OXFORD UNIVERSITY PRESS | Digital Scholarship in the Humanities

Humanities and Engineering Perspectives on Music Transcription

Journal:	*Digital Scholarship in the Humanities*
Manuscript ID	DSH-2021-0041.R2
Manuscript Type:	Full Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Holzapfel, Andre; KTH EECS, Division of Media Technology and Interaction Design Benetos, Emmanouil; Queen Mary University of London Killick, Andrew; The University of Sheffield Widdess, Richard; SOAS University of London
Keywords:	music transcription, automatic music transcription, ethnomusicology, music information retrieval, music notation

SCHOLARONE™ Manuscripts

Humanities and Engineering Perspectives on Music Transcription

Andre Holzapfel, KTH Royal Institute of Technology, Sweden

Emmanouil Benetos, Queen Mary University of London, UK

Andrew Killick, University of Sheffield, UK

Richard Widdess, SOAS University of London, UK

For Peer Review

Abstract

Music transcription is a process of creating a notation of musical sounds. It has been used as a basis for the analysis of music from a wide variety of cultures. Recent decades have seen an increasing amount of engineering research within the field of Music Information Retrieval (MIR) that aims at automatically obtaining music transcriptions in Western staff notation. However, such approaches are not widely applied in research in ethnomusicology. This paper aims to bridge interdisciplinary gaps by identifying aspects of proximity and divergence between the two fields. As part of our study, we collected manual transcriptions of traditional dance tune recordings by 18 transcribers. Our method employs a combination of expert and computational evaluation of these transcriptions. This enables us to investigate the limitations of automatic music transcription (AMT) methods and computational transcription metrics that have been proposed for their evaluation. Based on these findings, we discuss promising avenues to make AMT more useful for studies in the Humanities. These are, first, assessing the quality of a transcription based on an analytic purpose, second, developing AMT approaches that are able to learn conventions concerning the transcription of a specific style, third, a focus on novice transcribers as users of AMT systems, and, finally, considering target notation systems different from Western staff notation.

Key words: music transcription, automatic music transcription, ethnomusicology, music information retrieval, music notation

1 Introduction

Music transcription is a process of creating a notation of musical sounds, with music notation being the representation of musical sound through some other medium. This can take many forms, but the present article, and the field of Music Information Retrieval (MIR) in general, is concerned with static, visual representations of sound and specifically with Western staff notation. When performers use notation, it serves them as a set of instructions for producing certain sound-patterns; but in music transcription, the notation is produced from sound (usually from a recorded performance) rather than the reverse. Transcriptions may be made either by human transcribers or by machines such as computers, and may be human-readable (*e.g.* staff notation, guitar tab) and/or machine-readable (*e.g.* piano roll, MIDI file). If human-readable, they may be used as instructions for performers, but in ethnomusicology they are more often used to support analysis of music that was not previously notated (or not notated in a way that is useful for the analysis).

In ethnomusicology and its parent discipline, comparative musicology, transcription long played a central role (Ellingson, 1992). Initially it was the only means of communicating the sound and style of the music to readers who had never heard it. Over time, recordings took over this function, and transcription became more restricted to its other role, that of supporting and illustrating analyses (Nettl, 2015, p.75). As the comparative musicologists' agenda of arranging all music into global evolutionary schemes gave way to the more locally situated, fieldwork-based inquiries of ethnomusicologists, the role and focus of music analysis became more variable. For some, the whole project of transcription and analysis, and especially transcription into Western notation, felt uncomfortably close to the colonialist legacy that ethnomusicology was trying to leave behind (see Marian-Bălaşa, 2005). For others, transcription and analysis still had a part to play in characterising particular forms of music and thus helping to answer ethnomusicology's big question of why people make and use the particular forms of music that they do. But those who still practised transcription and analysis used it to address specific analytical questions relating to the particular music they were concerned with, and developed approaches and solutions that were tailored to these questions and often quite personal to the transcriber. Few were content with staff notation as used by classical composers: even the early comparative musicologists had recognised that modifications were necessary for transcribing music from outside that tradition (Abraham and Hornbostel, 1910). Some turned to technologies for producing automatic transcriptions that transcended the limitations of staff notation and human listeners (Metfessel, 1928; Seeger, 1958).

Research on computational - as opposed to more broadly technological - methodologies for music transcription emerged in the 1970s (*e.g.* Moorer, 1975). Research on the topic has intensified during the last two decades in the context of MIR research, resulting in development and increasing refinement of methods for automatic music transcription (AMT). Most of these methods assume Western staff notation as the final goal of the transcription process, and the related problems of this notation format have not been discussed to a large extent within MIR. Besides

AMT approaches, various metrics have been proposed within MIR that aim at the evaluation of the quality of AMT-produced transcriptions based on comparison. So far, ethnomusicologists did not make much use of AMT methods or evaluation metrics. Thus, a central question for the present paper is whether, and how, AMT might be made more useful to ethnomusicologists. The potential may lie within, for instance, the support of inexperienced transcribers, the discovery of melodic motives in large corpora, or the visualization of longer recordings in the form of notation. We believe that an increased awareness of MIR research among ethnomusicologists, and conversely of problems long discussed in ethnomusicology among MIR researchers, will help to provide answers to our central question and lead to an improved and more widely applicable AMT technology.

We therefore approach our central question through a combination of perspectives. In the following Section, we contrast perspectives on transcription in MIR and in Ethnomusicology, in order to identify aspects of proximity and divergence between the fields. In Section 3, we present our method that, first, extends a previous user study (Holzapfel and Benetos, 2019), which collected a large number of manual transcriptions for a collection of traditional dance tune recordings. Second, our method employs a combination of expert and computational evaluation of these transcriptions. This enables us to investigate the limitations of computational transcription metrics and AMT methods in Section 4. In Section 5, we discuss promising avenues to make automatic transcription more useful for music studies in the Humanities.

2 Background

2.1 Music Transcription in Ethnomusicology

If they transcribe music at all, ethnomusicologists usually aim to document a specific performance on the basis of a recording (so-called "descriptive music-writing"), rather than to provide a model for performance ("prescriptive music-writing"; Seeger, 1958). An ethnomusicologist may choose to make a "close transcription" that includes details of playing style - melodic or rhythmic nuances, ornaments, timbral effects *etc*; or a "broad transcription" in which such details are omitted in order to show the basic structure of the music. However, the distinction between "basic structure" and "details" is by no means always easy to make, and may require "insider" knowledge of the musical system in question. Ethnomusicologists can also choose an approach that is more or less "etic" (cf "phonetic"), whereby whatever is audible is included in case it should turn out to be significant, or "emic" (cf "phonemic"), including only those categories and distinctions considered significant in the culture concerned. The former was typical of early investigators working on sound recordings at a distance from the field; later, fieldwork, performance study and collaboration with performers made it possible to distinguish "structure" and "details" on the basis of "insider" knowledge of the musical system in question, and transformed transcription into a representation of performance in cultural context (Ellingson, 1992; Widdess, 1994).

Although characterised as an "unscientific" procedure (Seeger, 1958), most

For Peer Review

transcribers have continued to employ staff notation, with or without additional symbols or other modifications (Abraham and Hornbostel, 1910); automatic graphic representations such as the "melogram" (Seeger, 1958) offer an alternative with both the advantage and the disadvantage that they bypass the interpretive processes of human cognition, such as the ability to distinguish multiple simultaneous streams of sound (Jairazbhoy, 1977).

