaved, and participants were either asked to start transcribing their first segment manually or to edit an automatic transcription. The order of the 8 excerpts exposed to participants was randomized. Fig. 1 shows an example automatic transcription produced using ScoreCloud, compared with a reference and a study participant transcription.

Following the study, a short conversation with participants took place, in order to obtain qualitative feedback as well as information on their experience with automated tools for the task. All participant transcriptions that were produced on manuscript paper were re-transcribed by the authors in machine-readable music notation using MuseScore, in order to carry out quantitative evaluations.

2.5 Evaluation Metrics

2.5.1 Participant Questionnaire

Participants were asked to quantify their effort for every excerpt towards producing the transcription on a scale 1-10 (1: no effort, 10: very high effort). In addition, for every excerpt to be edited from an automatic transcription, participants were asked to rate the quality of the AMT (on a scale 1-10, with 10 being excellent). After completing the experiment, participants were asked to specify the most crucial mistakes present in the automatic transcriptions, and to comment on the possible value of AMT as a starting point towards producing manual transcriptions.

2.5.2 Quantitative metrics

In order to evaluate the performance of the automatic transcription methods, as well as to compare the participants' transcriptions with the reference transcriptions, we use the quantitative metrics for complete AMT proposed in [15] [3] . We chose to compare with these particular reference transcriptions, because their transcribers had extended experience with both transcription and the musical style.

These quantitative metrics are based on an automatic alignment of the estimated score to the reference score using the method of [16]. Following alignment, we are able to identify correctly detected notes, notes with pitch errors (also called *pitch substitution errors*), extra notes, and missing notes. Based on the above definitions, the following error rates are used, as per [15]: pitch error rate E_{p}, extra note rate E_{e}, missing note rate E_{m}, and onset time error rate E_{on}. We also define an average error metric E_{mean} as the arithmetic mean of all 4 aforementioned metrics. As additional quantitative metric, we also measured the time taken by each participant to transcribe each excerpt.

[3] Noting that typical metrics used in multi-pitch detection and note tracking [2] are not suitable for evaluating transcriptions in staff notation.

3. RESULTS

3.1 AMT system evaluation

Excerpt#	E_{p}	E_{e}	E_{m}	E_{on}	E_{mean}
1	18.75	18.18	15.62	48.15	25.18
2	10.71	3.85	10.71	36	15.32
3	20.69	28.57	31.03	45	31.32
4	10	29.03	26.67	18.18	20.97
5	2.5	38	22.5	25.81	22.20
6	7.69	7.14	0	23.08	9.48
7	13.64	3.12	29.54	61.29	26.90
8	40.91	41.18	9.09	70	40.29
Average	15.61	21.13	18.15	40.93	23.96

Table 1. AMT quantitative evaluation scores using ScoreCloud.

Excerpt#	E_{p}	E_{e}	E_{m}	E_{on}	E_{mean}
1	12.5	66.67	66.67	75	55.21
2	21.05	22.22	26.31	35.71	26.33
3	0	100	100	0	50.00
4	3.70	77.78	77.78	50	52.31
5	17.39	39.13	39.13	21.43	29.27
6	12.5	16.67	16.66	35	20.21
7	33.33	0	25.64	31.03	22.50
8	38.46	53.57	0	53.85	36.47
Average	17.37	47.00	44.02	37.75	36.54

Table 2. AMT quantitative evaluation scores using the method of [8].

Tables 1 and 2 depict the error rates for all eight examples used in the experiments, using ScoreCloud and the AMT system of [8], respectively. The ScoreCloud system performs significantly better than the system of [8] (based on a paired-sample t-test, $p < 0.05$) for the extra- and missing note rates, and for the mean error rate E_{mean}.

Ranking of the examples looks quite inconsistent between ScoreCloud and the method of [8]. Segment 6 has lowest error rates for both, but the highest error rates are for Segment 8 for ScoreCloud, and Segment 1 for [8]. However, when calculating the average E_{mean} of both algorithms, one could identify Segment 3 as the most challenging (2nd highest error rate for both algorithms).

