Performance Assessment Technologies for the Support of Musical Instrument Learning

Vsevolod Eremenko, Alia Morsi, Jyoti Narang, and Xavier Serra[a]

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

Keywords: Music Education, Music Performance Analysis, Music Assessment, Audio Signal Processing, Machine Learning, Music Information Retrieval.

Abstract: Recent technological developments are having a significant impact on musical instruments and singing voice learning. A proof is the number of successful software applications that are being used by aspiring musicians in their regular practice. These practicing apps offer many useful functionalities to support learning, including performance assessment technologies that analyze the sound produced by the student while playing, identifying performance errors and giving useful feedback. However, despite the advancements in these sound analysis technologies, they are still not reliable and effective enough to support the strict requirements of a professional music education context. In this article we first introduce the topic and context, reviewing some of the work done in the practice of music assessment, then going over the current state of the art in performance assessment technologies, and presenting, as a proof of concept, a complete assessment system that we have developed for supporting guitar exercises. We conclude by identifying the challenges that should be addressed in order to further advance these assessment technologies and their useful integration into professional learning contexts.

1 INTRODUCTION

Learning to play a musical instrument is a complex process that requires the acquisition of many interlinked skills. An expert musician is able to play with beautiful tone, intonation, note accuracy, rhythm precision, clear articulation, dynamic variation, and expressive inflection (Duke and Byo, 2012). But even more, a musician is able to execute all of them at the same time and in the context of music making. These skills need motor and cognitive capabilities that can only be acquired with time, practice, and commitment.

Two main contexts within which formal musical learning takes place are the classroom, through playing and interacting with a teacher, and at home, practicing alone. There are also informal contexts that are important for learning, like playing with friends or listening to music, but we will focus on the basic formal contexts. In the typical classroom setting, the teacher defines what the student should practice and gives feedback on what has been practiced. At home, the student plays following the teacher's advice and develops practicing habits adequate for the goals to achieve.

A software application cannot substitute a music teacher but it can be a complement in the learning process, especially in supporting the daily practice. There are many ways in which technologies can complement traditional formal learning and support practice. In the simplest form, recording and playing back a performance is available to everyone with a mobile, and it is a useful way to develop a more analytical listening perspective. Moreover, just the ability to have digital scores on a tablet computer and being able to share and annotate them digitally is quite valuable. Some software applications support even more specialized functionalities, like organizing the material needed for practice, mainly scores; offering music minus options, thus being able to play along with an audio accompaniment; or guiding the practice in a way that promotes engagement. In this article we are specially interested in the technologies that analyze audio recordings for assessing the student's playing.

Although performance assessment is only one of the aspects involved in the teaching of a musical instrument, it is a highly acknowledged manner of evaluating competence, which not only enables the assessment of student growth, but also the evaluation of the effectiveness of a learning program. It is cru-

[a] https://orcid.org/0000-0003-1395-2345

Eremenko, V., Morsi, A., Narang, J. and Serra, X.
Performance Assessment Technologies for the Support of Musical Instrument Learning.
DOI: 10.5220/0009817006290640
In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020)*, pages 629-640
ISBN: 978-989-758-417-6

Copyright © 2020 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

cial to identify the goals and purposes of assessments (Wesolowski and Wind, 2019; Pellegrino et al., 2015), because the design of an appropriate assessment strategy must correspond to its goals. We can talk about assessment *of* learning and assessment *for* learning (Mantie, 2019; Schneider et al., 2019; Wesolowski, 2014), thus we should differentiate the assessment metrics to measure competence from the ones to enhance learning, and also distinguish whether they are about improvement or accountability.

Next, we review some considerations and choices in music performance assessment practices, aiming to contextualize the scope in which we position the technological contributions of interest. Then, we go over the current state of the art in performance assessment technologies and present a case study of a complete assessment system that we have developed to support guitar exercises in an online course. We conclude by identifying some open research challenges that should be addressed in order to further advance these assessment technologies.

2 ASSESSMENT PRACTICES

To develop computational methods which can aid student learning by providing meaningful feedback, we must understand what should be borne in mind when developing assessment strategies, and how choices pertaining to a strategy affect its usefulness in terms of learner feedback. As already highlighted in section 1, the purpose and use are key decisions to which the design of an assessment task must align to (Schneider et al., 2019). We are interested in performance assessments for instructional purposes. In this section, we review various considerations in the development of performance assessment frameworks discussing the differing detail levels of expected feedback, and how they could affect subsequent choices of appropriate evaluation criteria and measurement tools.

2.1 Assessment Frameworks

Performance assessment is typically conducted through a set of tests, where a test could be defined as ‘the collection and interpretation of data representing a particular music behaviour using a systematic and uniform procedure’ (Wesolowski and Wind, 2019). There is much research discussing frameworks by which assessment systems should be created (Schneider et al., 2019; Wesolowski, 2012; Wesolowski, 2014). Designers of an assessment strategy should take the following conceptual decisions: a) what are the knowledge, skills, and attributes to be assessed; b) what is the evidence that will demonstrate them; and c) what tasks will be used to extract such evidence from learners (Schneider et al., 2019). It is out of our scope as music technologists to make the decisions on what any of the above 3 points should entail. Rather, our current research direction aims to connect the third point back to the first one, i.e., how does the student’s response to the defined tasks reflect their strengths and weaknesses within the knowledge, skills, and attributes that the test was designed to assess. In other words, how will we measure the skill levels or the progress of a student using their responses. Providing useful feedback to students would be a consequence of accurate documentation of performance measurement of skills.

2.2 Feedback for Instructional Benefit

Evaluations could be conducted with various levels of detail, ranging from evaluations based on an overall impression of the performance, to those based on more detailed criteria representing technical and expressivity levels, or to those based on micro-skills presented, such as adequacy of played notes, rhythmic correctness, appropriateness of dynamic changes, and quality of articulation (Mazur and Łaguna, 2017). The appropriateness of one level of detail over another is very context specific. For example, in music education literature, there is a distinction between assessments made for *formative* purposes from the ones for *summative* purposes.

