Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-12476-0

A perceptual measure for evaluating the resynthesis of automatic music transcriptions

Federico Simonetta[1] · **Federico Avanzini**[1] · **Stavros Ntalampiras**[1]

Received: 4 May 2021 / Revised: 26 December 2021 / Accepted: 25 January 2022

Abstract
This study focuses on the perception of music performances when contextual factors, such as room acoustics and instrument, change. We propose to distinguish the concept of "performance" from the one of "interpretation", which expresses the "artistic intention". Towards assessing this distinction, we carried out an experimental evaluation where 91 subjects were invited to listen to various audio recordings created by resynthesizing MIDI data obtained through Automatic Music Transcription (AMT) systems and a sensorized acoustic piano. During the resynthesis, we simulated different contexts and asked listeners to evaluate how much the interpretation changes when the context changes. Results show that: (1) MIDI format alone is not able to completely grasp the artistic intention of a music performance; (2) usual objective evaluation measures based on MIDI data present low correlations with the average subjective evaluation. To bridge this gap, we propose a novel measure which is meaningfully correlated with the outcome of the tests. In addition, we investigate multimodal machine learning by providing a new score-informed AMT method and propose an approximation algorithm for the p-dispersion problem.

Keywords Automatic music transcription · Audio resynthesis · Music perception · Music information retrieval

1 Introduction

Automatic Music Transcription (AMT) can be broadly defined as the process elaborating on digital audio recordings in order to infer a specific set of relevant musical parameters of

✉ Federico Simonetta
Federico.Simonetta@unimi.it

Federico Avanzini
Federico.Avanzini@unimi.it

Stavros Ntalampiras
Stavros.Ntalampiras@unimi.it

[1] LIM – Music Informatics Laboratory, Department of Computer Science, University of Milano, Milano, Italy

the sounds, and to convert them in some form of notation [33]. Nowadays, AMT is a broad signal processing field encompassing a wide gamut of tasks and approaches. As an example, the output of an AMT system can be a traditional score, a Standard MIDI File (SMF), or a set of ad-hoc features [3]. A traditional score is a sequence of symbols that describes music according to the western notation and focuses on expressing music in a human-readable way so that it can be easily reproduced. SMFs instead describe the performance itself, possibly sacrificing precision at the semantic level (which is useful to the musician for performing the piece) while gaining precision in the description of the physical events that happened during the execution – i.e. velocity and duration with which the keyboard keys were pressed, the pedaling timing, etc.

SMFs originate from the description of keyboard music and it hardly adapts to other instruments; moreover, the constantly increasing importance of Music Information Retrieval (MIR) created the need for a different symbolic representation of music sounds: hence, several AMT systems extract MIR features – e.g. $f0$ estimation, intensity levels, and timbral descriptors [17, 48]. The input to AMT systems is a variable itself: most authors focus on mono-modal methods which take as input only the audio recordings, while other methods tackle the problem with multimodal approaches [51] such as audio and scores (score-informed) [1, 12, 27, 61].

Regarding the resynthesis of MIDI transcribed recordings, many studies have shown that performers change their way of playing according to contextual conditions, such as physical properties of the instrument, reverberation and room acoustics, often even unconsciously [30]. Recently, the authors of [66] proposed a method to automatically transfer a piano performance across different contexts (instruments and environments) in order to make the reproduced sound as similar as possible to the original one, by adapting MIDI velocities and duration to the new context. However, they assume knowledge of the original piano parameters to carry out the adaptation; moreover, they use the same microphones and post-processing pipelines in every different context.

There are very few attempts addressing the subjective evaluation of AMT systems, with the notable exception of [67]. The authors prepared more than 150 questions asking subjects to chose the best transcription of a reference audio clip lasting 5-10 seconds and managed to collect 4 answers per question. We use the results of the above work as a main reference for the present study.

We propose a methodology for evaluating the resynthesis of MIDI recordings extracted through AMT systems, taking into account contextual conditions in the resynthesis. Specifically, we wish to adapt the performance to the new resynthesized context while having knowledge only of the target context and of the original recording. In doing so, we propose a conceptual framework which distinguishes between the actual performance and the underlying artistic intention (i.e., the interpretation), and we design a methodology assessing to which extent such interpretation is perceivable in the resynthesized recording. Possible applications that may be impacted by the proposed study are manifolds. The long-term objective of this study is the resynthesis of music for production and restoration purposes – see Section 2. Other possible use-case scenarios include musicological studies and music teaching applications, such as the analysis and comparison of the interpretation, which in turns can pave completely new paths for these research fields; moreover, the ability of transcribing both performance and interpretation would also allow the comparison of the manifold ways in which different performers adapt their interpretation. Not least, architectural studies could be impacted from robust context-aware AMT models.

The contributions of our work are 1. an indication of MIDI format's inability to completely capture the artistic intention of a music performance and 2. a perceptually-motivated

measure for the evaluation of AMTs. In addition, we investigate multimodal machine learning technologies applied to AMT by providing a new score-informed method and propose an approximation algorithm for the p-dispersion problem to optimally-chose the excerpts for the test.

For the purpose of comparability and reproducibility of the results, the code is made available online[1] and the full set of the computed statistics are available in the Supplementary Materials.

2 Restoration, performance and interpretation

One of the long-term motivations behind the present work is the automatic restoration of old and contemporary music recordings by reproducing the performances as accurately as possible. If audio restoration processes [21] usually target old and deteriorated operas, restoration of contemporary art is a relevant objective as well. In particular, the World Wide Web offers the opportunity of accessing large resources of low quality videos, images, and audios. Moreover, modern mobile devices allow people to record music and videos with inexpensive transducers that produce low quality data in respect to the expensive professional technology. Our restoration intent is therefore directly connected with the democratization of music production technologies and with the fruition of old audio recordings.

The audio restoration literature is dominated by two general approaches: the first aims at reconstructing the sound as it was originally "reproduced and heard by the people of the era", while the second and most ambitious one aims at reconstructing "the original sound source exactly as received by the transducing equipment (microphone, acoustic horn, etc.)" [21, 54]. However, an exact restoration is impossible in both cases. Particularly regarding the second approach, aiming at recovering the so-called "true sound of the artist" [42] exposes the restoration to subjective interpretations regarding the performer's artistic intention. Indeed, the artist's original intention is never completely captured by the recording because of the recording equipment limitations, such as microphones compression, noises, and degradation [66]. To get over this issue, several studies tried to exploit the timbral features of the audio recording to compute original sound characteristics such as note intensities [28, 37, 63], but this is a particularly challenging problem hindered by the variability of MIDI velocity mappings [10].

Thus, we propose not to recover the original artistic intention but the intention survived until today and perceivable by the listener, as this is the best case scenario which is not influenced by subjective factors. The proposed idea consists in a) analyzing a recording via an AMT system so as to estimate the parameters of the performance and b) resynthesizing it using modern technologies. In other words, we wish to retain the effort of the performer in a resynthesized version of the automatically-extracted music transcription.

Towards defining the specific problem, two concepts need to be distinguished, i.e. *interpretation* and *performance*. The *performance* is the set of physical events that result in the activity of playing a music piece. It is bijectively associated with a certain time, place and performer, so that it is a unique unrepeatable act. *Interpretation*, instead, refers to the *ideal* performance that the performer has in mind and tries to realize. Thus, an interpretation could be repeated in different performances and differs from the performance because it lacks the adaptation to the context. It comprises the ultimate goal of the performer and thus, we

[1] https://github.com/LIMUNIMI/PerceptualEvaluation.

identify it with the performer's artistic intention. During the restoration process, we seek to generate a new performance based on the interpretation extracted from the audio recording.