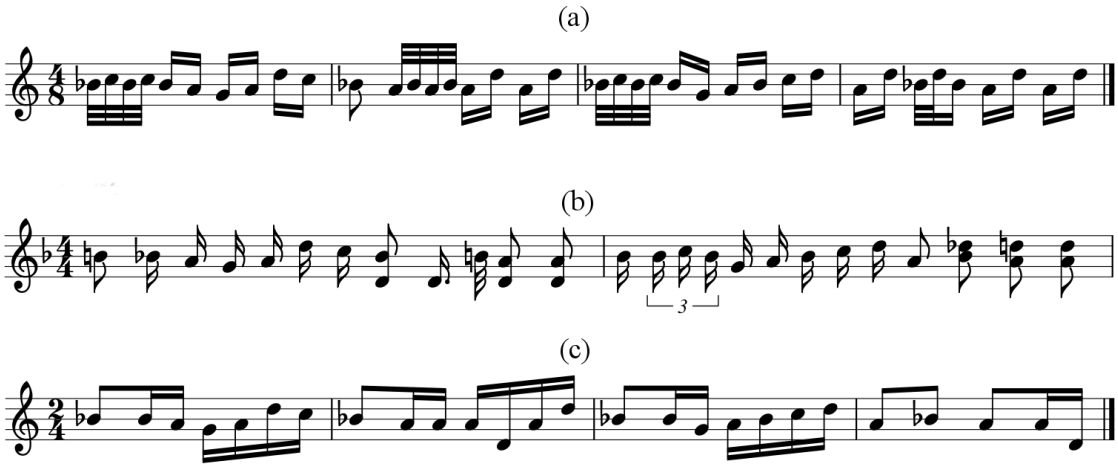

Figure 1. Transcriptions for bars 85-88 of segment 114 from the corpus of [8]. (a) Reference transcription. (b) Automatic transcription as presented to participant. (c) Manual transcription created by one of the expert participants.

3.2 Participant transcription evaluation

Table 3 depicts the transcription error rates obtained for each piece, averaged over all participants. The lowest mean error rate E_{mean} is obtained for Excerpt 6, and the highest for Excerpt 3, which is consistent with the ranking obtained from the two automatic transcription algorithms.

In Table 3, error rates with statistically significant differences (based on one-sample t-tests) to the error rates obtained from the ScoreCloud AMT (Table 1) are underlined. In addition, significantly lower error rates are emphasized using bold numbers. For the overall average (last row of Table 3), the only significant differences are increases in error rates (E_e, E_m) for the participant transcriptions, which indicates that participants' transcriptions include a larger number of extra and missing notes compared to the ScoreCloud transcriptions. Regarding significant changes of error rates for the individual pieces, four out of five for E_p, three out of five for E_{on}, and one out of two for E_{mean} are decreases in error rate. This indicates that at least some participant transcriptions were more consistent with the reference regarding meter (E_{on}) and pitch (E_p), compared to the ScoreCloud automatic transcriptions.

The inconsistent tendencies observed for the various metrics depicted in Table 3 motivate to investigate further which of the metrics most accurately reflect the notion of the quality of a transcription. Based on the quality ratings that the participants provided for each AMT, a conclusion was obtained which of the five error rates depicted in Tables 1 to 3 most correlated with the rated quality of a transcription. The highest correlation with the participants' stated AMT quality ratings was obtained for E_{mean} ($r = -0.91, p = 0.0019$). Correlations for E_p and E_{on} were still significant ($p < 0.05$), but correlations with both E_e and E_m were not. This motivates to focus on E_{mean} as the main metric for the rating of transcription quality in this paper, and to rather consider E_e and E_m as indicators for the chosen level of detail in the manual transcription.

Excerpt#	E_{p}	E_{e}	E_{m}	E_{on}	E_{mean}
1	14.58	42.15	24.42	40.85	30.50
2	**4.76**	20.07	15.37	**21.49**	15.42
3	**7.97**	54.05	48.45	45.83	39.07
4	18.16	38.50	35.61	44.13	34.10
5	19.52	33.75	16.52	37.16	26.74
6	**1.75**	18.84	11.50	16.90	12.25
7	17.53	15.33	30.68	**41.20**	26.18
8	**20.25**	37.81	13.97	**47.22**	**26.81**
Average	12.96	32.75	24.47	36.70	26.72

Table 3. Average error rates over all participant transcriptions. Underlined values emphasize statistical significant difference to the value in Table 1 (one-sample t-test, $p < 0.05$); bold values emphasize significant decrease in error rates over AMT.

3.3 Differences depending on the subject

Group	T(s)	Eff.	E_p	E_e	E_m	E_{on}	E_{mean}
Experts	430.33	4.1	**9.71**	31.58	28.60	**28.49**	24.60
Other	803.96	5.4	13.99	**33.12**	23.16	39.29	**27.39**

Table 4. Comparison of mean transcription times, T(s), rated efforts (Eff.), and error rates between experts and other participants. Statistically significant differences between experts and others are underlined (Welch's t-test, $p < 0.05$), and significant differences to average error rates in Table 1 are emphasized using bold numbers.

