A DATASET OF RHYTHMIC PATTERN REPRODUCTIONS AND BASELINE AUTOMATIC ASSESSMENT SYSTEM

Felipe Falcão[1,2] **Baris Bozkurt**[2,3] **Xavier Serra**[2]
Nazareno Andrade[1] **Ozan Baysal**[4]
[1] Universidade Federal de Campina Grande, Brazil
[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona
[3] Izmir Demokrasi University, Turkey
[4] Istanbul Technical University, Turkey

ABSTRACT

This work presents a novel dataset comprised of audio and jury evaluations for rhythmic pattern reproduction performances by students applying for a conservatory. Data was collected *in-loco* during entrance exams where students were asked to imitate a set of rhythmic patterns played by teachers. In addition to the *pass* or *fail* grades provided by the members of the jury during the exam sessions, a subset of the data was also evaluated by external annotators on a 4-level scale. A baseline automatic assessment system is presented to demonstrate the usefulness of the dataset. Preliminary results deliver an accuracy of 76% for a simple pass/fail logistic regression classifier and a mean average error of 0.59 for a linear regression grade estimator. The implementation is also made publicly available to serve as baseline for alternative assessments systems that may leverage the dataset.

1. INTRODUCTION

Automatic assessment of music performances is an important audio signal processing application drawing increased attention over the past few years. The *Massive Open Online Course* (MOOC) methodology has recently contributed to the growth of online music courses, attracting a large number of students. In this scenario, automatic assessment methodologies have the potential to largely reduce the instructor load of assessing student submissions. Moreover, due to its subjective nature, the task of evaluating students performances can be very difficult [4, 14, 24], sometimes even preventing different evaluators from reaching an agreement while assessing the same performance [17, 22]. Automatic evaluations may circumvent this obstacle by defining clear and objective goals that must be achieved in order to succeed in a musical performance.

While some authors leverage their evaluation tools using linear measures to quantify similarity between pairs of reference and performance [20], most of the recently proposed assessment systems rely on machine learning based models for this task. Earlier models were trained with labeled data and hand-crafted audio features targeting at the prediction of grades for performances. Such methodology is applied in Nakano et al. [13] for improving the state-of-art for singing voice assessment. The authors gathered data from the AIST-HDB dataset [9] to train a model able to predict *good/poor* classifications for singing performances. Bozkurt et al. [4] have also suggested a supervised learning method for singing voice assessment, but this time conducting their own data collection procedures inside a music conservatory. The collected data containing jury evaluations was fed into a machine learning model whose accuracy was reported as 74% for binary *pass/fail* predictions. Singing assessments following similar methodologies and including classification systems were previously proposed by Schram et al. [16] and Molina et al. [12]. Both authors aimed at the automatic evaluation of voice performances by training machine learning models with audio features and targeted scores.

Recent advances in unsupervised learning led researchers to also rely on learned metrics to support their assessment systems. Unlike the aforementioned supervised procedures, these techniques delegate to the model itself the task of figuring data patterns directly from raw audio data, clustering similar observations into equivalent groups and using such information to predict assessments. Examples of such methodology include the results discussed by Wu and Lerch in [24], where authors modeled a feature learning approach specially designed for assessing percussive performances recorded during band auditions. This very same data source leveraged Pati et al. [14] in their similar study also tackling the problem of modeling music assessment by means of learned features. They have proposed the application of deep neural networks capable of capturing non-linear aspects of performances that would better correlate with reference's features. These recent studies encourage the use of unsupervised feature learning rather than supervised methods, reporting that the former

outperforms the latter in most of the cases.

Audio corpora play a critical role in every assessment system, as the decision functions used for predicting evaluations are trained on the corpora. In their work, Li et al [10] delivered an important contribution by reviewing several commonly-used music datasets made publicly available for MIR tasks. This systematic review also describes important information regarding the nature of each dataset (e.g. available content, total audio durations, types of annotations) and points out to a lack of datasets with annotations related to rhythmic assessment, we argue. Specifically regarding the rhythmic dimension, the dataset provided by the Florida Bandmasters Association (FBA) has been commonly adopted [14,21,23,24] and is, to the extent of our knowledge, the only dataset that currently includes rhythmic performances that are annotated with grades. It is comprised of audio performances from band auditions recorded in the Florida state between 2013 and 2015, including jury assessments for different music aspects (musicality, note accuracy, rhythmic accuracy, tone quality, etc.).