This operative definition is completely unrelated to the musicological debate about what an interpretation is – e.g. when the notion "interpretation" was introduced with reference to music [13] or whether the interpretation is the "performer's idea of the music" [11]. We rather focus on tracing the difference between desired and realized performance, which differ due to external causes.

Such distinction is in line with the state-of-art research in the field of Music Performance Analysis that focuses on the acoustics of concert stages and rooms. The interest in the influence of the room acoustics on the performance dates back to 1968 [60], but it has not received significant attention by the research community in the successive decades with the exception of a limited amount of works [6, 41, 56]. All these studies showed that musicians (orchestra, choir, piano, and percussion players) adapt their music performance to the acoustic environment in which they perform. Several music psychologists hypothesized the existence of an interior representation of the sound that the musician wants to convey [18]. Such a perspective was further elaborated by various authors in an attempt to understand how musicians adapt their performance to various acoustic environments. First, subjective tests on musicians playing in different virtualized acoustic settings were explored and a circular feedback model between the performer and acoustical environment was proposed [57, 59]. Then, in the last decade, a few studies attempted to tackle the problem with objective evaluations. In 2010 and 2015, the same authors proposed two new studies in which physical features extracted from audio recordings were compared with the subjective self-evaluation of musicians and of listeners [32, 58]. From the comparison of objective and subjective evaluations, they argued that the feedback process was conscious. Other researchers tried to understand which factors of the room acoustics influence the performance and how [30], arguing that the way in which musicians change their execution is performer-specific. In recent years, the research in the room influence on the performance has continued with the analysis of singers [36, 43] and trumpet players [19]. The overall contribution of the previous studies is that the adaptations applied by musicians influence the timbre, the amplitude dynamics, and the timing. An overview of existing works and methodologies has been recently published [34]. However, all the existing studies, are directed towards the understanding of the factors characterizing room acoustics. At the same time, they rarely consider the listener perception and never take into account indirect factors that can effectively change the acoustics of the instrument, such as the temperature and the humidity.

Two theoretical concepts have been developed in the related literature, as outlined above. First, previous studies suggest the existence of a *circular feedback* between the performer and the surrounding environment [58, Fig. 1]. Second, the outlined literature is coherent with the existence of an *interior representation* of the music performance that has to be realized; this idea was proposed by psychological studies [18], and developed in the above-mentioned literature [31]. The definition of "interpretation" proposed in this work is undoubtedly similar to the concept of such interior representation. In Fig. 1, the difference between "interpretation" and "performance" is clearly outlined.

3 Designing the Test

This section analyzes the conducted experiment from the technical point of view as well as the reasoning behind the presented choices.

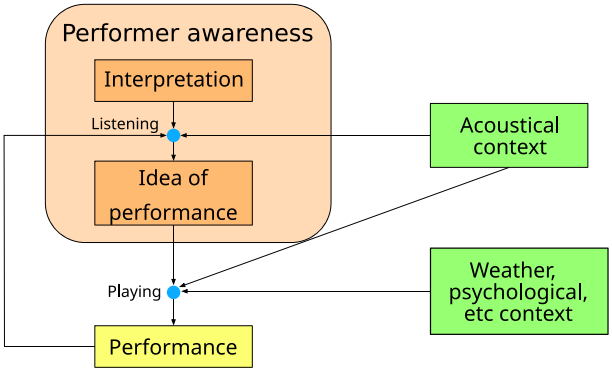

Fig. 1 Diagram showing the circular feedback involving the interpretation, the acoustical context, and the performance. The interpretation is influenced by the acoustical context based on the feedback coming from listening to the performance itself. The "idea of performance" is an intermediate representation that the performer consciously creates based on such feedback

3.1 Research questions

Given the definitions of Section 2, we assume the following:

1. MIDI – and consequently SMF – is able to record every aspect characterizing a piano performance;
2. using the same interpretation, a musician is going to create different performances given that the context (instrument, room, audience) changes;
3. during the audio recording process, there is information loss up to a certain extent, while a different type of information, related to the context (including microphones), may be introduced. Contextual alterations render practically impossible to extract the exact MIDI performance from the audio.

The first research question that we seek to answer with the listening test is to which extent the interpretation is still identifiable when changing the context and retaining the performance, i.e. the MIDI recording. According to our second assumption and to psychological studies described in Section 2, if the context changes and the performance stays constant, we assume that the two performances are generated from two different interpretations. Consequently, while retaining the same interpretation in two different contexts, we expect that the listener will perceive two different interpretations. With this research question, we aim at assessing whether MIDI is effective in representing aspects related to interpretation.

The second research question is whether state-of-art AMT systems, typically trained to extract performance parameters, are effective in the extraction of the interpretation as well. Since it is practically impossible to recreate the exact performance, it is interesting to investigate whether a slightly modified MIDI coming from AMT systems is able to encompass interpretation aspects. In case this is true, we could resynthesize a given interpretation with a different context and obtain a different performance of the same interpretation.

Finally, following the line of thought presented in [67], we want to provide a perceptual evaluation of AMT systems.

3.2 Tasks

Three different tasks were designed addressing the above-mentioned questions. They consist in assessing the similarity in the interpretation between a reference audio excerpt and several candidates. The tasks differed in the way the audio clips were generated, as follows:

1. in the first task, named "transcription", all audio excerpts including the reference were synthesized from MIDI using the same context (i.e., same virtual instrument and reverberation);
2. in the second task, named "resynthesis", all audio excerpts were still synthesized from MIDI but we used two different contexts for the reference and the candidates;
3. in the third task, named "restoration", the reference was a real-world recording, while the candidates were synthesized from MIDI with a virtual instrument. Since the original recording contains substantially more noise than the synthesized candidates, this specific task is representative of a restoration process.

All tasks used the same 5 excerpts extracted following the process explained in Section 4, where we also describe the reasoning behind the choice of the virtual instruments.

3.3 Protocol and interface

Due to COVID-19 restrictions, we designed an online test using the "Web Audio Evaluation Tool" (WAET) [29].

First, subjects were prompted with some introductory slides explaining the difference between interpretation and performance. Specifically, after the formal definitions, they were suggested to adopt the following way of thinking: the interpretation is associated with the pianist, while the performance is related to a particular concert. Then, they listened to the first 30 seconds of two different performances by Maurizio Pollini of the Sonata No.30 op.109 by L. van Beethoven, and to a performance of the same piece by Emil Gilels; they were told that the first two were the same interpretation but different performances while the latter was both a different performance and interpretation. Finally, they were asked to a) use headphones or headset, b) stay in a quiet place, and c) consent to the use of their answers in an anonymous form.