Formative assessments, sometimes described as assessments for learning, are assessments done with the intention of supporting a learner’s process of improvement (Mantie, 2019). They provide detailed feedback information, thus are well suited for instructional benefit. They can be used to systematically adjust the instruction for improvement (Schneider et al., 2019; Wesolowski, 2014), and can foster a student’s self-assessment ability (Schneider et al., 2019), an important skill to enhance autonomy during practice. Summative assessments, or assessments of learning, are those conducted with the intention of making evaluative conclusions for reasons other than learner development (Mantie, 2019). They are commonly used for the quantification of learning acquired over a time period or educational unit. Typical examples of summative assessments would include auditions, placement tests, large scale assessments, and graded performances Given their ability to document the level of student achievement, they can provide information about the final status of a student’s development Evaluations conducted to support formative assessment practices must be geared towards a higher capacity for capturing details. In contrast, evaluations meant to support summative assessment goals would provide bigger picture overviews rather than a documentation of micro-skills.

2.3 Measurement Approach

Developing a measurement approach entails: a) selecting the evaluation criteria and b) choosing the type of measurement/assessment tool. These choices are not disconnected from one another, nor from the aforementioned issues on the purpose of a test and the required type of feedback.

2.3.1 Evaluation Criteria

Conceptually, the broad content areas (dimensions) to be assessed are determined, then test items (criteria) representing each of the dimensions are chosen. Hallam and Bautista (2012) outline a list of skills acquirable when learning an instrument, including expected ones such as aural skills, technical skills, and performance skills to more abstract ones such as evaluative skills and self regulatory skills. Each of these skills could be observed through different behaviours. For example, the aural skill level of a student could manifest in several musical behaviours, including a student's ability to play by ear, and their sense of rhythmic accuracy (Hallam and Bautista, 2012). Therefore, to measure the aural skill level from a test, dimensions such as pitch and rhythm would be relevant. Choosing criteria reflecting these dimensions would depend on characteristics of the performer (i.e. the capabilities of the instrument) and the evaluation (i.e. the task and repertoire) (Mazur and Łaguna, 2017), where the research efforts mentioned in section 2.1 regarding frameworks for developing assessments would offer insight. One must also take into account the effect of criteria choices on the validity, reliability, and fairness of the assessment (Wesolowski and Wind, 2019). Practically, this process is far from straightforward, and it should be noted that vague or narrow criteria can result in the mismeasurement of student learning (Brookhart, 2018).

An approach to addressing difficulty and inherent subjectivity in determining the dimensions and criteria pertaining to assessment situations is facet factorial analysis, which has been applied for building measurement scales. Such approaches aim to identify performance factors that affect music performance assessment. Russel (2010) applies this approach to develop a general guitar rating scale, concluding that the performance criteria gathered in the study are best divided into 5 categories, which are: interpretation, tone, rhythm, technique, and intonation. In general, works of this type could help in defining evaluation criteria for assessments.

2.3.2 Measurement Tools

In addition to choosing the evaluation criteria, the type of measurement tool/instrument for capturing student level within such criteria needs careful thought and selection. In most sources, the use of checklists, rating scales, and rubrics is discussed. Techniques are sometimes combined or used in conjunction to develop an assessment strategy, as mentioned in Pellegrino et al. (2015).

Checklists: are lists of specific characteristics through which an evaluator can make a dichotomous decision (yes/no, or absent/present) regarding the existence of a criterion (Brookhart, 2013). They are suitable in cases where learning outcome is demonstrable merely by the absence or presence of objectives (Wesolowski, 2014), or when we want to communicate requirements to be followed. However, their instructional value is limited because of the responses are not indicative enough of exact student level, or what should be improved (Brookhart, 2013). Therefore, they are better suited for summative assessments (Pellegrino et al., 2015).

Rating Scales: differ from checklists in that they allow the indication of degrees to which the evidence of a particular criterion is displayed, thus capturing levels of proficiency (Brookhart, 2013; Wesolowski, 2014), so they are more beneficial from a developmental standpoint. Such proficiency degrees could be captured by frequency ratings or quality ratings (Brookhart, 2013), where a Likert-type scale could be appropriate to use. Rating scales are suitable for summative assessments, and in some cases they could be used for formative assessment as well. However, rating scales that capture student responses using quality ratings do not have a high instructional benefit because, they give judgement without a description of the evidence (Brookhart, 2013). However, augmenting a classic rating scale with comment sections would allow the addition of explanations or improvement feedback, making them much more useful from a formative sense (Pellegrino et al., 2015).

Rubrics: are defined to have two things: "criteria that express what to look for in the work, and performance level descriptions that describe what instantiations of those criteria look like in work at varying quality levels, from low to high" (Brookhart, 2018). Performance level descriptions are what differentiate rubrics from rating scales or checklists (Brookhart 2013, 2018). In addition, they are what make rubrics superior to their counterparts in terms of objectivity

(Wesolowski, 2012), and in their suitability for formative assessment goals (Wesolowski, 2014).

Rubrics can be classified by their generality (*general* or *task-specific*), and by their grouping of criteria (*holistic* or *analytic*). A general rubric is applicable to a family of related tasks, whereas a task-specific rubric is only applicable to a single task, since the proficiency level descriptions in the latter reflect characteristics that are only relevant in the context of a specific task (Brookhart, 2018). Analytic rubrics keep criteria separate, thus allowing feedback to be given on each alone, and are suitable when different performance elements should be measured separately. On the other hand, holistic rubrics group different criteria, and are suitable when evaluators perceive that criteria should be considered conjunctively. In holistic rubrics, the performance level descriptions address the grouped criteria simultaneously, whereby a single decision is made for the whole group (Brookhart, 2018; Schneider et al., 2019). Holistic rubrics are more practical when the focus of assessment is grading a student's overall level through a global rating of the performance rather than providing detailed feedback. Therefore, they are suited for summative assessments. Analytic rubrics better support formative assessments given the granularity of provided information.

2.4 Automating Feedback

Despite the plethora of pedagogical implications relating to assessment practices, it certainly is possible to take positive steps towards assessment automation, which is the goal of our research. Automatic measurement of student skill regarding a set of criteria is useful, even if at the moment such criteria are only applicable for the assessment of simple exercises.