In the past, considerable importance was attached to the evaluation of transcription, in terms of precision, significance and style (List 1963, 1974; England, 1964). Current debates in ethnomusicology rarely depend on transcription or analysis, focussing instead on social and political issues such as identity, power relations, colonialism, globalisation *etc*, and on musical experience and behaviour at individual or community levels. Transcription in Western notation can be viewed as an exercise in measuring other musics against Western norms, not necessarily favourably, or as a neo-colonialist imposition of those norms on other musics, or simply as an outdated methodology. In recent years, however, there has been growing interest in analytical approaches to world music outside mainstream ethnomusicology (*e.g.* Tenzer, 2006~~5~~; Tenzer and Roeder, 2011~~0~~), including computational or cognitive approaches; and this has encouraged both the continuation of conventional transcription, and new approaches to visual representation of music (*e.g.* Killick, 2020).

Meanwhile, the uses of transcription have diversified, both within and across disciplines. As noted (perhaps with some exaggeration) by Regine Allgayer-Kaufmann (2005, p.71), "ethnomusicologists today do not at all use their transcriptions for exploring, *i.e.*, they do not explore with their eyes; instead, they explore with their ears and their body. The purpose of transcriptions has turned out to be mainly to communicate knowledge that was obtained by these other means." This is what Ter Ellingson (1992, p.141-2) called "conceptual transcription", in which "essential features [of the musical system] are presumed to be already known through fieldwork, performance lessons, study of traditional written and aural notations and learning and leadership processes. The transcription then becomes a means, not of discovering, but of defining and exemplifying the acoustical embodiment of musical concepts essential to the culture and music." Few ethnomusicologists would now embrace what Ellingson calls "classical Hornbostelian transcription" (ibid.), in which the sounds of any music are first transcribed according to a standardised procedure and only then subjected to analysis.

Yet some ethnomusicologists and other exponents of sound analysis do continue to find the process of transcription valuable as a means of discovering aspects of musical sound organisation that might not have been noticed through other means. This is evident in the "Forum on Transcription" convened by Jason Stanyek (2014), which presents edited transcripts of conversations with pairs of scholars working in six areas: ethnomusicology, song lyrics, popular music studies, animal vocalisations, the culture industry and music information retrieval. One of the ethnomusicologists, Tara Browner, suggests that transcription "provides a way to engage with music with a kind of depth and intensity that, just listening to it, you don't get" (p. 112), and the other, Michael Tenzer, agrees: "there is much about music that you can't learn until you write it down" (p. 119). Similarly, song lyric

For Peer Review

specialist Dai Griffiths reports that when doing transcription “one always comes out noticing things one hadn’t spotted in the listening” (p. 130), and popular music analyst Anne Danielson finds that “through transcription work… my listening has become more precise” (p. 134). But each participant in the forum adopts a different approach to notation, some based on staff notation with various modifications or additions and others using entirely different graphic representations of sound. This reflects the different analytic and communicative agendas that the scholars have set themselves; for as Bruno Nettl has observed, “rather than simply providing a visual record of music, transcription has been used more to solve specialized problems, and for this, a variety of techniques, mechanical and manual, have been developed” (2015, p.86). Even among ethnomusicologists using a common system, such as staff notation, the quality of a transcription tends to be judged according to the specific analytical purpose that it is intended to support. Consequently, different transcriptions of the same music can be equally “good”, and there is no single “correct” transcription of a given performance (England, 1964; Nettl, 2015).

2.2 Music Transcription in MIR

In the MIR field, automatic music transcription (AMT) is typically defined as the process of converting an audio recording or audio stream into some form of human- or machine-readable notation (Klapuri and Davy, 2006). The first approaches for automatic transcription of musical audio to machine-readable notation originated from the 1970s (*e.g.* Moorer, 1975), with the problem gaining attention from the early 2000s with the development of signal processing and pattern recognition methods for analysing audio signals. AMT is generally considered to be a fundamental problem in the MIR field (Serra *et al.*, 2013), with numerous applications that include but are not limited to the creation of user-facing software systems for converting audio into Western staff notation (see Benetos *et al.* (2019) for an overview of commercial software for automatic music transcription), to the use of AMT technologies as a way of creating a compact and meaningful representation of an audio performance, to be used as a descriptor in downstream MIR applications (*e.g.* the automatic recommendation of music~~music recommendation, fingerprinting, audio tagging~~). Beyond the MIR field, automatic music transcription technologies have found applications in computer music (*e.g.* for automatic music accompaniment), music education (for automatic instrument tutoring), and computational musicology - the study of music with computational modelling and simulation.

Depending on the application and music style, works in the AMT literature have focused on addressing specific MIR tasks. These include but are not limited to melody transcription (in the presence of either monophonic or polyphonic[i] music signals - Salamon *et al.*, 2014), automatic transcription of polyphonic music (Benetos *et al.*, 2019), lead sheet transcription (*e.g.* melody and bass line/chord estimation - *e.g.* Ryynänen and Klapuri, 2008), lyrics transcription (*e.g.* Mesaros and Virtanen, 2010), and drum transcription (Wu *et al.*, 2018). Another dimension of AMT is on the form of the desired output representation: many works have focused on producing an output representation in terms of active pitches over physical time (often in the form of a *piano-roll representation* or MIDI file); more recently, works in AMT have

For Peer Review

attempted to produce a *complete transcription* in the form of a music score in Western staff notation (*e.g.* Román *et al.*, 2019). The latter assumes the automatic recognition of multiple concurrent pitches, their respective onsets and offsets with respect to a metrical subdivision (as opposed to physical time), instrument identities, time signature, key/mode, dynamics, phrasing/grouping, voice/staff separation, and estimation of expressive information (*e.g.* rubato, ornaments) amongst others.

The evaluation of AMT methods is inherently linked with the output representations these methods produce, and all evaluations involve a comparison between an output automatic transcription with a "reference" transcription. For approximately 15 years (2000-2015), the vast majority of AMT systems focused on producing output representations related to pitch and physical time, either in the form of detected pitches over small time segments or in the form of lists of notes characterised by a pitch, onset time, and offset time (in seconds). Such systems were typically evaluated using the benchmark metrics proposed as part of the Music Information Retrieval Evaluation eXchange (MIREX) public evaluation campaigns (Bay *et al.*, 2009). Informal observations have been made on how the evaluation of systems with respect to producing lists of notes (typically referred to as *note-based evaluation*) is more perceptually relevant compared to the evaluation of groups of pitches over small time segments (typically referred to as *frame-based evaluation*). However, community efforts towards proposing evaluation metrics for AMT that are linked with how humans would judge transcriptions are fairly limited and have not reached a broad consensus. An early perceptual study by Daniel *et al.* (2008) proposed evaluation metrics that take some common local errors related to automatic transcription into account (such as note insertions and octave errors), but do not take into account aspects related to meter or tonality. Recently, for the more relevant task of complete transcription, Nakamura *et al.* (2018) proposed a set of evaluation metrics that address local errors in a musical score (*e.g.* insertions, deletions). Higher-level evaluation metrics have also been proposed which draw knowledge from music theory and focus on typesetting (Cogliati and Duan, 2017; McLeod and Steedman, 2018), although their links with human assessment of music transcription are unclear. Finally, Ycart *et al.* (2020) proposed a single evaluation metric for AMT systems that produce outputs in physical time following perceptual evaluations; the focus of the study was however only on piano music, and the metric's ability to generalise to other instruments is as yet unclear.