Next, subjects were asked preliminary questions, namely:

1. what level of expertise they have with music (options: none/hobbyist vs. undergraduate/graduate/professional)
2. how often they listen to classical music (options: less vs. more than 1 hour per week)
3. how often they listen to music other than classical (options: less vs. more than 1 hour per week)
4. what is the cost of the headphones they were using (options: less vs. more than 20 euros)

At this point, subjects were exposed to the experimental interface through an example question, with guitar instead of piano recordings. Similarly to MUSHRA test [16], subjects could play back a reference audio file and four additional candidates containing the same musical excerpt resynthesized with various contexts and performances; they were asked to rate each candidate with a horizontal slider on a continuous variable evaluating the extent to which the candidate clip contained the same interpretation as the reference. Three labels

were put along the slider: “different interpretations”, “don’t know”, and “same interpretation” – see Fig. 2 for a screenshot. Users were instructed to experiment with the example question until s/he felt comfortable. As shown in Fig. 2, users could adjust the audio level at any time, even during the example question. Audio clips were normalized to -23 LUFS level in both the example and the actual test. However, authors are not aware of studies about loudness influence in Music Performance Assessment studies.

Finally, subjects started the actual test, with piano recordings. The order of questions and candidates were randomized, as well as the initial positions of the sliders to prevent biases. Since the test lasted approximately 30 minutes on average, if a subject decided to leave the session her/his answers were recorded. For this, we used a feature provided by WAET to first prompt new subjects with questions for which fewer answers were collected, so that the number of answers per question was uniform.

3.4 Number vs. duration of excerpts

In any listening test which deals with the artistic expression of the performer, an issue arises concerning the length of the excerpts. In general, it can be expected that longer duration will lead to more accurate subjective judgments. However, one second competing factor is the total duration of the test: the longer the test is, the more difficult is to find volunteers willing to take the full test and be able to keep their concentration for the entire duration [50].

One study observed that the duration of excerpts did not influence the emotional response of the listeners [4]. On the other hand, it was shown that in the context of piano performance assessment, graduate music students and faculty professors rated 60 s excerpts higher than 20 s excerpts, while no significant difference was observed for non-graduate students [62]. Another study conducted on wind bands took into account the level of the performers and showed that music majors rated 25 s and 50 s excerpts higher than 12 s excerpts in the case of university or professional level performances, while the opposite

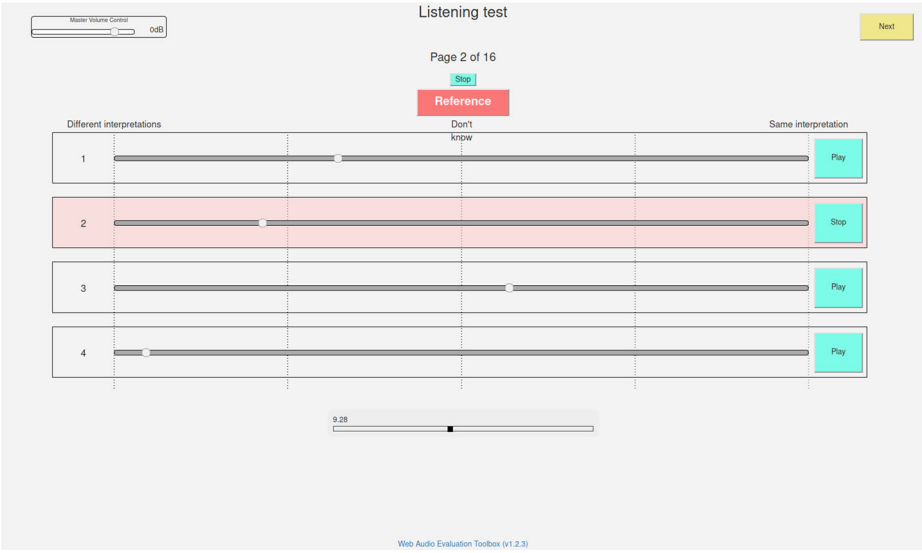

Fig. 2 Screenshot of the interface created using the “Web Audio Evaluation Tool” [29]

happened for performances at the high-school level [20]. A study on children chorale music revealed that 60 s excerpts were rated higher then 20 s excerpts by music majors [40]. Finally, another study on band performances was conducted with excerpts 12 s, 15 s, and 30 s long, and found that ratings from music majors were higher for long excerpts of bad performances [64].

It comes from the cited literature, that a minimum sufficient duration to observe the difference between ratings given by expert and non-expert listeners is in the range $15-25$ s. Thus, we used excerpts lasting 20 s. Since each subject performed all of the three tasks – see Section 3.2 – and we aimed at keeping the test less than 30 minutes long, we opted for 5 excerpts, resulting in 15 questions (5 questions per task) each with 5 audio clips (1 reference and 4 candidates) lasting 20 s, for a total minimum duration of 25 minutes.

The studies discussed above also suggest that expert listeners tend to base their judgements on longer time features with respect to non expert raters. We expect a similar behavior for the task of comparing two interpretations. As a consequence, if no significant difference is found between expert and non-expert ratings, a difference may still be observed when using excerpts longer than 20 s. The only way to rule out such hypothesis would be by using full song excerpts.

4 Generating excerpts and contexts

Since we used a limited number of excerpts, we wanted to chose them in a way that minimizes any type of subjective bias. The same also applies for the choice of the contexts. In Fig. 3, we show the overall workflow used for solve this problem.

4.1 The *p*-dispersion problem and uniform selection

For choosing the excerpts, we built a dataset where each sample was represented by features extracted from audio clips and MIDI symbols. Since we expected that the perception of music changes as these features vary, we aim at maximizing the distance between feature vectors of the chosen excerpts so that perceptual variations are revealed.

This means that we are looking for the p samples in the feature space which are distributed uniformly, while the distance between them is maximal. This problem can be seen as a variation of the *p-dispersion problem* [14] or *max-min facility dispersion problem* [45]. However, the p-dispersion problem does not impose any restriction with respect to data

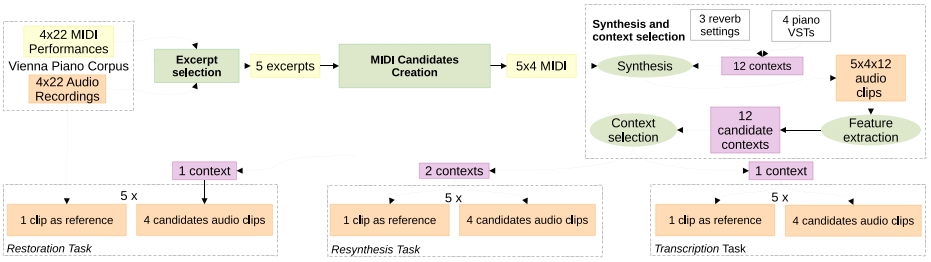

Fig. 3 The workflow used for creating the restoration, resynthesis, and transcription tasks. Legend: a) Yellow: MIDI data; b) Orange: audio data; c) Purple: contexts; d) Green: Operations

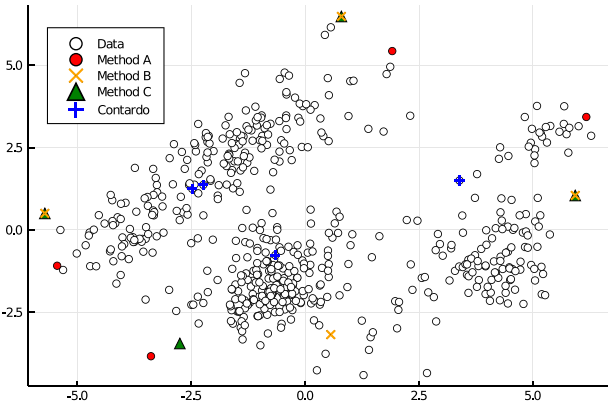

Fig. 4 Comparison of methods for the p-dispersion problem with $p = 4$. Data are the windows extracted from the *Vienna 4x22 Piano Corpus*. PCA was used in this plot to reduce the dimensionality from 15 down to 2 for demonstration purposes. For the listening test, we used *Method A*. *Contardo* is the state-of-art mathematically-proven method for the p-dispersion problem [9]. For the comparison, we used the original Julia code provided by the author

distribution, thus we derived four *ad hoc* algorithms solving the present problem based on Ward-linkage clustering [15]. Figure 4 compares them on our excerpt dataset.