The prior discussion of different feedback forms used in performance assessments is meant to provide a framework through which such choices can be positioned within an automatic assessment system, such as the one presented in section 4. Moreover, determining the measurement tools and criteria relevant for an assessment situation is a guiding factor to choose the adequate technologies to support the creation of automatic systems. Our focus is on audio-based student input, and our goal is to create models that can automatically map a student's input to the criteria that we are interested in assessing. Section 3 presents a technological pipeline for translating audio input into measurements of student skill level within a set of criteria. It is important to define in signal processing terms what needs want to be detected from audio as a representation for such criteria, noting that some criteria (e.g., pitch accuracy) are more straightforward than others (e.g., postural errors).

3 ASSESSMENT TECHNOLOGIES

The assessment of a musical performance requires the capturing and analysis of the sound produced by the player. This sound analysis typically involves extracting the relevant performance characteristics, then comparing them with a predefined set of criteria, and finally reporting the assessment results in a way that is understandable by the player.

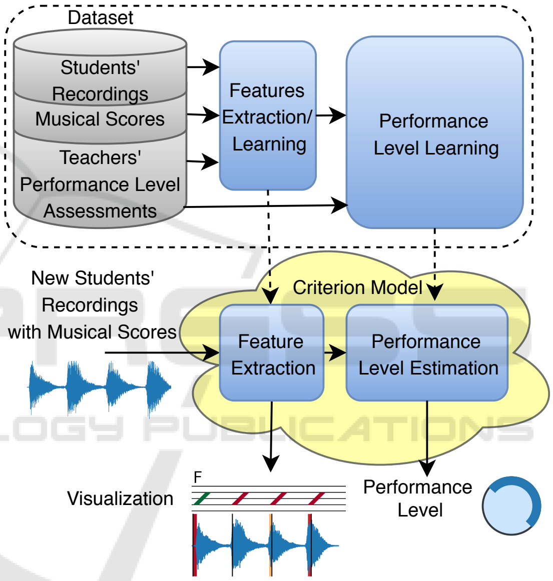

Dataset
Students' Recordings
Musical Scores
Teachers' Performance Level Assessments
Features Extraction/ Learning
Performance Level Learning
New Students' Recordings with Musical Scores
Criterion Model
Feature Extraction
Performance Level Estimation
Visualization
F
Performance Level

Figure 1: Proposed pipeline of a music performance assessment system. The top part (dotted line) includes the model creation part of the process (dataset, feature extraction, level learning) and the bottom part is the actual assessment of a recording (recording, feature extraction, level estimation).

Figure 1 shows a block diagram of our proposed assessment process based on audio analysis. This process includes two distinct parts: (1) the creation of the assessment model (top part) and (2) the actual assessment of a player performance (bottom part). The creation of the assessment model starts with an annotated audio collection (*Dataset*) designed for a specific type of performance exercise to be assessed, then it extracts the score-informed audio features relevant for the chosen task (*Feature Extraction/Learning*) and apply a machine learning method (*Performance Level Learning*) to capture the correspondences between audio recordings (*Students' Recordings*) and the specified assessments (*Teachers' Performance Level Assessments*). The output of the training part of the process is a *Criterion Model* that captures the characteristics of a good performance and includes the algorithms to do the audio analysis needed to model the criteria for different types of Measurement Tools (see 2.3.2). With this model, any new recording of a performance exercise type, *New Students' Recordings*, can be analyzed and assessed. The assessment results might include a *Visualization* of the student performance indicating the deviations from the reference performance as specified in the Dataset, and the Performance Level as predicted from the model.

We have grouped the audio analysis technologies of relevance to performance assessment according to the dimension to be evaluated: (1) rhythm, (2) pitch and chords, and (3) technique and expressivity.

3.1 Rhythm

Onsets, beat and tempo are rhythm related characteristics of a performance that are relevant in assessment and than can be studied from audio signals (Gouyon and Dixon, 2005).

Onset detection has been a basic task in rhythm analysis. For plucked or percussive instruments, like guitar, piano, or drums, detection of onsets is relatively easy, but for many other instruments, like singing voice or bowed strings, the current state of the art techniques still need improvements. Böck and Widmer (2013) used a spectral flux based model with vibrato suppression for non-percussive instruments (cello, violin, voice). The reported accuracy for cello and violin are quite good (greater than 80 percent), however, for singing, the accuracy is quite low (less than 70 percent).

The current state of the art techniques for beat estimation use dynamic bayesian networks (Krebs et al., 2015) and work quite well for dance music. For the related task of tempo estimation, convolutional neural networks using mel-spectrogram as features, have reported the best accuracy for some specific types of music (Schreiber and Müller, 2018). As it is typical with machine learning approaches, the developed models are limited by the datasets with which they have been trained.

3.2 Pitch and Chords

Pitch is a basic feature of a musical sound from which we can study many performance characteristics related to tuning, intonation, or note accuracy. Pitch is measured by detecting the fundamental frequency, but the specific signal processing approach to use is different depending on the type of signal. For example, pYIN (Mauch and Dixon, 2014) is a commonly used method in fundamental frequency extraction for monophonic music signals and Melodia (Salamon et al., 2014) is used for detecting the prominent pitch in polyphonic music fragments. The detection of several simultaneous pitches, i.e. multi-pitch detection, is much harder, but some useful results can be obtained with simple polyphonic signals (Klapuri, 2006). Other common approaches used to study pitch related characteristics are based on Chromagrams (Cho and Bello, 2014) which relate to the concept of pitch class profiles.

Chord analysis is related to multi-pitch estimation but the analysis methods used are typically different. Pauwels (2019) provides an excellent overview of chord detection methods and of the existing open challenges.

3.3 Technique and Expressivity

The analysis of the playing technique or of the expressivity of a performance requires more complex audio analysis methods. Machine learning methods are useful for capturing the subjectivity that the assessment of these playing characteristics require.

Technique refers to the control that a player has on the instrument and thus its study is instrument specific. The available studies have targeted a small number of instruments. For guitar, there have been efforts to extract playing techniques like bend, slide, pull-off and hammer (Reboursière et al., 2012). In a study on bass guitar, Abeßer et. al. (2010) implemented a feature based approach to distinguish 5 plucking techniques (finger-style, picked, muted, slap-thumb, slap-pluck) and 5 playing styles (normal, vibrato, bending, harmonics, dead-note). For flute, signal processing techniques were used to detect common mistakes in playing, like poor blowing or mis-fingering (Han, 2014). In another study on violin, support vector machines were used to identify common mistakes amongst novice players (Luo et al., 2015). The results of these studies look promising but more work is needed in this direction and they need to be adapted to use case of performance assessment.