This present work is an effort to address the shortage of music datasets designed for rhythmic assessment. The presented data was collected during the rhythmic sessions of entrance exams based in a music conservatory, where student performances were recorded and evaluated by members of a jury. Data curation procedures were applied over this raw dataset (including 1040 student performance recordings) in order to extract a subset of (80) performances which were submitted to an extra evaluation, this time in a 4-level scale (i.e. grades ranging from 1 to 4). The resulting subset featured with annotated data was fed to a simple machine learning rhythmic assessment system in order to demonstrate the use of the dataset in this scenario. A binary *pass/fail* logistic classifier and a linear regression grade estimator are implemented, the former delivering a maximum accuracy of 76% while the latter pointing to a minimum mean average error of 0.59 when tested over a 5-fold cross validation. Both the complete rhythmic dataset and the re-annotated subset are made publicly available. The implementation of the proposed rhythmic assessment system is also openly shared to serve as baseline for comparisons with similar approaches.

2. THE MAST RHYTMIC DATASET

The Musical Aptitude Standard Test (MAST) Rhythmic Dataset is a collection of rhythmic performances and references recorded in the Istanbul Technical University (ITU) Turkish Music Conservatory during entrance exams. In Turkey, these assessments are commonly applied to support the evaluation of the musical aptitude of applicants, determining whether or not they should be accepted to study in the institution. Categorizations are preferably achieved during jury-assisted exams when students are individually auditioned and evaluated according to multiple musical aspects (e.g. chord recognition, melodic singing, rhythm playing). For the rhythmic session students are asked to imitate reference performances by usually clapping hands or tapping a hard surface.

The rhythmic patterns included in the dataset are taken from the jury based qualification exams of the years 2015 and 2016. The rhythmic assessment portion of the entrance exams in these years was composed of two types of rhythmic pattern reproduction questions; rhythm one in a simple meter (4/4) and rhythm two in a compound meter (7/8, 9/8, 10/8 or a 5/4). In order to ensure the confidentiality of the questions asked in the exam and minimize the chance of a memorization and the leakage of these pattern outside of the examination areas, the applicants chose randomly from 10 different question packages, each package having a different version of the two types of rhythmic patterns stated. Thus, there are 20 different rhythms for each year making up the total of 40 distinct rhythmic patterns in our dataset.

Besides taking into consideration the two types of rhythms, the exam preparation committee designed the questions such that each pattern should be employing similar rhythmic values (quarter-note, eight-note, sixteenth-note and a triplet) and similar number of notes. The applicants were expected to perform above a threshold of success regardless of the package the selected. Later, it was observed that, while there may be differences in terms of difficulty for different packages which might affect an applicant's test score, there wasn't any significant relationship observed between the selected packages and the success of the candidates [1].

The jury committee consists of three members and the candidates are expected to reproduce the rhythmic pattern after it has been played two times by a member of the committee. The jury gives a full grade (10pts) and moves on to the next question if the candidate's performance is an exact reproduction (or nearly). If there are flaws in the reproduction, then the candidate is exercised by performing the rhythmic pattern divided to two halves separately, after which the rhythmic pattern is played for reproduction one last time. The evaluation at this stage may have three outcomes, either the participant receives a partial grade (8pts) if it is an exact reproduction (since he/she couldn't perform an exact reproduction at the beginning), if it still has 1-2 errors the jury gives a minor grade (4 pts), and no points are given if the performance has more than 2 errors. The evaluations of the jury member can show variances at this stage due to individual preferences (e.g. for some a consistent tactus may be more important than missing an attack, for some it is the accentuation and the phrasing). Due to these, in our dataset we have only selected those performances in which there was an unanimously consensus among jury members that it was an exact (or nearly exact) reproduction or a failure (all giving 0 pts.).

In general, to allow reinspection of the execution of these exams, each applicants performance is video recorded by the ITU conservatory directorate. For our purposes – and to ensure anonymity - these recordings were converted to wave files and then cropped so that each recording consisted of the candidates performances only.