4.2 *p*-dispersion problem

Here we briefly describe the algorithm and compare it with a state-of-art method for solving the p-dispersion problem but we leave to future works the mathematical study of the method[2].

Our approach consists in finding p subsets using hierarchical clustering. We used Ward method because it tries to minimize the variance inside each cluster and, consequently, each cluster is well represented by its centroid. In contrast to k-means clustering, instead, it is well suitable even for little sized datasets and does not depends on initialization heuristic. After having partitioned the data in p clusters with the Ward method, we chose one point per cluster as follows. We chose the point in each cluster which maximized the distance from the centroid of all the other points in the dataset (*Method A*). Successively, we have also considered other strategies to chose the point in each cluster, namely: the point which maximize the minimum distance from the centroids of the other clusters (*Method B*), the point which maximize the minimum distance from the points in the other clusters (*Method C*), the point which maximize minimum distance from all the other points (*Method D*). In Table 1 we compare these methods to the state-of-art method for the p-dispersion problem by Contardo, for $p = [4, 5, 10]$, using the code provided us by the author; we used the datasets of the Contardo's work containing less than 10.000 samples with the addition of our dataset created with PCA output of 2 features per sample [9]. For method [9], we used the Julia implementation provided us by the author. We have always used the euclidean distance – or sometime the sum of the absolute differences for improving the computation time.

[2]Code available at https://framagit.org/sapo/selection_test.

Table 1 Comparison of methods for solving the p-dispersion problem

	Min Dist	Time (s)	% wins vs all	% wins vs [9]
Method A	1.58E+04	2.00	9.52%	60.32%
Method B	1.33E+04	**1.66**	25.40%	61.90%
Method C	**1.90E+04**	9.34	**41.27%**	**63.49%**
Method D	8.43E+03	10.44	1.59%	50.79%
Contardo	9.21E+03	153.30	34.92%	100.00%

Columns are: average minimum distance in the output set, average time in seconds needed, percentage of instances in which each method had *Min Dist* grater or equal than any other (all) or than the Contardo's method [9]

4.3 Excerpt selection

The selection of the audio excerpts started from the *Vienna 4x22 Piano Corpus* [22], which consists of 88 audio and corresponding MIDIs recordings of 4 famous pieces highly representative of the classical-romantic music period, played by 22 professional and advanced student pianists. This corpus was useful in order to have a negative reference (NR) available for any chosen excerpt – i.e. a different interpretation. We used the *ASMD* framework [52] to handle the loading of files and dispatching parallel processing routines. Every audio clip was converted to monophonic, downsampled to 22050 Hz and normalized using ReplayGain [46]. MIDI files were loaded using pianorolls where each pixel contained the velocity value of the ongoing note and each column had a resolution of 5 ms. In order to compensate for temporal misalignments, for each pair (*audio, MIDI*), we identified the audio frame of first onset and last offset using an energy threshold of -60 *dB* under which the sound was identified as silence, and trimmed the files accordingly.

We split the audio recordings in windows and considered each window as a possible excerpt for the listening test. Each window lasted exactly 20 s with a hop-size of 10 s, resulting in 564 total windows. For each window we extracted a set of audio and symbolic features. The first were extracted from the audio recordings and consisted of high-level features among the most used in the MIR field, extracted using the state-of-the-art library Essentia [5].To take into account the timbral characteristics, we extracted 13 MFCCs [2]; we used a state-of-the-art onset detection method to extract 7 rhythmic descriptors [68]. Furthermore, we used the Essentia library to estimate BPM, along with the first and second peak values, spreads, and weights of the corresponding histogram; such features are related to the timing characteristics of the performance as rendered in the audio. Regarding the symbolic features, we used the non-zero pixels in the window pianorolls to extract information about the performance as recorded by the sensorized piano. Specifically, we extracted the average and standard deviation of pitches, velocities, duration, number of contemporaneous notes in each column and pitch difference in each column relatively to the lowest pitch in the column – which relates to the type of harmony. The resulting features were concatenated in an array of 30 features. Then, the 564 windows were standardized (mean removal and variance scaling) and passed through PCA to obtain 15 relevant features explaining 92% of variance.

We applied the methods described earlier to look for 4 dispersed windows. To ensure that the 4 selected points well represented the whole dataset, we chose the only method that managed to select one window for each of the 4 pieces in the *Vienna* corpus (*Method A*). Then, we added one excerpt computed as the medoid of the dataset, obtaining in total 5 excerpts.

It should be noted all the excerpts last exactly 20 s; this may be criticized as it implies that NR and realignments could lead to slightly different contents, since a given excerpt could be played and transcribed in a different time lapse. On the other hand, having the NR lasting a different amount of time would have produced a potential bias at the listener side, who would have been able to identify the NR based on duration rather than audio content. The variability of the features of the chosen excerpts are shown in Fig. 5.

4.4 MIDI candidates creation

For each question in the test, we used four candidates in addition to the reference excerpt: a negative reference (NR) containing a performance by another pianist of the same excerpt, a hidden reference (HR) containing the same performance as the reference, and two performances extracted using two different AMT systems described next. For the NR, we realigned its MIDI recording to the chosen reference MIDI using FastDTW [49], before trimming. For the *restoration* task, where the reference was a real-world recording, the HR was not the same audio recording (which would be immediately recognizable from the remaining candidates), but rather the associated MIDI available in the Vienna corpus.

Various AMT models were published in recent years; unfortunately, only the ones trained for solo piano music achieve satisfactory performances. In particular one model, here called *onsets and frames* (O&F), has been extensively evaluated on various datasets and has been shown to overcome the rest in almost every piece [25, 67]. Recently, it has been shown that AMT models could be enhanced by pre-stacking a U-Net [65]. U-Nets were first used for image segmentation and then for audio source separation. By pre-stacking a U-net, the network tends to learn knowledge regarding the sparse structure of the spectrogram-like input representation. However, we are more interested in understanding how audio-based AMT differs from score-informed AMT; intuitively, since score-informed AMT models exploit more information, the output should be more accurate. Thus, we compared O&F with a score-informed model (SI) which we developed based on a previous work [27].

In SI, inputs are a non-aligned score and an audio recording, while the output is a list of MIDI notes, each associated with onset, offset and velocity. SI performs audio-to-score alignment using a method based on Dynamic Time Warping that improves a previous system for piano music [35], and subsequently executes a Non-negative Matrix Factorization as source-separation method for each piano note. Then, it employs a neural network for estimating the velocities of each aligned note using as input the spectrogram of the note, computed thanks to the source-separation. Since the SI method requires the score, one was obtained from the World Wide Web for each of the 5 excerpts. For further details see the *Supplementary Materials*.