The expressivity of a performance can be studied by analyzing musical characteristics such as dynamic and timing changes, articulations, or vibrato. Widmer (1998) applied machine learning to estimate expression of piano music by analyzing dynamics and tempo from captured MIDI data of a player piano. Maestre et.al. (2005) studied the expressivity of saxophone jazz recordings by extracting audio features like energy and pitch contour. But musical expression is difficult to characterise and not much has been done in the context of music education.

3.4 Performance Assessment

For assessing a music performance, once we have analyzed the different musical dimensions of a recording, we have to compare the obtained features with the ones of reference recordings or with predefined guidelines. Lerch (2019) offers an overview of state of the art performance analysis technologies, including their use in assessment.

Most approaches to performance assessment use expert knowledge to derive hand-crafted features followed by some classification algorithm to predict expert ratings. Vidwans et al. (2017) predicted expert ratings on the basis of musicality, note accuracy, rhythmic accuracy and tone quality for alto saxophone recordings, concluding that hand-crafted features based on expert knowledge did not suffice to predict the ratings. Pitch based features are commonly used in evaluation tasks. For example, Schramm et al. (2015) used them for solfege assessment, Molina et al. (2013) for evaluating sung melodies, and Bozkurt et al. (2017) for evaluating examinations conducted by Turkish music conservatory. Nakano et al. (2006) used pitch intervals and vibrato features to assess singing of unknown melodies. These approaches provide useful feedback for pitch or rhythm related characteristics, but they are constrained to specific use cases. The limitations of these methods prompted work on automatic feature learning to capture audio characteristics that are difficult to obtain using hand-crafted features (Wu and Lerch, 2018). Pati et al. (2018) applied convolutional and recurrent neural networks to predict expert ratings for wind instruments, like Flute, Alto Saxophone and Bb Clarinet. The results show some promise, however, they offer limited scope for usage in a real music education context.

To develop actual learning systems, we need software implementations of the presented analysis methods that can work efficiently. Essentia[1] (Bogdanov et al., 2013) is an open source library for audio analysis developed and maintained by our research group which contains state of the art and robust implementations of most of the algorithms needed to develop music performance assessment systems. Using it, we have been involved in the development of two music learning mobile apps which include different performance assessment approaches. Cortosia (Bandiera et al., 2016) (Romani Picas et al., 2015) focuses on measuring the tone quality of instrumental playing and Riyaz[2] teaches how to sing in the classical Indian music traditions, measuring the pitch accuracy of the users' singing. Well known commercial applications developed for supporting musical practice include Smartmusic[3] and Yousician[4] which also include the type of technologies mentioned here. However, given the current state of the art, they just use pitch and rhythm features to analyze the users' playing.

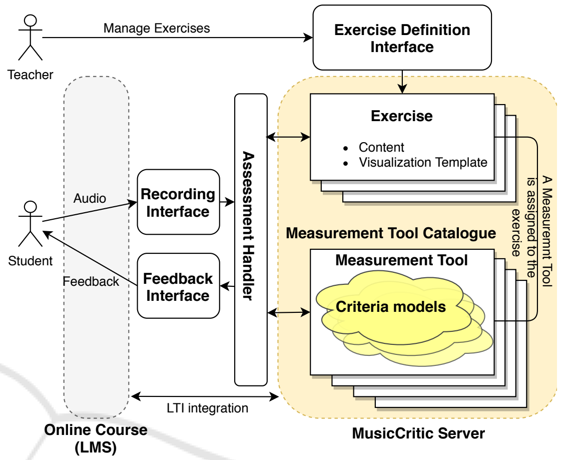


Figure 2: MusicCritic framework. A teacher defines an excercise, which is stored in the MusicCritic server, together with the analysis tools to assess it. A student submits a recording of an exercise through the Online Course platform, which is uploaded and assessed in MusicCritic server. The assessment results are sent to the Online Course platform for the student to view them.

4 CASE STUDY

To study the possibilities and challenges of applying the above mentioned concepts and technologies in a real educational context, we carried out a case study in which we developed a complete system for the formative assessment of guitar exercises. The system is based on Essentia and MusicCritic,[5] and was deployed and tested in the MOOC "Guitar for Beginners" by Berklee College of Music.[6]

Six exercises of the online course were chosen. Three of them are "picking" exercises (single line playing): an arpeggio exercise, a simple phrase exercise, and a chromatic scale. The other three are "strumming" exercises (chord playing): from two to

[1] https://essentia.upf.edu/

[2] https://riyazapp.com/

[3] https://smartmusic.com/

[4] https://yousician.com/

[5] http://musiccritic.upf.edu/

[6] https://kadenze.com/courses/guitar-for-beginners/

eleven chords played as even quarter notes, four beats per chord.

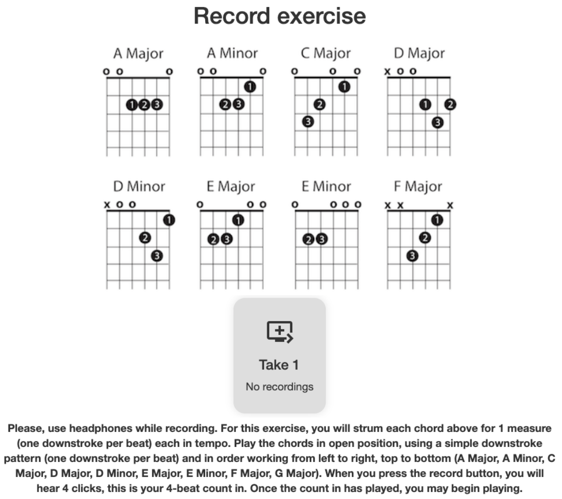
4.1 MusicCritic Framework

To create the pilot system we used MusicCritic, a software framework that we have developed to facilitate the assessment of musical exercises in the online education context (Bozkurt et al., 2018). It provides a way to create and assess tailored exercises, supporting personalized rubrics. Using a Web API, MusicCritic acts as an external music assessment tool via the Learning Tools Interoperability (LTI) standard, thus able to communicate with any Learning Management System (LMS), for example MOOC platforms.