An important issue of AMT methods refers to their implicit or explicit algorithmic biases. The development of AMT methods using supervised machine learning methods (which are the most commonly used methodologies nowadays) assumes the presence of a "target" or a "reference" transcription that the transcription system should try to approximate. This however can bias any resulting systems to music recordings for which notation exists -- most commonly pieces that have been composed in written form. A second point which is also linked to the previous one refers to the availability of data to train AMT systems. The availability of music recordings with corresponding annotations of notes and musical events over physical time is scarce, since the process of producing such annotations is an extremely laborious task (Su and Yang, 2015). This has led to the creation of datasets for AMT research that have been created using acoustic musical instruments that can automatically create such annotations - most commonly acoustic pianos with sensors that can capture music performance characteristics, such as Yamaha or Bösendorfer Disklavier pianos. This has led to a large imbalance in the diversity of AMT datasets towards the piano, *e.g.* through the commonly used MAPS (Emiya *et al.*, 2010) and MAESTRO (Hawthorne *et al.*, 2019) datasets, and therefore to the creation of AMT methods that are focused on piano transcription. Combined with the greater availability of reference scores for Western art music compared to other musical cultures or styles, this has led to the vast majority of AMT systems being exclusively focused on transcribing Western art music performed on the piano. This stylistic focus leads to the fact that the analytic purpose of a transcription has largely not been taken into account in MIR until now.

On the value of automatic music transcription methods for studies in musicology, in recent years studies have been attempted in the intersection of digital and computational musicology and MIR, mostly aiding musical study for large corpora. For example, Tidhar *et al.* (2014) used AMT methods as a first step towards a large-scale analysis of temperament profiles and temperament trends over time in harpsichord recordings. Similarly, AMT methods have been used towards estimating trends in tuning frequencies in the context of archival Western art music recordings (Abdallah *et al.*, 2017). In the context of jazz studies, melody estimation algorithms have been used to group melodic patterns in jazz improvisation (Höger *et al.*, 2019). In the field of computational ethnomusicology, AMT methods have been used in the context of Turkish makam music (Benetos and Holzapfel, 2015), where a discrepancy between music theory and practice was observed with respect to the pitch values implied by notation. Finally, melody estimation methods have been used to infer music similarity and dissimilarity in folk and traditional music recordings in a global sample (Panteli *et al.*, 2017).

2.3 Music Transcription in between the Fields

In ethnomusicology, mechanical devices for automatic transcription and analysis have been used since before the advent of computer technology (Ellingson, 1992; Cooper and Sapiro, 2006), but their application has been limited mainly to the analysis of pitch (tonometry) and melodic contour (melography). The melograph devices developed and advocated by Seeger (1958) and Hood (1971; 1993) were a response to the limitations of staff notation and manual transcription: a melogram or spectrogram could show pitch movement “between the notes” (Seeger, 1958), and other subtle nuances of timbre, loudness, and rhythm, which were presumed to be important indicators of style and cultural identity (“essential performance idioms”, Hood, 1993), but which were impossible to notate in staff notation, and liable to escape the culturally preconditioned ear of the Western transcriber. Other musicologists pointed to problems including the difficulty of reading spectrographic representations of sound, the inability of the technology to distinguish multiple simultaneous sounds and melodic lines, and inherent differences between human auditory perception and that of a machine (Jairazbhoy, 1977). Since the 1990s, the advent of digital sound analysis software for personal computers (now including the phonetics package *Praat*, and the music analysis programmes *Sonic Visualiser* (Cannam *et al.*, 2006) and *Tony* (Mauch *et al.*, 2015)) made the means to create a

For Peer Review

spectrogram or fundamental pitch trace more widely available than ever before; but it remains a specialised tool for very specific problems, used mainly in relation to monophonic music. In studies of Indian music, for example, sound analysis software has been used to represent elaborate pitch contours, clarify subtleties of pitch inflection between or around scale-degrees, and map rhythm against time in the absence of a clear metrical pulse (Rao and van der Meer, n.d.; Sanyal and Widdess, 2004), and to compare multiple renditions of the same melody, by the same performer on different occasions (Tallotte, 2017). Rao and van der Meer's "Automated transcription for Indian music" (AUTRIM) methodology, based on *Praat*, produces a precise, accurate and vividly realistic visual representation of the melody, with its complex vocal inflections. But it is not a method that can easily be employed by other ethnomusicologists, as it depends on specially created recordings, with sound isolation between the performers, and a degree of editorial intervention.

So far, automatic transcription into staff notation has not been widely employed, if at all, by ethnomusicologists. One possible reason for this may be that the process of equating the sounds of one musical system with the symbols of another is considered highly problematic by many ethnomusicologists, even though staff notation remains the most usual system for manual transcription. Indeed, the need for a global system of music notation, not prioritising any one musical tradition, has long been felt in ethnomusicology (Hood's "Laban solution", Hood, 1982), and such a system has recently been proposed (Killick, 2020). For ethnomusicological purposes, MIR approaches to AMT may in future need to be adapted to this or some other new notation, assuming that intrinsic bias towards any specific musical system can be avoided in the software design. Furthermore, the potentially diverse analytic purposes of individual transcriptions described in Section 2.1 question the concept of validation based on single reference transcriptions that is common in MIR. The present study will investigate such diversity in a larger corpus of transcriptions, and it will identify shortcomings of MIR evaluation measures by relating them to expert quality ratings.

3 Method

The methodology employed in this paper consists of four main steps, which will be discussed in the following subsections. First, a previously conducted user study (Holzapfel and Benetos, 2019) has been extended, in which a group of musicology students with experience in transcription were asked to compile transcriptions for a series of short music excerpts. Secondly, two senior ethnomusicologists assessed all the transcriptions available after the user study, which comprised not only the transcriptions by the students, but also algorithmic and pre-existing expert transcriptions. As a third step we investigate the correlations between the obtained expert assessments and a series of computational metrics. Finally, we analyse the discrepancies between human and automatic assessment.

3.1 User Study

3.1.1 Participants

Participants for the proposed study were recruited from the Institute of Musicology in Vienna (Austria), SOAS University of London (UK), City, University of London (UK), Queen Mary University of London (UK), and Royal Holloway, University of London (UK). In total, 18 participants participated in our study, ten male and eight 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. All participants filled consent forms, and the study was conducted following the Declaration of Helsinki and local ethics regulations.

The participants had 17 years of music training on average, with a standard deviation of 10 years. In terms of their interests, 8 participants closely identified with Western classical music, and 10 participants identified with world/folk/traditional music. In terms of their professional practice, 11 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, 7 with Transcribe!, 7 with Sibelius, and 2 with Sonic Visualiser.

3.1.2 Material

For this study, we use audio recordings and corresponding transcriptions to Western staff notation collected as part of the Crinnos project (Institute of Mediterranean Studies, 2005), which were also used as part of the *Sousta Corpus* for AMT research by Holzapfel and Benetos (2016). All recordings used in this study were recorded in 2004 in Crete, Greece, and all regard a specific dance called *sousta*. 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 transcribed in Western staff notation using time signatures of 2/4 or 4/8 by musicologists and local musicians (Institute of Mediterranean Studies, 2005; Andreoulakis and Petrakis, 2013), 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 bowed lute) playing the main melody, and usually two Cretan *laouto* (a long-necked plucked lute) playing the accompaniment.

Eight audio excerpts[ii] 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 3-4 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 in such a way as to provide study participants with a complete musical phrase, in order to aid transcription.

We did not assume that participants were familiar with the music culture used in this study. Therefore, one complete recording from the *Sousta Corpus* was also selected in order to familiarise participants with the music style prior to the start of the study.