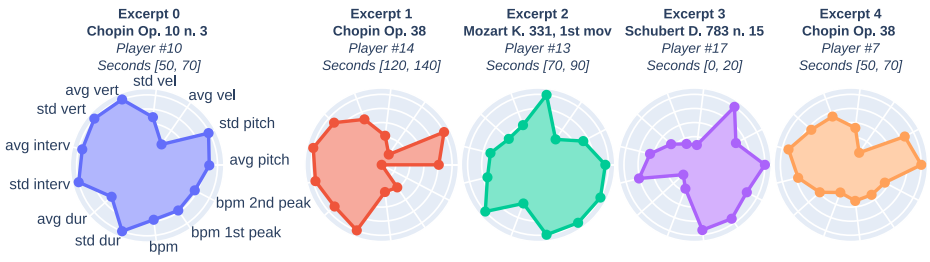

Fig. 5 Features extracted from the chosen excerpts; they are normalized in [0,1]. Seconds are relative to the audio recordings

Both the AMT systems predict pitches, onsets, offsets and velocities, while no other MIDI parameter (e.g., pedaling, etc.) is considered.

4.5 Synthesis and context selection

After producing the MIDI files, we synthesized them using 4 high quality different virtual pianos: a) the free *Salamander Grand Piano*,[3] b) two Pianoteq instruments freely available for research purposes (*Grand Steinway B* and *Grand Grotrian*), and c) the *Steinway* piano from Garritan Personal Orchestra 4. We post-processed every synthesized MIDI using SoX[4] in order to add reverberation using two different settings (values 50 and 100 of the SoX's `reverb` option). Thus, 12 different contexts, i.e. 4 without and 8 with reverberation, were formed.

We synthesized all the MIDI files with each context obtaining 12 different sets of audio clips. We extracted 13 MFCC and 7 rhythmic descriptors from each audio clip and computed the mean features to represent each context. We chose the medoid context for the *transcription* task and the most distant context from the average features of the original audio recordings for the *restoration* task. For the *resynthesis* task, two contexts were needed: in this case PCA explaining 99% of variance was applied to obtain a 10 dimensional representation from the original 20, and then searched for the two farthest points in the feature space based on the euclidean distance.

As a result of this process, the selected contexts were:

a) the Pianoteq *Grand Steinway* with SoX reverb set to 100 for the transcription task (44.1KHz/16bit stereo audio for both reference and candidates);
b) the Pianoteq *Grand Steinway* without reverb and the *Salamander Grand Piano* with SoX reverb set to 50 for the resynthesis task, candidates and reference respectively (44.1KHz/16bit stereo for the reference and 44.1KHz/24bit stereo for candidates);
c) the *Salamander Grand Piano* without reverb for the restoration task (44.1KHz/16bit mono for the reference and 44.1KHz/24bit stereo for candidates).

Note that no perceptual difference is known between 24 and 16 bit depth at the same sample rate [38, 39, 47].

5 Results

The listening test was communicated through mailing lists, chats, university courses, etc.; in total, 91 subjects responded to the entire test. Thanks to JavaScript-based WAET, we observed the subjects' behavior during the test, so that we were able to discard the answers where subjects listened to the excerpt for less than 5 seconds or where they did not move the cursor. After such filtering, we obtained more than 40 answers per question. Since these did not result in enough answers for each class of the initial questionnaires described in Section 3.3, we focused on two groupings only:

– subjects listening to classical music less than 1 hour per week (50) vs. subjects listening to classical music more often (41);

[3] free as in speech: https://musical-artifacts.com/artifacts/533.
[4] sox.sourceforge.net.

– subjects who have never studied music/hobbyists (57) vs. subjects who studied music professionally or having a degree or working as musicians (34);

We observed a general trend of non-experts providing higher ratings, meaning that, with respect to more experienced listeners or musicians, they rated candidates to be more like the same interpretation as the reference. However, this difference was not always statistically significant, thus not useful for the sake of our research questions. According to the literature discussed in Section 3, we could expect that by using longer excerpts the difference in the ratings would become significant and that expert listeners could give more accurate ratings.

We collected an imbalanced number of answers per type of headphones, namely 22 for headphones costing less than 20 euros and 69 for headphones costing more than 20 euros. Since we have found contrasting studies in literature about possible correlations between headphone retailing cost and sound quality, we have decided to disregard this factor during the successive analysis [7, 24].

During control group based analysis, we had more than 20 answers available per question and control group. We first applied Shapiro-Wilk normality tests with Bonferroni-Holm correction and $\alpha = 0.05$ to the collected responses for each control group, question, and method. We observed that the null-hypothesis – i.e. the collected answers are normally distributed – was rejected depending on the method and on the question. Consequently, we have performed the whole statistical analysis with both parametric and non-parametric methods and leave to the reader the ability to decide which test should be taken into account based on the single case. The following discussion and conclusions, however, hold in both the two cases – i.e. parametric and non-parametric analysis.

We computed error margins at 95% of confidence, that is a quantification of the accuracy for the estimated mean $\hat{\mu}$: we can say that by resampling the distribution, 95% of the collected populations will have the mean in $\hat{\mu} \pm e$, where e is the error margin. Error margins were computed with both normal distribution assumption [55] – i.e. parametric estimation – and bootstrapping methods [8] – i.e. non-parametric estimation. The full set of error margins is available in the *Supplementary Materials*. For our discussion, we can say that parametric and non-parametric error margins were rarely different when rounded at the 2nd decimal digit and that they ranged between 2% and 17%. Without using control groups, the error margins were between 3% and 10%, and between 2% and 5% when we average the ratings over the questions of the same task.

We analyzed results using ANOVA and Kruskal-Wallis tests with $\alpha = 0.01$ and we rejected the null hypothesis in all considered questions. We further analyzed the data using the Wilcoxon and the Student's t-test for related variables with the Bonferroni-Holm correction and $\alpha = 0.05$. In case the test condition is not satisfied, we cannot reject the hypothesis according to which the perception of two candidates is explainable by the same model. This is usually observable for non-expert subjects. Table 2 shows a general overview of the p-values computed for each pair of methods.

Figure 6 illustrates the average ratings for every task. The HR and the NR were recognized in all tasks, and O&F performed always better than NR and SI. In a post-experimental

Table 2 Number of questions with $p > 0.05$ for each pair of methods for both Sudents' t and Wilcoxon tests