Based on the participant information, participants were grouped into experts and non-experts. The group of experts comprises three participants, who were either instructors of transcription courses, or had several decades' experience in transcribing folk music. Table 4 depicts the mean transcription times, rated transcription efforts, and error rates obtained from the transcriptions of these two groups. Whereas the expert group's transcription times

and effort ratings were significantly lower, the results regarding the quantitative error rate metrics remain inconclusive. Only regarding the onset error rate (E_{on} - which assesses the metrical correctness of the transcriptions when compared with the reference), the expert group had significantly better values that the non-expert group. In comparison with the average error rates obtained using the ScoreCloud algorithm (Table 1), the other transcribers have significantly higher error rates for E_e and E_{mean}, whereas the experts have significantly lower error rates regarding E_p and E_{on}. This implies that the tendency towards decreased error rates in the latter two metrics observed in Table 3 is more emphasized among the expert group.

3.4 Differences between editing and manual transcription

Case	T(s)	Eff.	E_p	E_e	E_m	E_{on}	E_{mean}
AMT	733.14	4.98	11.91	29.51	21.58	36.02	24.75
Man.	695.44	5.20	14.02	35.99	27.35	37.37	28.68

Table 5. Comparison of mean transcription times, T(s), rated efforts (Eff.), and error rates between AMT editing and manual transcription. None of the differences between the cases were found statistically significant (Student's t-test, all p-values > 0.18).

Table 5 shows the error rates over all participant transcriptions when editing automatic transcriptions as a starting point and when carrying out manual transcriptions, respectively. In order to address the question if any difference in the quality of the obtained participant transcriptions exists between the manual transcriptions and the editing of the automatic ones, the distributions of the error rates from the two cases were compared using two-sample t-tests. However, no significant differences were observed, indicating that the transcription times, efforts, and quality neither improved, nor deteriorated by using automatic transcriptions as a starting point. Significantly decreased variances were observed for two error rates (E_p, E_m) when using the AMT as a starting point (two-sample F-test, $p < 0.05$). This indicates that the usage of AMT as a starting point – at least in our experiments – led to transcriptions that are more similar, which may be interpreted as a bias imposed on the transcribers by using the AMT.

Since we observed in Table 4 that experts transcribe generally faster, we investigated if some gain in using AMT in terms of transcription time and effort can be observed at least for particular participants. To this end, we computed the relative changes in transcription time comparing manual transcription with editing per participant, and the absolute differences in the effort ratings per participant. For each participant, negative values for transcription time difference indicate that editing AMT was faster, whereas negative values for rating difference imply less effort when editing AMT. Figure 2 shows a correlation between the differences in effort and transcription times, which indicates that participants who had a tendency to rate a decreased effort in editing tend also to be those spending less time when editing. The approximately equal number of points in the lower-left and the upper-right quadrant reflects the absence of an overall effect of using AMT on transcription times and effort. The fact that all three expert transcribers (emphasized by circles) spend longer when editing AMT may indicate that providing AMT is not of practical use for experienced transcribers, a point further discussed in the following Section.

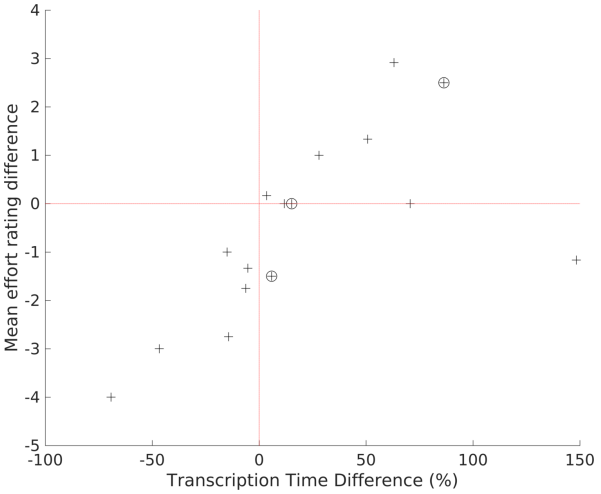

Figure 2. Scatter plot of absolute differences in the effort ratings and change in transcription time. Expert participants are emphasized by a circle.