3.1.3 Procedure

For each excerpt, participants were asked to "transcribe the basic melody as produced by the lyra, not the accompaniment (if any exists), and leave out minute transcriptions of embellishments". 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 eight excerpts per participant, presented in randomized order. We provided AMT outputs obtained from the state-of-the-art transcription software ScoreCloud[iii] in printed and machine-readable format for four of the excerpts chosen at random, to be used as a starting point for the transcription process. The other four excerpts were transcribed completely manually. The order of manual and AMT-informed transcriptions was interleaved, and participants were either asked to start transcribing their first segment manually or to edit an automatic transcription. The motivation for including these two transcription modes was to be able to investigate how AMT may influence the participants' transcriptions.

In the study 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. Following the study, a short conversation with participants took place, in order to obtain additional 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 staff notation using MuseScore.

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 the participant), manuscript paper, and a study questionnaire. Participants were video recorded in order to assist with the subsequent annotation process.

For Peer Review

3.2 Expert Evaluation

The user study described in Section 3.1 resulted in 116 transcriptions by the participants. Not all participants managed to complete transcriptions for all eight provided segments, but for each segment 13 to 17 transcriptions were obtained. In addition to these 116 transcriptions by our 18 participants, we added transcriptions that had been compiled by the ethnomusicologist of the Crinnos project. These transcriptions had been used as reference for evaluation by Holzapfel and Benetos (2019), and subjecting them to an evaluation by experts enabled us to investigate the quality of these transcriptions. Finally we added the automatic transcriptions obtained from the two algorithms used in our user study (ibid.). This results in a corpus of 140 transcriptions by 21 different transcribers, including the two algorithms as transcribers into the collection.

Two senior ethnomusicologists (the third and fourth authors of the present paper, AK and RW) were provided with all the 140 transcriptions along with the audio recordings of the eight segments. Since neither of the two ethnomusicologists have expertise in the particular style of music, they familiarized themselves with the music material using the training recording (see Section 3.1.2) and the eight short audio segments. After that, the experts were asked to assess the transcriptions for each segment. The experts conducted the process independently of each other. As a first step, the two experts ranked the eight segments in order of difficulty. They then ranked all transcriptions for each segment, placing the best transcription on top of the list. After completing the ranking for a segment, they provided a score for each transcription, which ranged from 1 (transcription "completely unrelated to recording") to 10 ("extremely accurate transcription"). In addition, experts provided information regarding the motivations for their ranking. They were asked to specify what made the high-ranked transcriptions better than the low-ranked, which aspects of the transcriptions they considered in the ranking, and what general problems they saw in the transcriptions.

3.3 Computational Analysis

Having obtained the expert evaluations for all transcriptions, we analysed the ratings by the experts under two main aspects. First, we compared the expert ratings with various MIR evaluation metrics for transcription assessment, described below. The goal of this process is to obtain a measure of how much various MIR metrics correlate with experts' evaluations. Since all MIR evaluation metrics involve comparison between a transcription to be evaluated and a reference transcription, the question needed to be addressed how to choose such reference(s). Whereas we previously (Holzapfel and Benetos, 2019) relied on the authority of the ethnomusicologists of the Crinnos project, in the present paper we are able to base the choice of references on the expert ratings. To this end, we chose the N transcriptions with the highest mean average rating by the two experts, with the value of N to be determined based on the distribution of these ratings. For a transcription to be assessed automatically using an MIR metric, we computed its comparisons with the N reference transcriptions of the segment ($N-1$ if the transcription to be assessed is among the N best-rated). Then we used the best metric value, motivated by the assumptions that the reference transcriptions are characterized by slight mutual differences, and that a good transcription may be most similar to one of the references.

To obtain an idea of the consistency of the reference transcriptions, we computed the mutual agreements between transcriptions using the MIR metric found most strongly correlated with the expert evaluations. The mutual agreements were computed in two groups for each segment: the N highest-rated and the N lowest-rated transcriptions. The main question to be explored here is whether mutual agreement of metrics correlates with the rating of the experts. Our hypothesis is that transcriptions that are highly rated by experts have a higher mutual agreement than lower rated transcriptions. This would imply that good transcriptions tend to be more consistent regarding note onset times, pitch values, and duration values.

Metrics: In the MIR literature, the performance of automatic music transcription systems is typically evaluated using metrics that compare the output of the transcription system with a "reference" transcription. Evaluations are typically carried out by comparing lists of notes in terms of pitch, onset, and offset in physical time, or by comparing binary "piano-roll" representations (see Section 2.2 for more information on AMT evaluations). However, such metrics are not suitable for evaluating transcriptions in staff notation, since converting a piano-roll representation to staff notation is a non-trivial process.

To that end, in this work we focus on metrics for evaluating transcriptions in staff notation. The first set of metrics was proposed in Nakamura *et al.* (2018) and was originally used for evaluating the performance of a system that automatically transcribed Western art music performed on a piano. The above-mentioned metrics first perform an automatic alignment of the automatically transcribed score to the reference score; following the alignment step, evaluation is carried out by identifying 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 derived: pitch error rate *Ep*, extra note rate *Ee*, missing note rate *Em*, and onset time error rate *Eon* (for all above metrics, smaller is better). The onset time error rate *Eon* is based on the minimum number of scale and shift operations for onset score times.

The second set of metrics considered was proposed in McLeod and Steedman (2018). This set of metrics jointly termed as MV2H includes figures of merit for evaluating multi-pitch detection, voice separation, metrical alignment, note value detection, and harmonic analysis performance. For the purposes of the present study, we focus on the multi-pitch detection F-measure *F-mp* and the note value recognition score *S-val* (for the above metrics, larger is better). The remaining MV2H metrics are not used since this study does not include multiple voices in the transcriptions and does not assume tonal harmony.

3.4 Identifying Limitations

Expert evaluations (Section 3.2) enable us to evaluate the quality of transcriptions by human and algorithmic transcribers, and the computational analysis described in Section 3.3 establishes connections between MIR metrics and human evaluation. Towards the main goals of this paper - analysing limitations of MIR metrics and automatic transcription outcomes - we specifically analyse a set of algorithmic transcriptions for their problems, and investigate cases in which we observe large discrepancy between MIR metrics and expert evaluations. These investigations will add to our previous findings, and identify a series of blind spots that until now have not been taken into account by evaluation metrics and AMT procedures.

4 Results

In order to establish the basis for our analysis, we investigate the consistency

For Peer Review

between the two experts in the ratings of the transcriptions. As depicted in Fig. 1, there is a large consistency between the two experts in their ratings, with a correlation coefficient of 0.754. The ratings cover a large range of the overall scale, with about ⅔ of the ratings being in the upper half of the rating scale.

Place Fig. 1 here: Distribution of the ratings between the two experts. Correlation coefficient is 0.754 (significant, p<10e-26).

4.1 Ranking of Transcribers

The ratings on the level of the 21 individual transcribers are summarized in Table 1. The transcribers have been ranked based on the average ratings obtained from both experts for all the transcriptions by the individual transcribers. In the upper and lower parts of Table 1, transcribers are emphasized who according to the standard deviation of their ratings are unlikely to produce a low-rated or high-rated transcription (with 5.5 being the border between low and high-rated). According to this criterion, there is a larger group of good transcribers, and a smaller group of poor transcribers, which reflects the general distribution of the ratings as stated above. Additional insight can be obtained by looking more closely at the list of identified good and bad transcribers. Among the seven good transcribers, Transcriber 21 is ranked on the seventh position, obtaining the lowest mean rating in this group. This transcriber was the source of the reference transcriptions that were used by Holzapfel and Benetos (2019). A closer investigation for the relatively low ranking for this transcriber revealed that both experts rated this transcriber lower than others based on the analytic purpose of the transcriptions stated in the user study to transcribe the main melody and leave out ornamentations. Transcriber 21 had compiled the transcriptions with a different analytic purpose, and had included a greater amount of ornamentations than the transcribers from our user study. This is consistent with the principle that the evaluation of a transcription has to take into account the stated analytic purpose, a principle firmly established in ethnomusicology but widely ignored in MIR.