Figure 2 shows a block diagram of the MusicCritic framework as it is used in an online education context. The teacher, through the *Exercise Definition Interface*, prepares an exercise, which is composed of three elements: *Content*, *Visualization Template*, and *Measurement Tool*. The exercise's content is a machine-readable representation of the exercise's ideal performance. For example, our guitar exercises include information about pitch, rhythm, and tempo, and plan to include information about technique and dynamics. The visualization template is a music score with placeholders to be automatically filled with the assessment results (e.g., note heads coloring or marks on waveform drawing). Finally, the *Measurement Tool* is a computational model based on a general Measurement Tool (section 2.3.2), specifying the feature extraction procedures and the automatic criteria estimators to be used. The MusicCritic Server includes a catalogue of *Measurement Tools* with varied *Criteria Models* that have been developed to analyze and assess different performance characteristics. The exercise specification is also used to configure the recording and feedback interfaces for that particular exercise. The actual assessment starts from the course site, where the student plays the proposed exercise (e.g., Figure 3), which is uploaded to the MusicCritic Server. Once the recording is assessed using the Measurement Tool, the report of the assessment, performance visualization plus performance level obtained, is sent to the course site for the student view (e.g., Figure 4).

4.2 Measurement Tool Design

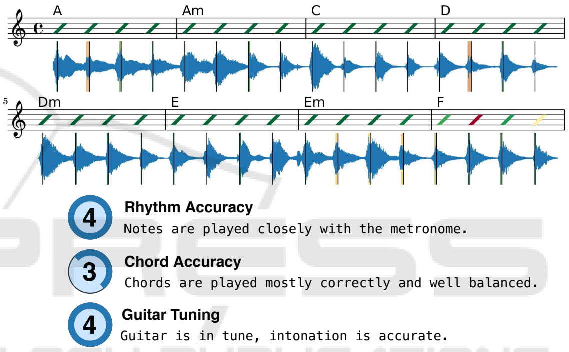
As explained in 2.3.2, a measurement tool is the core of the assessment process and captures the criteria to be assessed. To develop it for our case study, we first studied the course content and made some initial analysis of the type of playing errors that the students make on the chosen exercises. Then we organized the playing errors using the assessment dimensions proposed by Russel (2010): interpretation/musical effect, tone, technique, rhythm/tempo, and intonation. Within these dimensions we specified assessment criteria of relevance for the selected exercises. After a few iterations, and taking into account technical feasibility, we came up with the criteria shown in Table1.



Record exercise

Please, use headphones while recording. For this exercise, you will strum each chord above for 1 measure (one downstroke per beat) each in tempo. Play the chords in open position, using a simple downstroke pattern (one downstroke per beat) and in order working from left to right, top to bottom (A Major, A Minor, C Major, D Major, D Minor, E Major, E Minor, F Major, G Major). When you press the record button, you will hear 4 clicks, this is your 4-beat count in. Once the count in has played, you may begin playing.

Figure 3: Recording interface for a strumming exercise.



4 Rhythm Accuracy
Notes are played closely with the metronome.

3 Chord Accuracy
Chords are played mostly correctly and well balanced.

4 Guitar Tuning
Guitar is in tune, intonation is accurate.

Figure 4: Feedback interface of a strumming exercise with the assessment results, performance visualization plus performance level obtained.

After evaluating the implementation feasibility, we selected three criteria to focus on in the initial prototype: (1) closeness to the metronome, (2) notes/chords accuracy, and (3) tuning/intonation. We decided to use a rating scale with 4 levels for each criteria: from 1 (inaccurate, worst) to 4 (accurate, best) and wrote descriptive explanations for each criterion level. For example, a student who achieves pitch accuracy level 2 receives an explanation that "Incorrect or inaccurate notes are sometimes played," and for level 3: "Notes are played mostly correctly."

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 1: Criteria used in the assessment of the guitar exercises. In boldface are the criteria selected for automation.

Dimensions	Criteria
Rhythm/ Tempo	**closeness to the metronome**, rhythm pattern correctness
Pitch/ intonation	**notes/chords accuracy**, **tuning/intonation**
Technique	“explosive” attack quality (for fingerpicking strumming), attack sharpness and cleanness, legato quality
Tone	guitar tone beauty
Interpretation/ Musical effect	notes/chords loudness stability

4.3 Dataset

To develop and evaluate our audio analysis methods, we created a dataset of 233 recordings, including the 6 chosen exercises played by students with different skill levels, from absolute beginners to conservatory graduates. To cover diverse contexts we used various recording setups (hardware, OS, and browser), guitars (acoustic and electric), and amp setups for electric guitar (clean, overdriven). We involved two guitar teachers in the manual assessment of each recording, using 4 evaluation levels for the following criteria: Rhythm Accuracy (“closeness to the metronome”), Pitch Accuracy, Guitar Tuning, and Overall Performance. We are conscious that this dataset is small, but it was sufficient for developing our proof of concept.

4.4 Automatic Estimation of Criteria

For each assessment criteria selected, we devised the appropriate audio feature extraction method and evaluated its performance by using the created dataset. Since the assessment results should be properly explained in order to improve pedagogical effectiveness and help to develop trust in the assessment feedback (Conati et al., 2018), following Rudin (2019, May), our system is based on musically meaningful features, and the machine learning architectures used obey structural knowledge of the domain.

4.4.1 Rhythm

To estimate how “closely” a student plays to the metronome, the system measures the distances between note/chord onsets and expected metrical positions. Ideally we should identify and use perceptual attack times (Gordon, 1987), but for single guitar notes the difference between onset times and perceptual attack times is not large, so we used onset times. However, a played chord might have multiple onsets (Freire et al., 2018) and we decided to use as chord onset position the last onset present, typically the last picked string.

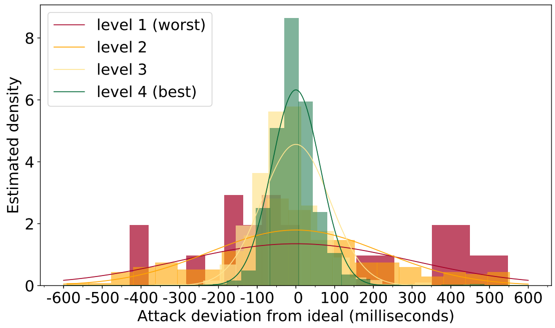


Figure 5: Histograms of onset deviations from ideal position for different performance levels.