3.5 Qualitative Results

The experiment was designed to be flexible in terms of providing the participants with exactly those tools for transcription that they would normally use outside of the context of this experiment. Therefore, the choices regarding these tools provide valuable insights into the transcription practice in the context of musicology. Out of the 16 participants, eight decided to transcribe the segments on paper, using mainly the software *Transcribe!* as a tool to loop certain phrases, and to decrease the speed of the playback. The other eight transcribers used notation software, four of them *MuseScore*, and four of them *Sibelius*; the latter exclusively applied by transcribers based in the UK, which indicates differences between the transcription practices based on the local musicology education.

Even though many participants transcribed on paper, all but one participant provided a positive response to the question if AMT tools are able to provide a valuable starting point for a manual transcription. Four of the participants expressed their opinion that use of AMT would be helpful mainly for inexperienced transcribers, and another four explicitly mentioned the potential to save time when using AMT.

A thematic analysis [4] was applied to the questionnaire responses in order to obtain the main reasons for criticism and appraisal of AMT. Four main themes emerged as depicted in Figure 3, three expressing criticism, and one expressing the value of AMT. Most frequently, participants

criticized rhythmic aspects of the AMT, referring most of the time to note durations contained in the AMT. The participants' second most frequent criticism concerns the omission of notes sounding in the recordings, and the addition of notes not heard by the transcriber, with omissions and additions being similarly frequent in the comments. Finally, participants criticized the simultaneous notation of notes, resulting from the notation either of accompaniment notes played on the lute or of overtones of the main instrument. Despite the fact that our questions focused on criticism of the AMT, one positive theme emerged as well, as participants emphasized the value of the AMT to obtain an understanding of the overall shape of the melody.

Summing up the qualitative observations, the most important finding is that current AMT technologies in the context of musicology may have a generally positive value for inexperienced transcribers in terms of pitch information. This value is, however, diminished by the frequent problems related to rhythm, addition/omission of notes, and poor separation of main melody and accompaniment.

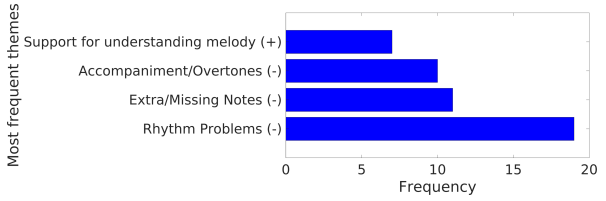

Figure 3. Most frequent themes in the discussions of qualities (+) and problems (-) of the AMT.

4. DISCUSSION

There are several aspects of the proposed study that need to be taken into account before making any claims on the usefulness of AMT in the context of (ethno)musicology. Firstly, the sample size in terms of participants is relatively small, which indicates the difficulty in locating subjects who are trained in transcription and are willing to work with automated methods. It is also difficult to rate the participants' transcription skills: future work could enlist the assistance of transcription instructors and to ask them to rate the participants' transcriptions. Another aspect to take into account is the bias introduced by the AMT system when asking participants to edit transcriptions.

The quantitative evaluation metrics proposed in [15], which were used as part of this study are not error-free: they rely on automatic score-to-score alignment that has been designed to align performance MIDI with reference scores. In particular, the symbolic alignment step could fail in the case of "abstract" transcriptions which could only focus on transcribing notes that are on strong beats, e.g. ignoring any passing notes. Therefore, additional work can be done towards improving the automatic symbolic alignment approach of [16] towards supporting the alignment automatic and manual transcriptions with reference scores.

Additionally, it should be stressed that the present study is focused on the usefulness of AMT in the context of (ethno)musicology, thus not taking into account potential uses of AMT in other application domains. The focus of this study was also on monophonic transcription (despite the presence of polyphony in certain segments); therefore, the usefulness of polyphonic AMT, and also of multiple-instrument AMT technologies, remains to be explored.

Finally, the question posed in the study on the value of AMT could be viewed as suggestive and could have biased participants towards providing a positive answer. There might also be a bias on participants who agreed to take part in the study, since their participation could indicate their general interest into the subject of automatic music transcription, and more generally on the use of technology in the transcription process.