On the lower end, only four transcribers can be identified that consistently provided transcriptions with low ratings. Two out of these are the algorithmic transcribers (19 and 20). We will turn our attention to a closer analysis for the motivations for these lower rankings in the final part of this Section. In general, the low ranking for the algorithmic transcribers confirms our previous results (ibid.) that neither of the two state-of-the-art algorithms can be expected to provide a high quality transcription for the present musical style, despite the fact that at least one of the algorithms (19) has been developed and evaluated using musical samples of the style that is the focus of this paper.

For Peer Review

Mean Rating	STD	Transc. ID
7.94	1.06	**11**
7.56	1.09	**16**
7.17	0.83	**14**
7.06	1.06	**3**
7.06	0.93	**18**
6.81	1.52	**12**
6.75	1.18	**21**
6.38	1.75	10
6.36	1.55	7
6.10	1.85	4
5.88	0.83	5
5.44	1.55	17
5.06	1.69	2
4.80	2.78	1
4.75	1.82	15
4.50	1.69	8
4.44	1.09	**20**
4.25	1.75	6
3.83	1.85	**9**
3.00	0.97	**19**
2.20	0.79	**0**

Table 1: Transcribers sorted by their transcription rating average (over all their transcriptions, and both rating experts). Transcribers are emphasized whose standard deviations do not transcend the border between a low-rated and a high-rated transcription (5.5).

For Peer Review

4.2 Bias through AMT

At this point, the ratings from the experts enable us to investigate an important question: Despite the apparent low quality of the automatic transcriptions, does using them as a starting point for transcription affect the outcome in any way compared to a completely manual transcription? As explained in Section 3.1.3, all participants were assigned to transcribe half of the segments completely manually, and the other half in a process that uses the AMT obtained from the algorithmic transcriber 20 as a starting point. Figure 2 presents the mean ratings for all segments separated into the two groups of editing and manual transcription. Whereas there is no significant difference in the means of the ratings, a general tendency towards smaller variance in the ratings for the editing case can be observed. This decrease has been found statistically significant for two segments (4 and 6, $p<0.05$, 2-sample F-test). As depicted in Fig. 2, this is caused by the fewer very low ratings for edited transcriptions (dashed-dotted box plots in Fig. 2). This finding is consistent with the previously observed decreased variance in transcription metrics (Holzapfel and Benetos, 2019) for the editing case. We previously associated this observation with an algorithmic bias introduced by the AMT, but now we arrive at an additional interpretation based on the expert ratings: providing an initial transcription even of low quality seems to help especially less skilled transcribers. To further support this interpretation, we divided the transcribers into two groups based on the mean rating in Table 1. The group of less skilled transcribers (mean rating below 5.5) were found to produce higher rated transcriptions when editing an AMT instead of manually transcribing ($p=0.037$, two-sample t-test), whereas for the higher skilled group no significant difference was found.

Place Fig. 2 here: Mean expert ratings for the eight segments, separated into two groups for complete manual transcription (solid-line box plots) and editing AMT (dashed-dotted-line box plots).

4.3 Human Compared to Computational Assessment

4.3.1 Corpus Perspective

After having obtained the above series of insights from the expert ratings, we now proceed to investigate correlations between these human ratings and ratings that can be automatically derived using the MIR metrics identified as appropriate for our analysis (see Section 3.3). To this end, we follow the procedure described in Section 3.3 to compute the metrics for all individual transcriptions. In order to define a set of reference transcriptions for each segment, we chose the set of N=4 transcriptions for each segment that obtained the highest average ratings from the two experts. This set of reference transcriptions is related to average ratings of 7 or larger, and will be used in the remainder of this Section to compute MIR metrics for a transcription to be assessed. To simplify representation in the remainder of the paper, all metrics have been converted to the range of the expert ratings (1-10), and metrics that assign low values for high quality were inverted. The correlation coefficients and p-values between the metrics and the average expert ratings are listed in Table 2. In the rightmost column, the two strongest correlated metrics have been combined, achieving a further improved correlation. This combined metric takes into account complementary information of onset time errors (*Eon*) and of the f-measure of pitch detection (*F-mp*), indicating that this combination leads to the highest correlation with human ratings. This finding confirms the lack of correlation between metrics that focus on added or missing notes (*Ee*, *Em*) and human quality ratings, which was observed in Holzapfel and Benetos (2019).

Metric	*Ep*	*Ee*	*Em*	*Eon*	*F-mp*	*S-val*	*Eon+F-mp*
R	0.485	0.068	0.337	0.616	0.549	0.341	0.682
p-value	<1e-10	.427	<1e-06	<1e-17	<1e-13	<1e-06	<1e-20

Table 2: Correlation coefficients and p-values between computational metrics and average expert ratings.

It is interesting to observe that the correlation between combined metric and mean expert ratings (.682) is still lower than the correlation between the two human experts (.754). This observation motivates our investigation of cases where ratings and metrics diverge as a starting point to identify blind spots in the current MIR evaluation metrics. A more advanced regression to obtain a further improved combined metric was not conducted on our style-specific corpus, but we consider it as a valuable investigation for future research based on a larger and more diverse dataset.

4.3.2 Close Analysis

We can obtain insights into the differences between the ratings by human experts and computational metrics by, first, analysing the criteria stated by the experts and comparing them with the aspects metrics take into account. Both experts state that the analytic purpose motivates the ranking, in which melodic criteria are primary (RW), and AK considered it essential that transcriptions capture a melodic idea. Both agree that high ranking necessitates precision and detail in pitch and rhythm, a clear distinction between main melody notes, ornaments and accompanying sounds, readability through appropriate use of notational conventions (*e.g.* beaming and use of accidentals), and use of expressive signs (*e.g.* bowing markings and indications of vibrato). Notational conventions and use of expressive signs are aspects ignored by all existing computational metrics, and therefore constitute two blind spots of MIR evaluation when applied in the context of evaluating transcriptions.

For Peer Review

As a second way to investigate which aspects of a transcription play a larger role for the human experts than for the computational metric, we selected five transcriptions with large discrepancy between expert ratings (avg.: 4.00) and the combined computational metric (avg.: 9.31). Both experts provided detailed explanations of the problems that they identified in these transcriptions, and the first author analysed the obtained texts. The problems were grouped into themes, and on a higher level these themes were assigned to the categories of problems related to rhythm/meter, pitch, and notation. For instance, wrong note durations were regarded as rhythm/meter problems, whereas missing notes were considered to cause problems of both rhythm/meter and pitch. In total, 14 problems related to rhythm/meter were identified, but only 6 problems related to pitch, and one problem related to notation (which was the wrong use of accidentals). Hence, at least in our body of samples, the discrepancy is related to the computational metric being less sensitive to rhythmic problems than the human experts.

Two examples of transcriptions with divergence between expert ratings and computational metrics are depicted in Figures 3 and 4. For both excerpts, the pitch contour is captured quite well by the transcriptions in the bottom staves. The identified problems relate mainly to rhythmic aspects. Mistake A causes a metrical shift, Mistakes B, C, D and G are incorrect rhythms, whereas mistake F is a missing note that causes both rhythmic and melodic distortion. Mistake E relates to a wrong use of accidentals by the transcriber. Two additional insights emerged from our close analysis: First, mistakes D and E recur due to the structure of the melody. Whereas in an assessment by a human expert such repeated mistakes are easily identified, computational metrics would count them as separate instances, resulting in an apparently larger number of mistakes. And, second, whereas we pointed out several divergences between human and computational rating, the basic process of comparing to a reference transcription was common to both. As can be seen from the two examples, AK and RW agree closely in their transcriptions. Hence, shared reference and shared analytic goal seem to be fundamental for their documented agreement. The conclusion concerning computational metrics is that they disregard certain criteria and do not facilitate the statement of an analytic purpose, but are coherent in the basic idea of comparison with a reference transcription.