To properly detect onsets from audio recordings made on a mobile phone or computer, our method had to be robust to issues such as: recording conditions, microphone quality, computer latencies, and interfering sounds. After trying several state of the art approaches, we decided to use a heuristic approach which involves a SpectralFlux onset detection function (Böck and Widmer, 2013), zeroing the output when the finite difference of the smoothed energy RMS (Schedl et al., 2014) is negative, to filter out string release sounds and other interfering noises. The actual onset positions are the peaks of the SpectralFlux function (Böck et al., 2012) and we used a threshold, based on the averaged spectral centroid (Schedl et al., 2014), to reject string squeak noises.

Results of our approach are shown in Figure 5. We used our dataset and computed the distance between measured onsets and metrical positions for each recording. We show a histogram for each of the 4 labeled performance levels. The distribution of the deviations for the highest performance level is peaked near zero and concentrated mostly between -100 and 100 ms. In contrast, for lower performance levels, the distribution is flatter and has a larger spread.

In order to assess a student’s recording, we measure the distances between onsets and metrical positions using the method proposed, showing them in the assessment visual feedback (Figure 4). Superposed to the waveform of the audio recording, black vertical bars designate metrical positions and the colored rectangles show the deviations. Green and yellow colors correspond to acceptable and tolerable deviations, while the red corresponds to substantial deviations. We predict the student’s performance level for this particular criterion applying isotonic regression

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

(Gupta et al., 2016) to the variance of the sampling distribution of individual note deviations.

4.4.2 Pitch

To analyze pitch accuracy of the played notes and chords, we model NNLS chroma vector probability distributions for each kind of musical event (i.e., single note or chord) based on a chord detection pipeline (Cho and Bello, 2014).

To make our system robust to changes of dynamics or noise, and to reduce the dimensionality of the chroma vectors, we model l^1-normalized chroma distribution as a product of two independent distributions: (1) the ratio of the sum of in-the-chord tone chromas to sum of out-of-the-chord tone chromas (“in/out tones ratio”), and (2) renormalized vector of in-chord tone chromas (“in-chord chromas”). From the distributions obtained by analyzing the dataset recordings, we can visualize the pitch accuracies of the different performance levels in Figure 6. The left-most picture shows that in/out tones ratio histogram has a prominent peak for the top performance level recordings and more flat for lower levels. Similarly, ternary plots (van den Boogaart and Tolosana-Delgado, 2013) of the joint distribution of in-chord chromas show a prominent peak for the top performance level recordings and bigger spread for recordings with lower levels.

For modeling “in/out tones ratio”, we use Beta distribution; “in-chord chromas” vector is transformed with additive log-ratio transformation (van den Boogaart and Tolosana-Delgado, 2013) and modeled with multivariate Gaussian distribution. During the training phase, we estimate parameters of these distributions for each musical event kind using only the recordings in our dataset having the best assigned performance level.

Then during the automatic assessment process, the posterior probability of a musical event given the observed chroma is used as the measure of pitch accuracy. Probabilities for individual events are used for the visual feedback as shown in Figure 4. The color of the note represents pitch accuracy, where green means high accuracy and red represents wrong notes or chords. The yellow and shades in-between represent various non-fatal errors, e.g., unnecessary strings are slightly ringing during single note picking or chord strumming, or chord is imbalanced.

The criterion performance level is predicted with isotonic regression, which uses mean squared log of individual events probabilities as the independent variable.

4.4.3 Guitar Tuning

Chroma features give us a semitone resolution, ignoring smaller frequency deviations, thus they are not adequate for identifying guitar tuning flaws and intonation errors. Our approach for tuning assessment is based on comparing spectral peaks with equal temperament values. Firstly, we sample from the intervals between the most prominent spectral peaks in each analysis frame (Schedl et al., 2014) and then calculate deviations of these intervals to the equal temperament grid. Figure 7 shows that the histogram of deviations to the equal temperament grid is peaked for top tuning performance level and square for low performance level, thus variance for good tuning is significantly lower. For each recording, we take the variance of these deviations and use it as an input for isotonic regression to predict tuning quality level.

4.5 Evaluation

To evaluate our assessment approach we employ a fivefold cross-validation procedure to compare the predicted assessment values with the human assessments. In Table 2 we show the mean squared difference between automatically estimated performance levels and human annotations for the three criteria we have studied. The differences between human grades (individual and average) and automatic grades are less than the difference between human annotators themselves. Thus, automatically predicted levels lie in the same range than human assessments.

However, we have to properly interpret the results. It might be that the high degree of human subjectivity is caused by insufficient assessment instructions or criteria definition; perhaps the element of personal subjectivity should not be counted at all and a single consensus grade should be given to each factor for every recording. Also, since the final goal of the feedback is to develop the students metacognitive skills, such as the ability to self-learn and being self-critical, the effectiveness of the system should be evaluated by measuring the improvement of the students’ skills.

Table 2: Mean squared difference between numeric grades for Pitch, Rhythm and Tuning performance levels produced by our automatic system (auto) and two human annotators (hum. 1 and hum. 2).

Subjects	Pitch	Rhythm	Tuning
auto/hum.1	0.57	0.55	0.38
auto/hum.2	1.02	1.06	0.93
auto/hum. average	0.58	0.79	0.44
hum.1/hum.2	1.33	1.16	1.16

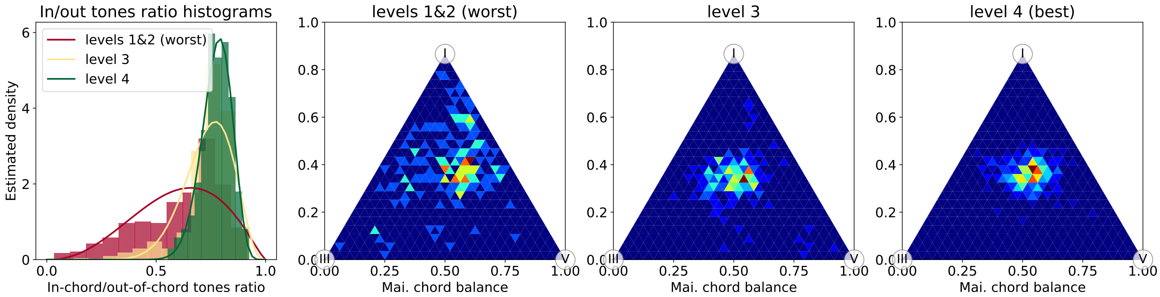


Figure 6: Distributions of chroma-derived features of major triad chords for different pitch accuracy levels. Leftmost: in/out tones ratio histograms and their β-distribution approximation. To the right: ternary plots for joint in-chord chroma distributions.