5. CONCLUSIONS

This paper presented a user study on AMT in the context of ethnomusicology. Participants were asked to manually transcribe four segments of folk dance tunes, and to transcribe four different segments of the same style using the output of an AMT system as a starting point. Quantitative analysis shows: a comparative quality between automatic and manual transcriptions; differences between expert and non-expert transcribers, in terms of the time required to carry out transcriptions and also on the metrical quality of the resulting transcriptions; a correlation between the differences in stated effort between manual and edited transcriptions and transcription times; and a correlation between AMT quality ratings and some of the employed quantitative metrics. Finally, qualitative results show support for AMT to obtain an understanding of the overall melodic shape, although combined with criticism of the AMT related to harmonic errors, missing/extra notes, and rhythmic problems. Importantly, however, using an AMT output as a starting point for a transcription did not result in any quantifiable differences regarding quality, transcription time, or effort.

Future work will investigate similarity/dissimilarity of participants' transcriptions, in particular between those by the expert participants and the reference transcriptions, and will liaise with transcription instructors towards grading the resulting transcriptions. We will investigate ways to improve the quantitative metrics of [15] towards a more robust symbolic alignment of automatic and manual transcriptions with reference scores, and conduct evaluations using the newly-proposed metrics of [13]. Exploring whether the conclusions of this paper hold more broadly across other AMT systems and musical repertoires will have important impact on AMT research and on the applicability of AMT in ethnomusicology. We believe that this paper provides a viable and effective framework for user-based evaluation of AMT methods.

6. ACKNOWLEDGMENTS

EB is supported by UK RAEng Research Fellowship RF/128 and a Turing Fellowship. AH is supported by NordForsk's Nordic University Hub "Nordic Sound and Music Computing Network - NordicSMC" (proj. nr.

86892). The authors would like to thank Sven Ahlbäck, Stephen Cottrell, Emir Demirel, Michael Hagleitner, Eita Nakamura, August Schmidhofer, Richard Widdess, and Adrien Ycart for support and feedback.

7. REFERENCES

[1] Website of the Crinnos project. `http://crinnos.ims.forth.gr`. Accessed: 2019-03-26.

[2] M. Bay, A. F. Ehmann, and J. S. Downie. Evaluation of multiple-F0 estimation and tracking systems. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 315–320, 2009.

[3] E. Benetos, S. Dixon, Z. Duan, and S. Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30, 2019.

[4] V. Braun, V. Clarke, N. Hayfield, and G. Terry. Thematic analysis. In Pranee Liamputtong, editor, *Handbook of Research Methods in Health Social Sciences*, chapter 48, pages 843–860. Springer, 2019.

[5] R. G. C. Carvalho and P. Smaragdis. Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 151–155, 2017.

[6] A. Demetriou, M. A. Larson, and C. Liem. Go with the flow: When listeners use music as technology. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 292–298, 2016.

[7] P. E. Goddard. A graphic method of recording songs. In *Anthropological papers written in honor of Franz Boas*, page 137. New York, 1906.

[8] A. Holzapfel and E. Benetos. The Sousta Corpus: beat-informed automatic transcription of traditional dance tunes. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 531–537, 2016.

[9] J. H. Lee and S. J. Cunningham. The impact (or non-impact) of user studies in music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 391–396, 2012.

[10] J. H. Lee and S. J. Cunningham. Toward an understanding of the history and impact of user studies in music information retrieval. *Journal of Intelligent Information Systems*, 41(3):499–521, 2013.

[11] Z. Li, Q. Xiang, J. Hockman, J. Yang, Y. Yi, I. Fujinaga, and Y. Wang. A music search engine for therapeutic gait training. In *ACM International Conference on Multimedia*, pages 627–630, 2010.

[12] G. List. The reliability of transcription. *Ethnomusicology*, 18(3):353–377, 1974.

[13] A. McLeod and M. Steedman. Evaluating automatic polyphonic music transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 42–49, 2018.

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

[15] E. Nakamura, E. Benetos, K. Yoshii, and S. Dixon. Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 101–105, 2018.

[16] E. Nakamura, K. Yoshii, and H. Katayose. Performance error detection and post-processing for fast and accurate symbolic music alignment. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 347–353, 2017.

[17] M. Piszczalski and B. A. Galler. Automatic music transcription. *Computer Music Journal*, pages 24–31, 1977.

[18] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza. An end-to-end framework for audio-to-score music transcription on monophonic excerpts. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 34–41, 2018.

[19] M. Schedl, A. Flexer, and J. Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539, 2013.

[20] J. Stanyek. Forum on transcription. *Twentieth-Century Music*, 11(1):101–161, 2014.

[21] D. Weigl and C. Guastavino. User studies in the music information retrieval literature. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 335–340, 2011.