SI	4		
O&F	6	0	
HR	7	1	9
	NR	SI	O&F

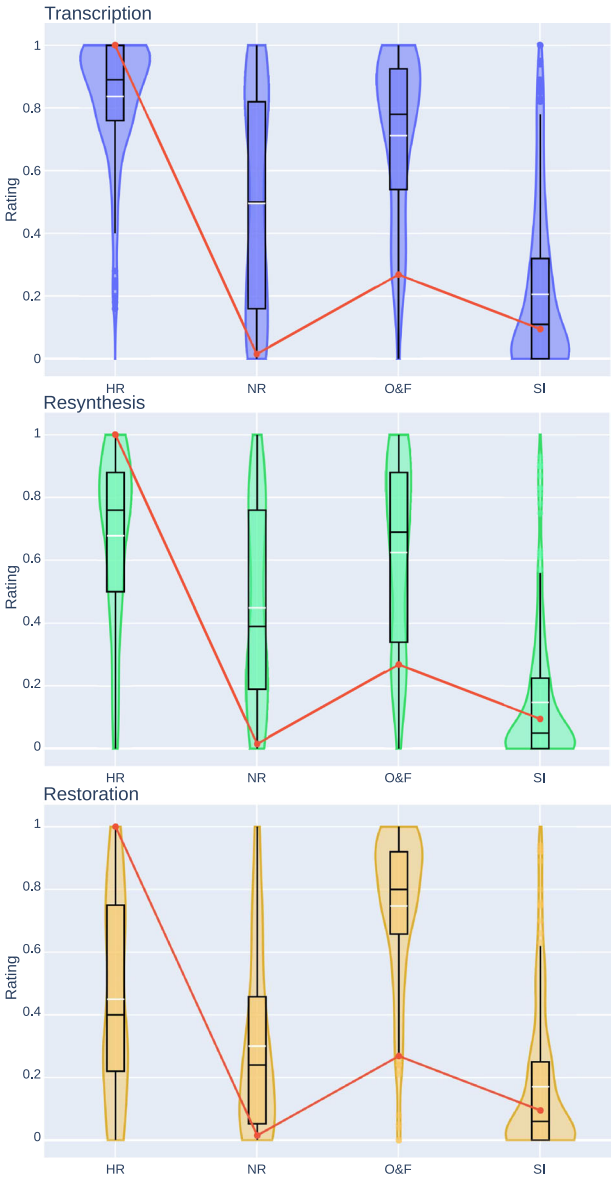

Fig. 6 Ratings per task averaged over all the excerpts. The red line identical in all tasks is the objective F-measure. White horizontal line is the mean, the black horizontal line is the median. All plotted distributions pass the pairwise significance tests against the other distributions in the same task, except for O&F and HR in the resynthesis task. Plots are created with [26]

interview, several subjects reported that SI was hardly comparable to the reference due to bad alignment. Indeed, in some excerpts, notes were distributed by SI in a very short time, producing a correspondingly long-lasting silence; this was often caused by missing/extra notes appearing in the music score. Thus, we conclude that the main reason for which SI was always rated worse than O&F is related to misalignments; we can consequently answer

the third research question by stating that score-informed approaches are generally limited by the alignment stage and that, as of now, monomodal AMT approaches provide improved performance assessed from a perceptual point of view.

In the *restoration* task, O&F was rated higher than HR. Since this behavior is not observable in the *transcription* and *resynthesis* tasks, and since the MIDI files were identical throughout all tasks, we attribute this outcome to the specificity of the *restoration* task, where the HRs (MIDI) were different from the references (audio), unlike the other tasks. In particular, at the time of writing, performance annotations in the *ASMD* framework do not include information about the pedaling used by the players, and recorded in the audio. Thus the HRs in the *restoration* task were synthesized from MIDI with no sustain control changes, whereas the audio references contained them. On the other hand, O&F is not able to transcribe the pedaling, but its authors enlarged the duration of sustained notes in the training ground-truth, so that the prediction of note duration is temporally tied to the duration of the resonance of the note rather than the onset/offset of the key. Such a durational enlargement allows O&F (and SI as well, as it uses O&F for the alignment) to predict duration perceptually more accurate than the HR in the *restoration* task.

The *resynthesis* task is also worth some further discussion. In this task, HR and O&F are perceived similarly, especially by non-expert listeners, and there are no statistically significant differences between the distributions of their ratings – see Fig. 6. Even though HR is rated slightly higher than O&F, it can be noted to score lower than in the *transcription* task. This suggests that the whole reference interpretation was hardly recognised in the HR, and that part of the interpretation was perceptually lost. Based on this outcome, we can try to address negatively the first research question, that is: when the context changes, MIDI representation seems not adequate to reproduce the same interpretation. However, other experiments are needed to confirm this hypothesis.

Analysing results question by question, we discovered an interesting behavior in the *transcription* task for excerpt 3 – see Figs. 5 and 7. There, O&F is associated with lower ratings than NR with $p = 0.38$ for Wilcoxon test and $p = 0.47$ for Student's t-test, meaning that the transcription is so inadequate that another interpretation resembles better the original one. Similar results are derived when investigating excerpt 0 in *resynthesis* and *transcription* task, where NR and O&F present almost identical ratings ($p > 0.23$). This behavior is more evident when looking at less expert listeners. Such results can also answer negatively

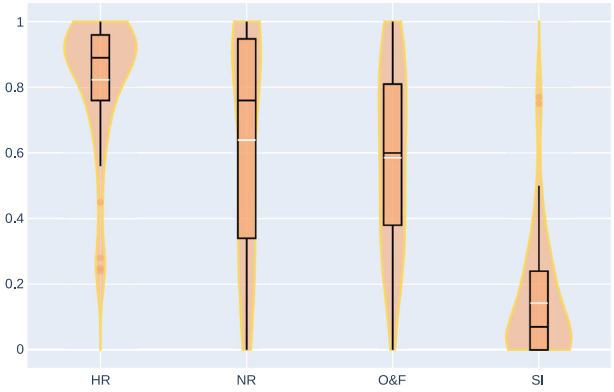

Fig. 7 Ratings for Excerpt 3 in the Transcription task. For this excerpt, all distributions pass the pairwise significance test except O&F and NR. Plots are created with [26]

to the second research question: the state of the art for piano AMT may not be able to extract parameters usable for reproducing the original interpretation, regardless changes in context (resynthesis and transcription tasks).

6 A new measure

Having answered our research questions, we looked for correlations between ratings and typical measures used for evaluating AMT systems. In particular, we adopted the widely-used measures available in a Python package [44]. In Fig. 6, the red line represents the F1-score computed considering as matches notes whose parameters lie within a certain range around the true value; in this case, we considered as parameters 1. the onset and offset times with a range of $\pm$ 50ms, 2. the linearly re-scaled velocity so that the L2 error is minimized with tolerance of 10%, and 3. the pitch with tolerance of 1 quarter-tone [25]. This measure, hereafter denoted as *OBJ*, demonstrated low correlation with subjective ratings, mainly due to the following factors:

- excerpt 0 has a low OBJ rating for O&F (almost 0); however, subjects rated it much higher than O&F,
- in the restoration task, O&F received higher ratings than HR, which always has OBJ equal to 1 – see Section 5, and,
- occasionally, O&F was rated lower than NR, which always has OBJ equal to 0.

Another interesting measure, named *PEAMT*, reflecting subjective ratings was proposed in [67]. We computed Pearson and Spearman correlation coefficients between OBJ and PEAMT measures and the median and average values of the collected ratings. It was discovered that the average generally presents slightly higher correlation than the median, therefore the former was used for subsequent analyses. Table 3 shows that PEAMT correlates more strongly than OBJ to subjective ratings, especially for the Pearson coefficient. Nevertheless, especially in the *restoration* task, correlation remains poor, motivating the search for an alternative measure.

We considered the features already used for the excerpt selection phase – see Section 4 – except for audio-based features to be consistent with the existing evaluation methodologies. We computed BPMs using the MIDI representation by counting how many onsets were present in sliding windows of size 0.1, 1.0, and 10 s with a hop-size of 50%. More precisely, we counted the mean and standard deviation for each window size. Importantly, to improve the portability of our measure, the features were first standardized using parameters computed on a large set of piano MIDI created by extracting piano solo performances from various datasets using *ASMD* [52]. After standardization, features of the predicted performances were subtracted from the target and the OBJ measure was appended. We performed linear regression on the dataset that we collected using various methods: Bayesian Ridge, Automatic Relevance Determination, Lasso Lars, Lasso, ElasticNet, Ridge and basic Linear regression. ElasticNet provided the best performance in terms of average L1 error. To further improve the generalization ability, we trained a model using ElasticNet while removing features with low weights, i.e. < 0.1.