Place Fig. 3 here: Example transcription of Segment 2 (bottom stave) with a large divergence between rating by experts (AK: 3, RW: 4) and computational metric (9.1). Transcriptions by AK and RW are depicted in the upper two staves. The pickup measure was added to the transcriptions of AK and RW for alignment purposes. Dashed boxes denote mistakes that the experts specified as motivation for their low rating.

Place Fig. 4 here: Example transcription of Segment 6 (bottom stave) with a large divergence between rating by experts (5) and computational metric (9.7). Transcriptions by AK and RW are depicted in the upper two staves. Dashed boxes denote mistakes that the experts specified as motivation for their low rating.

4.4 Problems of AMT

To further explore the types of mistakes in algorithmic transcriptions, six of the transcriptions by algorithms were analysed in the same way as the five human transcriptions in the previous paragraph. The average ratings by the two experts for these six algorithmic transcriptions (avg.: 4.08) were almost identical to the average for the five human transcriptions (4.0), whereas the average computational metric for the algorithmic transcriptions (avg.: 7.81) was lower than for the human transcriptions (avg,: 9.31) . Some interesting contrasts to the type of mistakes observed above emerged. Problems of rhythm/meter (11) did not outnumber problems related to pitch (16). Among the pitch-related problems, wrong pitches played a stronger role, being either octave errors or microtonal inflections that were notated as chromatic alterations. The most prominent difference, however, was additional notes, a phenomenon absent from the analysed human transcriptions. These were either related to the inclusion of what human transcribers would identify as ornamentation into the main melody, or to the spilling over of accompaniment notes into the melody. Hence, we can conclude that algorithms tend to make mistakes that no human transcriber would make. They tend to make both rhythm and pitch contours more complex; better results might be produced by providing tighter constraints regarding rhythm and pitches as an additional learning stage for the algorithms based on a corpus of example transcriptions.

4.5 Agreement between Transcriptions

The analysis in this paper is able to profit from the availability of expert ratings, and obtains additional insight by specific expert assessment of individual transcriptions. In the absence of such expert evaluation and established reference transcriptions, it may be of advantage to estimate the quality of a group of transcriptions automatically. The question is whether we can automatically identify a set of good transcriptions, based on their mutual agreement. As a first step in this direction we investigate how the mutual agreement among the *N* best transcriptions compares with the mutual agreement among the *N* lowest rated. In order to compute the mutual agreement in a group of transcriptions, we employ the combined metric depicted in Table 2 between all transcriptions in a group, and compute the mean of the obtained values. Our comparison demonstrates that the *N* best transcriptions agree mutually more than *N* lowest rated transcriptions (Fig. 5). This implies that high quality transcriptions tend to agree more in their basic pitch, note onset and duration characteristics than low quality transcriptions. The difference is significant over the whole data set, and only for segment 8 can an overlap be observed. This is the segment with the generally lowest quality ratings by the experts, and it was rated as the most difficult to transcribe by the 18 participants in our user study. Stylistically, it is highly idiosyncratic compared to the other segments, characterizing it as a special case among our eight segments. Figure 6 depicts the two transcriptions with the highest average rating for this segment, in which relatively large differences in the interpretation of both pitch and rhythm are apparent.

For Peer Review

Place Fig. 5 here: Mutual agreement among the highest rated (dashed-dotted line boxes) and the lowest rated (solid line boxes). The combination of two metrics found to correlate most with the expert rating was used to mutually compare each group of transcriptions.

Place Fig. 6 here: Transcriptions of Segment 8 that received the highest average ratings by the experts. The tempo of the transcribed segment is about 120 beats per minute.

In a real-world scenario, quality ratings on the level of individual transcriptions will not be available, and shaping a reference committee of a group of transcribers that are assumed to have high expertise may be a viable alternative. We evaluated such an alternative and were able to confirm that the mutual agreement among a group of transcribers is strongly correlated with the average expert rating of that group of transcribers. This implies that the mean mutual agreement among a group of transcribers or transcriptions can be used as an indicator for the choice of reference transcriptions in a corpus. It remains to be explored if the findings in our case study generalize to other musical styles and analytic tasks.

5 Conclusion

By comparing ratings of human experts with computational metrics through corpus and close analysis, we documented differences in how the quality of a transcription is assessed in ethnomusicology and in MIR. We revealed several aspects that the metrics seem to be “missing” in Section 4.3. Computational metrics are only partially correlated with human ratings. Specifically, the highest correlation between metrics and human ratings can be found for metrics that focus on onset times and pitch detection errors, and computational metrics are less sensitive to rhythmic problems compared to experts. An important methodological aspect that is shared is the assessment procedure, which is based on comparison with a reference transcription, which indicates that the MIR procedure is not substantially wrong. A conceptual aspect that is missing is the consideration of the analytic purpose, which importantly guides the shape of the reference transcription for evaluation. In applications where such purpose is not clearly stated - such as the development of generic transcription tools in MIR - we recommend to use more than one reference transcription to cover a range of such purposes. To identify such a group of references in a larger corpus, the mean mutual agreement among a group of transcriptions can be considered as an indicator. However, further research is required to investigate to what extent our findings generalize to other musical styles, which requires the extension of the present work to a larger diversity of musical repertoire and analytical purposes.

The evaluation in the present paper agrees with the finding in Holzapfel and Benetos (2019) in that AMT does not achieve transcriptions of sufficient accuracy for the present corpus. The quality of algorithmic transcriptions is still fairly low (see Section 4.1) and algorithms make mistakes that no human transcriber would make (see Section 4.4). Edited transcriptions that use an automatic transcription as a starting point have a tendency to be biased, although automatic transcriptions can assist less experienced transcribers. The documented lack of accuracy of AMT seems to be one reason why ethnomusicologists have not made much use of automatic transcription into staff notation. By contrast, some ethnomusicologists do find automatic transcription into other forms of notation (*e.g.* melographic) accurate enough for their purposes, even when they require manual correction. What may be further reasons why so few ethnomusicologists use AMT to derive transcriptions? One aim of the melograph is to reveal things that we might not perceive just by listening, whereas automatic staff notation rather aims to match what a competent human transcriber would hear. Hence, in contrast to the melograph, automatic staff notation rather aims at suppressing things that cannot be heard from the algorithmic output. It may therefore be a valuable effort to consider how such potential scepticism by ethnomusicologists towards such suppression can be addressed.

To this end, suggestions for future directions of research in MIR relate to, first, the evaluation of transcriptions using evaluation metrics that consider several levels. On the first and lowest level (signal level), metrics such as the ones used in this paper are applied that consider onset times and durations. Beyond a local approach of estimating onsets and duration, contextual information needs to be included that accounts for the overall structure and repetition when assessing a transcription. The second level takes the analytic purpose into account. This may be facilitated by choosing among a group of reference transcriptions, for instance based on the level of detail they provide. One important conclusion of our paper is therefore the rejection of the idea of a single "ground truth" transcription. A third level should take performance aspects into account, such as the use of expressive signs in a transcription. Finally, on the fourth level, notational style and conventions are considered by evaluating how well a transcription adheres to conventions of the notation system. The resulting metric would thus combine three dimensions, expanding on the metric by McLeod and Steedman (2018) who proposed a multidimensional approach previously: First, a dimension of note detection accuracy that considers context and analytic purpose, second, a dimension that rates use of expressive signs, and finally, a dimension that evaluates the adherence to notational conventions.