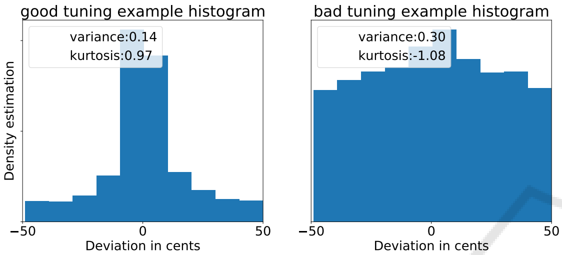


Figure 7: Histograms of inter spectral peaks intervals deviations from equal temperament grid for good tuning and bad tuning recordings.

5 OPEN CHALLENGES

Despite great advances, there are still many challenges to be addressed before we can claim to have assessment technologies with a broad educational impact.

A musical concept or process has to be understood before developing computational methods to study them. In section 2 we reviewed music performance assessment practices from which we based our work. However, the music learning process goes far beyond performance assessment alone. The results of assessment are an aid to determining the position of the student within a wider curriculum or learning plan meant to teach them competences needed for musicianship. Learning requires the development of meta-cognitive (learning to learn) skills, thus the planning, monitoring, and evaluation of learning. We need to understand how these meta-cognitive skills are learned and how we can help in their acquisition. Also is important to present the assessment results to a student in a way that promotes learning. Further research efforts should be made to conceptualize the different phases of music learning (Hallam and Bautista, 2012).

Most music education research has focused on western classical music, thus leaving out most music traditions and styles. We need to support our cultural diversity and thus promote the learning of the variety of musics that exists. In the project CompMusic[7] (Serra, 2011) we have been working on the analysis of several non-western music traditions and developing educational tools for them, but much work remains to be done.

We have presented advance signal processing and machine learning techniques capable of analyzing audio signals, but there is still many open problems when analyzing and characterizing the different dimensions and criteria of relevance in a musical performance. Especially difficult is to assess expressivity and, in general, to assess the more advanced levels of playing.

In this article we have focused on the analysis of audio signals, but the study of music can benefit from analysing other relevant and complementary signals, such as gesture movements or neurophysiological data. Some work has used the analysis of non-audio signals in the context of music education (Ramirez et al., 2018) but not much has been done on integrating different signal modalities into a unified analysis process of relevance for assessment.

Given the complexity of most musical concepts, to study them computationally we need to take advantage of the most advanced Artificial Intelligence approaches, specially machine learning methods. The availability of large and well annotated datasets, adequate for all the diverse assessment needs is the single most important current limitation. Also, given the characteristics of music learning, it is fundamental that the AI models developed be interpretable (Rosé et al., 2019), thus, that we can understand what they learn and that the analysis results help in understanding the complexities of a musical performance.

A part from developing adequate technologies to support music learning tasks, we need to create complete systems that support students in their learning. We need to include a user-centered perspective

[7] http://compmusic.upf.edu/

(Amershi et al., 2014). Each learner requires a different teaching method, thus a system should adapt to the needs of each individual student and it should include functionalities and technologies that go beyond what we have introduced in this article.

6 CONCLUSIONS

In this article we reviewed core topics related to music performance assessment while proposing specific technological solutions. We introduced specific technologies that we have developed to support the assessment of introductory guitar exercises and presented a prototype that has been deployed and tested in a real music education context. This prototype can be extended to support other musical instruments and can be adapted to a variety of educational contexts. The results show the usefulness and potential of our approach.

In order to further advance in this music education topic there is a need to involve different disciplines and expertises, which makes the task quite hard. Music Education is already a very multidisciplinary field, and if we add the technological component it is even more so. The collaboration between experts of all the different disciplines is the only way to address the challenges identified in this article.

ACKNOWLEDGEMENTS

This research was partly funded by the European Research Council under the European Union's Seventh Framework Program, as part of the TECSOME project (ERC grant agreement 768530).

REFERENCES

Abeßer, J., Lukashevich, H., and Schuller, G. (2010). Feature-based extraction of plucking and expression styles of the electric bass guitar. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2290–2293.

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120,

Bandiera, G., Romani Picas, O., Tokuda, H., Hariya, W., Oishi, K., and Serra, X. (2016). Good-sounds.org: A framework to explore goodness in instrumental sounds. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 414–419.

Böck, S., Krebs, F., and Schedl, M. (2012). Evaluating the online capabilities of onset detection methods. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 49–54.

Böck, S. and Widmer, G. (2013). Maximum filter vibrato suppression for onset detection. In *Proc. of the Int. Conf. on Digital Audio Effects*, pages 55–61.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 493–498.

Bozkurt, B., Baysal, O., and Yüret, D. (2017). A dataset and baseline system for singing voice assessment. In *Proc. of the Int. Symposium on Computer Music Multidisciplinary Research*, pages 25–28.

Bozkurt, B., Gulati, S., Romani, O., and Serra, X. (2018). MusicCritic : A technological framework to support online music teaching for large audiences. In *Proc. of the World Conf. of Int. Society for Music Education*, pages 13–20.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Association for Supervision & Curriculum Development, Washington, DC.

Brookhart, S. M. (2018). Appropriate Criteria: Key to Effective Rubrics. *Frontiers in Education*, 3(April),

Cho, T. and Bello, J. P. (2014). On the relative importance of individual components of chord recognition systems. *IEEE Transactions on Audio, Speech and Language Processing*, 22(2):477–492,

Conati, C., Porayska-Pomsta, K., and Mavrikis, M. (2018). AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling.