We finally measured the average L1 error in a leave-one-out experiment, which provided 0.12 for our measure, 0.19 for PEAMT and 0.34 for OBJ. Table 3 (bottom rows) shows correlation coefficients for each task and measure.

Table 3 Correlations of various measures with the average rating of the subjects. Values are percentages

			OBJ	PEAMT	Ours
		min	49	62	**95**
		max	78	**99**	97
	Pearson	avg	75	89	**97**
		min	40	40	**80**
		max	**100**	80	80
Transcription	Spearman	avg	**80**	**80**	**80**
		min	41	64	**95**
		max	73	**100**	99
	Pearson	avg	66	85	**98**
		min	40	**60**	**60**
		max	**100**	**100**	80
Resynthesis	Spearman	avg	**80**	**80**	**80**
		min	-4	55	**78**
		max	57	94	**100**
	Pearson	avg	29	78	**89**
		min	0	0	**80**
		max	60	60	**100**
Restoration	Spearman	avg	60	60	**100**
	Pearson		45	71	**85**
Average (leave-one-out)	Spearman		44	54	**74**

When comparing PEAMT and our measure, one needs to consider differences in the design of the related tests. PEAMT is based on a test using audio clips lasting 5 – 10 s, while, following the discussion summarized in Section 4, this work employs clips lasting 20 s. PEAMT authors' created 150 questions and collected 4 answers for each one; we instead preferred to collect more than 20 answers per question for plurality, while considering a control group which led us to reduce to 15 the total number of questions. At the same time, the space of possible note combinations is covered optimally (see Section 4). Furthermore, PEAMT is based on categorical questions – subjects could chose between two systems – with no HR and NR, while we measured a linear variable and included hidden and negative references. Finally, we focused on changing the context of the recordings and synthesis, but we included in our Transcription task the scenario used by PEAMT.

In general, we can state that PEAMT results agree with ours in finding low correlations between subjective ratings and OBJ, and that the two evaluation measures that we have built are rather similar in our preliminary tests. However, our test highlights new aspects that we think fundamental for audio restoration and that only our measure is able to tackle.

7 Conclusion

After conducting a thoroughly designed perceptual test, this work proposed a new approach for audio restoration: in the light of recent developments in audio signal processing, it

becomes imaginable to recreate performances in the real world or through virtual instruments. We have therefore designed a perceptual test to assess to which extent existing technologies allow for such a methodology. It was discovered that the main limit lies in the usage of the MIDI format itself. Nonetheless, we proposed a new evaluation measure that seems consistent with the perception of context changes.

In case Standard MIDI Format is used as basis for the resynthesis, knowledge regarding contextual factors is required. Consequently, we argue that the future challenge for resynthesis-based audio restoration is in the conversion of the existing audio and music score in a new restored audio without the use of mid-level representations such as MIDI.

In this work, we have also identified limits for score-informed AMT, that, despite exploiting more information, lacks an effective feature fusion stage. Audio-to-score alignment should therefore become a main challenge for score-informed AMT; overcoming this problem could lead to improvements related to the exact knowledge of pitches and timings, leaving space for focusing on other parameters. Finally, we proposed a generic method to meaningfully choose excerpts when conducting music listening experiments.

Future works include further experiments to assess the proposed experimental strategy and to confirm the presented results, as well as the development of new technologies for feature fusion and context-aware music transcription, such as Paraconsistent Feature Engineering [23], Wavelet transforms, and novel audio-to-score alignment methods [53]

8 Reproducibility and Supplementary Information

To the sake of reproducibility, the whole code used for excerpt creations, answer collection, and statistical analysis is available online at the web address https://github.com/LIMUNIMI/PerceptualEvaluation.

Two Supplementary files are also provided:

1. `supplementary01.pdf`: contains the detailed description of the SI method used in this work – see Section 4.4;
2. `supplementary02.pdf`: contains extensive screenshots of the statistical analysis report used in this work. All the screenshots are generated using the code made available at the above URL. Specifically:
   - the analysis of all responses – no control groups – per each task, averaged across the excerpts (page 1) and not (page 2);
   - the analysis of the expertise control groups, averaged across the excerpts (page 3) and not (page 4);
   - the analysis of the listening habits control groups, averaged across the excerpts (page 5) and not (page 6).

Acknowledgment We greatly acknowledge the support of Pianoteq for providing a license of the physical modeled virtual piano; NVIDIA Corporation for the donation of a Titan V GPU; Julia Project, Python Software Foundation, and Linux Foundation for their invaluable work.

References

1. Akbari M, Cheng H (2015) Real-time piano music transcription based on computer vision. IEEE Trans Multimedia 17(12):2113–2121
2. Alías F, Socoró J, Sevillano X (2016) A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Appl Sci
3. Benetos E, Dixon S, Duan Z, Ewert S (2019) Automatic music transcription: An overview. IEEE Sig Proc Magazine, 36(1)
4. Bigand E, Vieillard S, Madurell F, Marozeau J, Dacquet A (2005) Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition and Emotion
5. Bogdanov D, Serra X, Wack N, Gómez E, Gulati S, Herrera P, Mayor O, Roma G, Salamon J, Zapata J (2013) Essentia: An open-source library for sound and music analysis. ACM Int Conf Multimedia
6. Bolzinger S, Warusfel O, Kahle E (1994) A study of the influence of room acoustics on piano performance. Journal De Physique Iv 4
7. Breebaart J (2017) No correlation between headphone frequency response and retail price. The Journal of the Acoustical Society of America 141(66):EL526–EL530
8. Chernick MR, González-Manteiga W, Crujeiras RM, Barrios EB (2011) Bootstrap Methods. Springer, Berlin, pp 169–174
9. Contardo C (2020) Decremental clustering for the solution of p-dispersion problems to proven optimality. INFORMS Journal on Optimization
10. Dannenberg RB (2006) The interpretation of MIDI velocity. ICMC
11. Davies S, Sadie S (2001) Interpretation. Grove Music Online. https://doi.org/10.1093/gmo/9781561592630.article.13863
12. Devaney J, Mandel MI (2017) An evaluation of score-informed methods for estimating fundamental frequency and power from polyphonic audio. ICASSP
13. Dreyfus L (2020) Beyond the interpretation of music. J Musicol Res. https://doi.org/10.1080/01411896.2020.1775087, https://doi.org/10.1080/01411896.2020.1714442
14. Erkut E (1990) The discrete p-dispersion problem. Eur J Oper Res
15. Everitt BS, Landau S, Leese M, Stahl D (2011) Hierarchical clustering. chap. 4
16. Feiten B, Wolf I, Oh E, Seo J, Kim H-K (2005) Audio adaptation according to usage environment and perceptual quality metrics. IEEE Trans Multimedia 7(3):446–453
17. Fu Z, Lu G, Ting KM, Zhang D (2011) A survey of audio-based music classification and annotation. IEEE Trans Multimedia 13(2):303–319
18. Gabrielsson A (1999) The Performance of Music, second edition edn., chap. 14, p. 501602. Cognition and Perception. Academic Press
19. Garí SVA, Kob M, Lokki T (2019) Analysis of trumpet performance adjustments due to room acoustics
20. Geringer JM, Johnson CM (2007) Effects of excerpt duration, tempo, and performance level on musicians ratings of wind band performances. J Res Music Educ
21. Godsill SJ, Rayner PJW (1998) Digital Audio Restoration. Springer, London
22. Goebl W (1999) The vienna 4x22 piano corpus. https://doi.org/10.21939/4X22
23. Guido RC (2019) Paraconsistent feature engineering [lecture notes]. IEEE Signal Proc Mag 36(1):154–158
24. Gutierrez-Parera P, Lopez JJ (2018) Perception of nonlinear distortion on emulation of frequency responses of headphones. The Journal of the Acoustical Society of America 143(44):2085–2088
25. Hawthorne C, Elsen E, Song J, Roberts A, Simon I, Raffel C, Engel J, Oore S, Eck D (2018) Onsets and frames: Dual-objective piano transcription. ISMIR
26. Inc. PT (2015) Collaborative data science. https://plot.ly
27. Jeong D, Kwon T, Nam J (2020) Note-intensity estimation of piano recordings using coarsely aligned midi score. JAES 68
28. Jeong D, Nam J (2017) Note intensity estimation of piano recordings by score-informed nmf. Int Conf on Semantic Audio