Further steps in MIR research include the development of AMT approaches that are able to learn conventions concerning the transcription of a specific style from a compact and representative collection of example transcriptions. Alternatively, in the context of algorithmic composition it has been attempted to have an algorithm produce a larger set of compositions and then have a human choose from these (Sturm and Ben-Tal, 2017), and such a method of selection by a human would be equally applicable for the problem of AMT. In turn, these choices can be used to make the system learn further, *i.e.* to constrain it. The imposing of particular formal, high-level, conceptual rules on AI models is an ongoing topic of research in many machine learning domains (Hu *et al.*, 2016; Marra *et al.*, 2019). In the context of AMT, such rules could comprise the possible note durations or interval sizes to be used, which should be parameters accessible to the user of an AMT software. Alternatively, the integration of a music language model (Ycart *et al.*, 2019) combined with an acoustic model can also constrain the resulting transcriptions to follow a specific music style or specific transcription conventions.

For Peer Review

The large diversity of possible transcriptions of one piece, caused by the large diversity of possible analytic purposes, could be catered for by considering a target notation system that is mid-way between melographic and staff notation, such as the Global Notation System[iv]. This represents pitch in a continuous scale over time, but also maps it onto specific pitch and duration categories that can be read in terms of a background music system - whichever system is in operation in the music in question. With an increasing number of MIR methods focussing exclusively on audio-to-notation transcription (Carvalho and Smaragdis, 2017; Nakamura *et al.*, 2018; Roman *et al.*, 2019), it seems timely to consider the use of such methods for research in ethnomusicology by rethinking the targeted notation system, the user interactions when training an algorithm, and the evaluation process. MIR approaches that consider these three aspects are then likely to be of higher value for research in ethnomusicology, by providing flexible means of visualization and analysis of larger corpora with AMT as a starting point.

6 Funding

This work was supported by the Swedish Research Council [2019-03694 to A.H.]; the Marianne and Marcus Wallenberg Foundation [MMW 2020.0102 to AH]; EB is supported by a Turing Fellowship under the EPSRC grant EP/N510129/1.

For Peer Review

7 References

Abdallah, S., Benetos, E., Gold, N., Hargreaves, S., Weyde, T. and Wolff, D. (2017). The digital music lab: A big data infrastructure for digital musicology. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, **10**(1): 1–21.

Allgayer-Kaufmann, R. (2005). From the innocent to the exploring eye: Transcription on the defensive. *The World of Music*, **47**(2): 71–86.

Andreoulakis, I. and Petrakis, S. (2013). *Σκοποί και μαντινάδες της Κρήτης*. Athens: Filipos Nakas.

Bay, M., Ehmann, A. F., and Downie, J. S. (2009). Evaluation of multiple-f0 estimation and tracking systems. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 315–320.

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

Benetos, E. and Holzapfel, A. (2015). Automatic transcription of Turkish microtonal music. *The Journal of the Acoustical Society of America*, **138**(4): 2118–2130.

Cannam, C., Landone, C., Sandler, M.B. and Bello, J.P. (2006). The Sonic Visualiser: A visualisation platform for semantic descriptors from musical signals. In *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR),* pp. 324–327.

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

Cogliati, A. and Duan, Z. (2017). A metric for music notation transcription accuracy. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 407–413.

Cooper, D. and Sapiro, I. (2006). Ethnomusicology in the laboratory: From the tonometer to the digital melograph. *Ethnomusicology Forum*, **15**(2): 301–313.

Daniel, A., Emiya, V. and David, B. (2008). Perceptually-based evaluation of the errors usually made when automatically transcribing music. In *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 550–556.

Ellingson, T. (1992). Transcription. In Myers, H. (ed), *Ethnomusicology: An Introduction*. London: W. W. Norton & Company, pp. 110-152.

Emiya, V., Badeau, R. and David, B. (2010). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio,*

For Peer Review

Speech, and Language Processing, **18**(6): 1643–1654.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.Z.A., Dieleman, S., Elsen, E., Engel, J. and Eck, D. (2019). Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proceedings of the International Conference on Learning Representations (ICLR)*, arXiv:1810.12247.

Holzapfel, A. and Benetos, E. (2016). The Sousta corpus : Beat-informed automatic transcription of traditional dance tunes. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 531–537.

Holzapfel, A. and Benetos, E. (2019). Automatic music transcription and ethnomusicology : A user study. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 678–684.

Hood, M. (1982 [1971]). *The ethnomusicologist*. Kent, OH: Kent State University Press.

Hood, M. (1993). The untalkables of music. *Annuario degli Archivi di Etnomusicologia dell'Accademia Nazionale de Santa Cecilia*, **1**: 137–142.

Höger, F., Frieler, K. and Pfleiderer, M. (2019). Digging into pattern usage within Jazz improvisation (pattern history explorer, pattern search and similarity search). In *Digital Humanities Conference (DH2019)*. Available from <https://dev.clariah.nl/files/dh2019/boa/0723.html> (accessed 24 February 2021).

Hu, Z., Ma, X., Liu, Z., Hovy, E. and Xing, E. (2016). Harnessing deep neural networks with logic rules. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.2410–2420.

Institute of Mediterranean Studies (2005). *Website of the Crinnos project* [online]. Available from <http://crinnos.ims.forth.gr> (accessed 26 March 2019).

Jairazbhoy, N. (1977). The 'objective' and subjective view in music transcription. *Ethnomusicology*, **21**(2): 263–73.

Killick, A. (2020). Global notation as a tool for cross-cultural and comparative music analysis. *Analytical Approaches to World Music*, **8**(2): 235–279.

Klapuri, A. and Davy, M. (eds) (2006). *Signal Processing Methods for Music Transcription*. New York: Springer.

Marian-Bălaşa, M. (2005). Who actually needs transcription? Notes on the modern rise of a method and the postmodern fall of an ideology. *The World of Music*, **47**(2): 5–29.

Marra G., Giannini F., Diligenti M. and Gori M. (2020). Integrating learning and reasoning with deep logic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 517–532.

For Peer Review

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J. and Dixon, S. (2015). Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation (TENOR)*, pp. 23–31.

McLeod, A., Steedman, M. (2018). Evaluating automatic polyphonic music transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 42–49.

Mesaros, A. and Virtanen, T. (2010). Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, article number 546047.

Metfessel, M. (1928). *Phonophotography in Folk Music: American Negro Songs in New Notation*. Chapel Hill: University of North Carolina Press.

Moorer, J.A. (1975). On the transcription of musical sound by digital computer. In *Second USA-JAPAN Computer Conference*, pp. 32-38. Reprinted in the Computer Music Journal (1977), **1**(4): 32–38.

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

Nettl, B. (2015). *The Study of Ethnomusicology: Thirty-Three Discussions*. Third edition. Urbana: University of Illinois Press.

Panteli, M., Benetos, E. and Dixon, S. (2017). A computational study on outliers in world music. *Plos one*, **12**(12), p.e0189399.

Suvarnalata, R. and van der Meer, W. (n.d.). *Music in Motion: the automated transcription for Indian music (AUTRIM) project by NCPA and UvA* [online], Available from <https://autrimncpa.wordpress.com/about/> (accessed 26 August 2020).

Román, M.A., Pertusa, A. and Calvo-Zaragoza, J. (2019). A holistic approach to polyphonic music transcription with neural networks. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 731–737.

Ryynänen, M.P. and Klapuri, A.P. (2008). Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, **32**(3): 72–86.