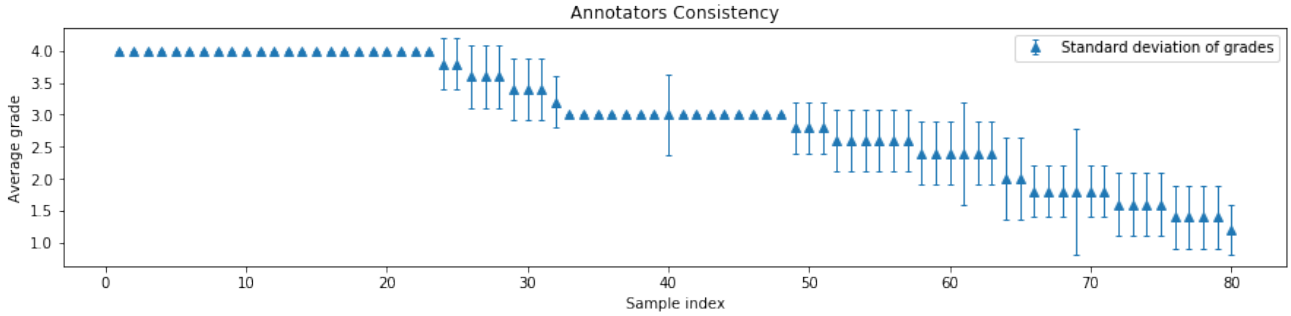

Figure 1. Distribution of grades assigned by the distinct annotators. The x-values indicate the performance index while the y-values stand for the average grade assigned. Error bars describe the standard deviation for all evaluations

3. USER ANNOTATIONS

In order to contribute to a more complete validation over the collected data, the original dataset - which originally only contained *pass/fail* classifications provided by members of a jury - was also annotated with a higher resolution (4-level) grading. Since the re-assessment of all the 1040 student performances comprised by the full dataset would require high human resources, the original data was sampled. This convenience sampling initially filtered 20 references with low rhythmic complexity, followed by the selection of four noise-free student performances for each reference - two from each *pass/fail* class, totaling 80 sampled student performances and 20 jury member performances (that serve as the reference/target rhythmic pattern for grading a student performance via comparison).

This subset of performances (from now on addressed as re-annotated subset) was submitted to evaluation sessions completed by seven annotators (male: six, female: one) and aided by a custom evaluation tool (depicted in Figure 2) developed for a similar data collection task. During these sessions, the annotators could hear in sequence the rhythmic reference and student performance (with a one second silence in-between) as many time as desired until feeling comfortable to choose between one of the available grades: *1 - Completely off, 2 - Major errors, 3 - Minor errors, 4 - Perfect*. Although no advanced music skills were considered mandatory to support the assessment of quite simple rhythmic patterns, authors tried to select annotators with some relevant music background. Besides, the graders were provided with a custom rubric documenting the musical aspects that should be taken into consideration when assigning grades. They were asked to assess the similarity of the student performance to the reference in terms of the beat and duration patterns, discarding tempo differences.

Evaluation sessions could be interrupted and resumed by annotators at any moment using the session control feature provided by the tool. Figure 1 presents the distribution of averaged grades assigned by annotators to all the sampled performances. Nearly half (38) of evaluated performances had unanimous assessments while the rest of them presented some deviations (mean: 0.49, max: 0.97).

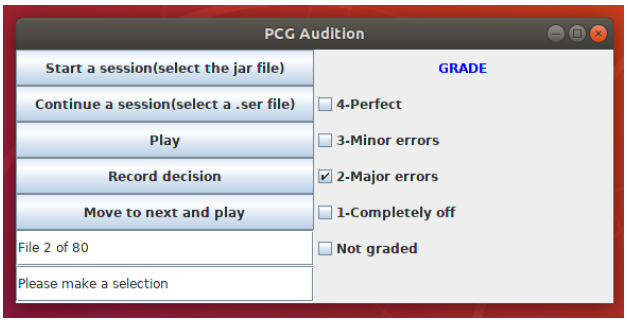

Figure 2. Annotation tool used during the custom evaluation sessions

4. DATA PREPARATION

The detection of rhythmic events is an issue recurrently addressed via extraction of onset times from raw audio [6–8]. Onset features are primarily encoded as onset vectors containing the moments when signal-disturbing events (e.g. chord attack, drum kick) happen. Multiple similarity measures have already been proposed for the comparison of onset vectors, including distance between vectors [18] and error measures [15]. In this present work we propose a hybrid model that benefits from both aforementioned similarity measures to predict rhythmic assessments.