29. Jillings N, Moffat D, De Man B, Reiss JD (2015) Web Audio Evaluation Tool: A browser-based listening test environment. SMC
30. Kalkandjiev ZS, Weinzierl S (2015) The influence of room acoustics on solo music performance: An experimental study. Psychomusicology 25(33):195–207
31. Kalkandjiev ZS (2015) The influence of room acoustics on solo music performances: An empirical investigation. Ph.D. Thesis, TU Berlin
32. Kato K, Ueno K, Kawai K (2015) Effect of room acoustics on musicians' performance. part ii: Audio analysis of the variations in performed sound signals. Acta Acustica united with Acustica 101(44):743–759
33. Klapuri AP (2004) Automatic music transcription as we know it today. Journal of New Music Research 33(3)
34. Kob M, Amengual Garí SV, Schärer Kalkandjiev Z (2020) Room effect on musicians' performance. pp 223–249, Springer International Publishing
35. Kwon T, Jeong D, Nam J (2017) Audio-to-score alignment of piano music using rnn-based automatic music transcription. SMC
36. Luizard P, Brauer E, Weinzierl S, Bernardoni NH (2018) How singers adapt to room acoustical conditions
37. Marinelli L, Lykartsis A, Weinzierl S, Saitis C (2020) Musical dynamics classification with cnn and modulation spectra. SMC
38. Mizumachi M, Yamamoto R, Niyada K (2017) Discussion on subjective characteristics of high resolution audio. Journal of The Audio Engineering Society
39. Mörtberg J-E (2007) Is dithered truncation preferred over pure truncation at a bit depth of 16-bits when a digital re-quantization has been performed on a 24-bit sound file?
40. Napoles J (2009) The effect of excerpt duration and music education emphasis on ratings of high quality children's choral performances. Bull Counc Res Music Educ
41. Naylor GM (1992) A laboratory study of interactions between reverberation, tempo and musical synchronization. Acta Acustica
42. Orcalli A (2001) On the methodologies of audio restoration. Journal of New Music Research 30(4). https://doi.org/10.1076/jnmr.30.4.307.7496
43. Potocan Z (2020) Aesthetic perception of the singing voice in relation to the acoustic conditions. Ph.D. Thesis, University of Ljubljana
44. Raffel C, McFee B, Humphrey EJ, Salamon J, Nieto O, Liang D, Ellis DPW (2014) Mir_eval: A transparent implementation of common mir metrics. ISMIR
45. Ravi SS, Rosenkrantz DJ, Tayi GK (1994) Heuristic and special case algorithms for dispersion problems. Oper Res
46. Replaygain 1.0 specification. http://wiki.hydrogenaud.io/index.php?title=ReplayGain_1.0_specification
47. Repp R (2006) Recording quality ratings by music professionals. In: ICMC, Michigan Publishing
48. Rizzi A, Antonelli M, Luzi M (2017) Instrument learning and sparse nmd for automatic polyphonic music transcription. IEEE Trans Multimedia 19(7):1405–1415
49. Salvador S, Chan P (2007) Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis
50. Schwarz D, Lemaitre G, Aramaki M, Kronland-Martinet R (2016) Effects of test duration in subjective listening tests. ICMC
51. Simonetta F, Ntalampiras S, Avanzini F (2019) Multimodal Music Information Processing and Retrieval: Survey and Future Challenges. Int Work on Multilayer Music Representation and Processing
52. Simonetta F, Ntalampiras S, Avanzini F (2020) Asmd: an automatic framework for compiling multimodal datasets. SMC
53. Simonetta F, Ntalampiras S, Avanzini F (2021) Audio-to-score alignment using deep automatic music transcription. In: Proceeddings of the IEEE MMSP 2021
54. Storm W (1980) The establishment of international re-recording standards. Phonographic Bulletin
55. Tanur JM (2011) Margin of Error. Springer, Berlin Heidelberg, pp 765–765
56. Ternström S (1989) Long-time average spectrum characteristics of different choirs in different rooms. Voice (UK) 2:55–77
57. Ueno K, Kanamori T, Tachibana H (2005) Experimental study on stage acoustics for ensemble performance in chamber music. Acoust Sci Technol 26(44):345–352
58. Ueno K, Kato K, Kawai K (2010) Effect of room acoustics on musicians' performance. part i: Experimental investigation with a conceptual model. Acta Acustica united with Acustica 96(3333):505–515
59. Ueno K, Tachibana H (2005) Cognitive modeling of musician's perception in concert halls. Acoust Sci Technol 26(22):156–161

60. Von Békésy G (1968) Feedback phenomena between the stringed instrument and the musician. Rockefeller University Press
61. Wang S, Ewert S, Dixon S (October 2017) Identifying missing and extra notes in piano recordings using score-informed dictionary learning. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 25(10)
62. Wapnick J, Ryan C, Campbell L, Deek P, Lemire R, Darrow A-A (2005) Effects of excerpt tempo and duration on musicians' ratings of high-level piano performances. J Res Music Educ
63. Weinzierl S, Lepa S, Schultz F, Detzner E, von Coler H, Behler G (2018) Sound power and timbre as cues for the dynamic strength of orchestral instruments. The Journal of the Acoustical Society of America, 144(3)
64. Williams M (2016) Effect of excerpt duration on adjudicator ratings of middle school band performances. Research Perspectives in Music Education
65. Wu Y, Chen B, Su L (2019) Polyphonic music transcription with semantic segmentation. ICASSP
66. Xu M, Wang Z, Xia GG (2019) Transferring piano performance control across environments. In: ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 221–225
67. Ycart A, Liu L, Benetos E, Pearce MT (2020) Investigating the perceptual validity of evaluation metrics for automatic piano music transcription. TISMIR
68. Zapata JR, Davies MEP, Gómez E (2014) Multi-feature beat tracking. IEEE/ACM Trans on Audio, Speech, and Language Processing