Salamon, J., Gómez, E., Ellis, D.P. and Richard, G. (2014). Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, **31**(2): 118–134.

Sanyal, R. and Widdess, R. (2004). *Dhrupad: tradition and performance in Indian vocal music*. Aldershot, UK: Ashgate (SOAS Musicology Series).

For Peer Review

Seeger, C. (1958). Prescriptive and descriptive music-writing. *Musical Quarterly*, **44**: 184–95.

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez Gutiérrez, E., Gouyon, F., Boyer, H., Jordà Puig, S. and Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H. and Widmer, G. (2013). *Roadmap for Music Information ReSearch*. MIReS Consortium, available from <http://www.mires.cc> (accessed 24 February 2021).

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

Sturm, B. L. and Ben-Tal, O. (2017). Taking the models back to music practice: Evaluating generative transcription models built using deep learning. *Journal of Creative Music Systems*, **2**(1). doi: https://doi.org/10.5920/JCMS.2017.09.

Su, L. and Yang, Y.H. (2015). Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *International Symposium on Computer Music Multidisciplinary Research*, pp. 309–321.

Tallotte, W. (2017). Improvisation, creativity, and agency in South Indian temple rāga performance. *Asian Music*, **48**(2): 24–61.

Tenzer, M. (2006). *Analytical studies in world music*. New York: Oxford University Press.

Tenzer, M. and Roeder, J. (2011). *Analytical and Cross-Cultural Studies in World Music*. New York: Oxford University Press.

Tidhar, D., Dixon, S., Benetos, E. and Weyde, T. (2014). The temperament police. *Early Music*, ***42***(4): 579–590.

Wu, C.W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Müller, M. and Lerch, A. (2018). A review of automatic drum transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(9): 1457–1483.

Ycart, A., McLeod, A., Benetos, E. and Yoshii, K. (2019). Blending acoustic and language model predictions for automatic music transcription. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 454–461.

Ycart, A., Liu, L., Benetos, E. and Pearce, M. (2020). Investigating the perceptual validity of evaluation metrics for automatic piano music transcription. *Transactions of the International Society for Music Information Retrieval*, **3**(1): 68–81.

Notes

For Peer Review

i With the term “polyphonic” we refer to the presence of multiple concurrent pitches, as opposed to “monophonic” which assumes the presence of a single pitch for a given time segment.

ii All material used for the study (along with all obtained transcriptions) is accessible at https://kth.box.com/v/DSH2021Transcription

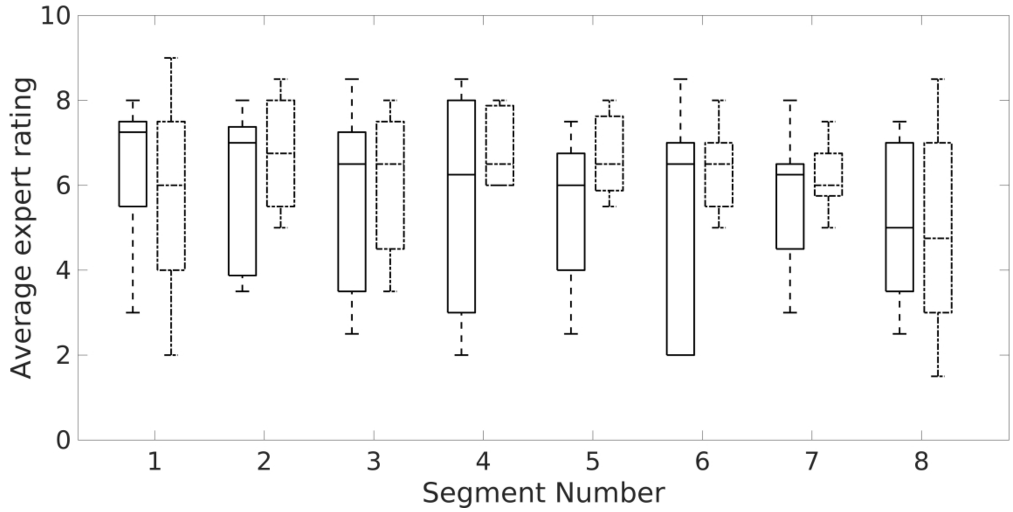
iii https://scorecloud.com/. This software is characterized by competitive performance in monophonic transcription of violin recordings, an instrument with timbre characteristics similar to the Cretan lyra.

iv The Global Notation System, and its ongoing development and extensions, are documented at http://globalnotation.org.uk

For Peer Review

Distribution of the ratings between the two experts. Correlation coefficient is 0.754 (significant, p<10e-26).

874x746mm (72 x 72 DPI)



Mean expert ratings for the eight segments, separated into two groups for complete manual transcription (solid-line box plots) and editing AMT (dashed-dotted-line box plots).

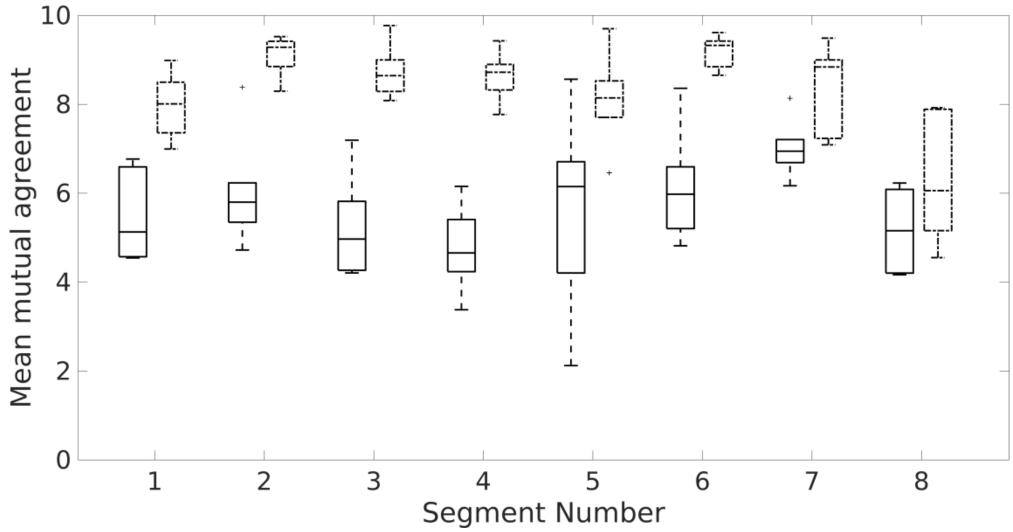
1761x886mm (72 x 72 DPI)

Example transcription of Segment 2 (bottom stave) with a large divergence between rating by experts (AK: 3, RW: 4) and computational metric (9.1). Transcriptions by AK and RW are depicted in the upper two staves. The pickup measure was added to the transcriptions of AK and RW for alignment purposes. Dashed boxes denote mistakes that the experts specified as motivation for their low rating.

155x49mm (299 x 299 DPI)

Example transcription of Segment 6 (bottom stave) with a large divergence between rating by experts (5) and computational metric (9.7). Transcriptions by AK and RW are depicted in the upper two staves. Dashed boxes denote mistakes that the experts specified as motivation for their low rating.

154x51mm (300 x 300 DPI)



Mutual agreement among the highest rated (dashed-dotted line boxes) and the lowest rated (solid line boxes). The combination of two metrics found to correlate most with the expert rating was used to mutually compare each group of transcriptions.

1761x920mm (72 x 72 DPI)

Transcriptions of Segment 8 that received the highest average ratings by the experts. The tempo of the transcribed segment is about 120 beats per minute.

151x30mm (300 x 300 DPI)