Duke, R. A., & Byo, J. L. (2012). Building musicianship in the instrumental classroom. In *The Oxford handbook of music education.*

Freire, S., Armondes, A., Viana, J., and Silva, R. (2018). Strumming on an acoustic nylon guitar: Microtiming, beat control and rhythmic expression in three different accompaniment patterns. In *Proc. of the Sound and Music Computing Conference*, pages 543–548.

Gordon, J. W. (1987). The perceptual attack time of musical tones. *Journal of the Acoustical Society of America*, 82(1):88–105,

Gouyon, F. and Dixon, S. (2005). A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54,

Gupta, M., Cotter, A., Pfeifer, J., Voevodski, K., Canini, K., Mangylov, A., Moczydlowski, W., and Van Esbroeck, A. (2016). Monotonic calibrated interpolated look-up tables. *Journal of Machine Learning Research*, 17:1–47.

Hallam, S. and Bautista, A. (2012). Processes of instrumental learning: The development of expertise. In *The Oxford handbook of music education*.

Han, Y. (2014). Hierarchical Approach to Detect Common Mistakes of Beginner Flute Players. In *Proc. of the Int. Society for Music Information Retrieval Conference*, number October 2014, pages 77–82.

Klapuri, A. (2006). Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. of the Int. Society for Music Information Retrievaltrieval Conference*, pages 216–221.

Krebs, F., Böck, S., and Widmer, G. (2015). An efficient state-space model for joint tempo and meter tracking. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 72–78.

Lerch, A., Arthur, C., Pati, A., and Gururani, S. (2019). Music Performance Analysis: A Survey. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 33–43.

Luo, Y. J., Su, L., Yang, Y. H., and Chi, T. S. (2015). Detection of common mistakes in novice violin playing. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 316–322.

Maestre, E. and Gómez, E. (2005). Automatic characterization of dynamics and articulation of expressive monophonic recordings. In *Proc. of the Audio Engineering Society Convention*, volume 1, pages 26–33.

Mantie, R. (2019). The Philosophy of Assessment in Music Education. *In The Oxford handbook of assessment policy and practice in music education, Volume 1* (pp. 32–55). Oxford University Press.

Mauch, M. and Dixon, S. (2014). PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal*, number 1, pages 659–663.

Mazur, Z. and Łaguna, M. (2017). Assessment of instrumental music performance: definitions, criteria, measurement. *Edukacja*, pages 115–128,

Molina, E., Barbancho, I., Gómez, E., Barbancho, A. M., and Tardón, L. J. (2013). Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 744–748.

Pati, K. A., Gururani, S., and Lerch, A. (2018). Assessment of student music performances using Deep Neural Networks. *Applied Sciences (Switzerland)*, 8(4),

Pauwels, J., O'hanlon, K., Gómez, E., and Sandler, M. B. (2019). 20 Years of Automatic Chord Recognition From Audio. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 54–63.

Pellegrino, K., Conway, C. M., and Russell, J. A. (2015). Assessment in Performance-Based Secondary Music Classes. *Music Educators Journal*, 102(1):48–55,

Ramirez, R., Volpe, G., Canepa, C., Ghisio, S., Kolykhalova, K., Giraldo, S., Mayor, O., Perez, A., Mancini, M., Volta, E., Waddell, G., and Williamon, A. (2018). Enhancing music learning with smart technologies. In *ACM Int. Conf. Proceeding Series*, number June, pages 1–4.

Reboursière, L., Lähdeoja, O., Drugman, T., Dupont, S., Picard, C., and Riche, N. (2012). Left and right-hand guitar playing techniques detection. In *Proc. of the Int. Conf. on New Interfaces for Musical Expression*, pages 1–4.

Romani Picas, O., Rodriguez, H. P., Dabiri, D., Tokuda, H., Hariya, W., Oishi, K., and Serra, X. (2015). A real-time system for measuring sound goodness in instrumental sounds. In *Proc. of the Audio Engineering Society Convention*, volume 2, pages 1106–1111.

Rosé, C. P., McLaughlin, E. A., Liu, R., and Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. *British Journal of Educational Technology*, 50(6):2943–2958,

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Russell, B. E. (2010). The Development of a Guitar Performance Rating Scale using a Facet-Factorial Approach. *Bulletin of the Council for Research in Music Education*, (184):21–34.

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,

Schedl, M., Gómez, E., and Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2-3):127–261.

Schneider, M. C., McDonel, J. S., and DePascale, C. A. (2019). *Performance Assessment and Rubric Design*, pages 628–650.

Schramm, R., de Souza Nunes, H., and Jung, C. R. (2015). Automatic solfège assessment. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 183–189.

Schreiber, H. and Müller, M. (2018). A single-step approach to musical tempo estimation using a convolutional neural network. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 98–105.

Serra, X. (2011). A Multicultural Approach in Music Information Research. In *Proc. of the Int. Society for Music Information Retrieval Conference*, pages 151–156.

Tomoyasu, N., Goto, M., and Hiraga, Y. (2006). An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features. In *Int. Conf. on Spoken Language Processing*, volume 8, pages 2–4.

van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Fundamental Concepts of Compositional Data Analysis*, pages 13–50. Springer Berlin Heidelberg,

Vidwans, A., Gururani, S., Wu, C. W., Subramanian, V., Swaminathan, R. V., and Lerch, A. (2017). Objective descriptors for the assessment of student music performances. In *Proc. of the AES Int. Conference*, pages 116–123.

Wesolowski, B. C. (2012). Understanding and Developing Rubrics for Music Performance Assessment. *Music Educators Journal*, 98(3):36–42,

Wesolowski, B. C. (2014). Documenting Student Learning in Music Performance. *Music Educators Journal*, 101(1):77–85,

Wesolowski, B. C. and Wind, S. A. (2019). *Validity, Reliability, and Fairness in Music Testing*. In *The Oxford handbook of assessment policy and practice in music education* (pp. 437–460).

Widmer, G. (1998). Applications of Machine Learning to Music Research: Empirical Investigations into the Phenomenon of Musical Expression. *Machine Learning, Data Mining, and Knowledge Discovery: Methods and Applications*, (March):269–293.

Wu, C.-W. and Lerch, A. (2018). Learned features for the assessment of percussive music performances. In *Proc. of the Int. Conf. on Semantic Computing*, pages 93–99.