For the onset detection, our audio processing module relies on the *OnsetDetection* algorithm implemented by the Essentia library [3] [1]. All references and performances comprised by the re-annotated subset were sampled at a rate of 44.1kHz and the resulting frames were provided to the onset detection algorithm to allow for feature extraction. The onset extractor is parameterized with default values (window size: 1024 samples, hop size: 512 samples) and three methods for onset extraction are tested and compared: High Frequency Content detection (HFC) [11], Spectral Flux detection (FLUX) [19] and Complex-Domain spectral difference (COMPLEX) [2].

The recordings in the dataset are not aligned in time, nor cropped with a fixed offset before/after the first/last onset. Hence the first and the last onset are considered as boundaries of each performance. For the use of vector distance measures applied to same-sized vectors, binary vectors are

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

computed applying a fixed-numbered (60) grid on the time axis (i.e. for each recording, the duration between the first and last onset is divided into 60 time bins and a binary value (onset/non-onset) is stored in the vector for each bin). We opted for 60 bins for each rhythmic performance, since this number is divisible by two, three, four, five, and six, which are the common multipliers for most rhythmic patterns.

Figure 3 presents an example of a visual representation for onsets times before and after the quantizing procedures that we have just described. The relative positions for the original onsets are plotted in dashed lines and circle stems, while the quantized information is drawn in solid lines and triangle stems. All the quantized, unquantized onsets and scaling information are also included into the re-annotated dataset for further use.

5. BASELINE ASSESSMENT SYSTEM

This work introduces a baseline automatic assessment system for rhythmic performances of students. Regression models are trained with vector similarities aiming at the modeling of the rhythmic evaluations detailed in the re-annotated subset.

The feature set for both models described below includes eight different similarity measures: two strictly related to the rhythmic domain (beat difference and Percival's similarity [15]), four text-distance approaches (Levenshtein, Damerou-Levenshtein, Jaro and Jaro-Winkler) and two vector-distance measures (Hamming and Yule). This last subset of features was selected according to their reported benefits in comparing vector with boolean data [5]. Since the nature of these measures resulted in different distance ranges, all features were normalized in order to range within a common scale.

Two different types of evaluations are targeted by the proposed assessment systems. A grade estimation is implemented through a linear regression model while categorizations between *pass/fail* classes are predicted by a logistic regression classifier. The overall 'true' grade of a performance is calculated via removing the highest and lowest grades and averaging the rest, which is an approach similar to the one usually applied in music conservatories. Binary categorization is modeled by judging as accepted (*pass*) all performances whose average grades are greater than or equal to three, rejected (*fail*) otherwise.

All machine learning implementations are written in Python 3.6.7 using scikit-learn [2] modules.

6. RESULTS

Both mentioned datasets are now made publicly available for further investigation. The complete rhytmic dataset (MAST rhythm dataset) [3] is a collection of 3721 audio files cropped from recordings of conservatory entrance examinations in Turkey (summer 2015 and summer 2016). 1040 of the recordings are student performances graded by a jury of 3 instructors as pass or fail. The rest of the recordings are jury performances of the same rhythmic patterns in various sessions. The re-annotated subset (MAST rhythm re-annotated subset) [4] is a balanced sample (in terms of pass-fail graded samples) extracted from the complete dataset, assessed by seven annotators in a 4-level grid and onset information for 80 performances, accounting for 20 distinct rhythmic patterns (references). All the code supporting the implementation of our automatic assessment systems is also shared as Jupyter notebooks at Github [5].

Classifier	Ac.	Pr.	Rc.
Logistic Regression (FLUX)	63%	66%	83%
Logistic Regression (COMPLEX)	72%	75%	83%
Logistic Regression (HFC)	**76%**	**79%**	**78%**

Table 1. Performance measures (accuracy, precision and recall) for classifiers trained with different onset features

The designed models are evaluated according to how well they predict assessment for unseen test data. Our evaluation relied on a 5-fold cross validation (test size: 20%) for both models. The performance results for the logistic classifier are summarized in the learning curves shown in Figure 4 and states a maximum accuracy of 76% when trained with HFC data. The complete comparison stating accuracy, precision and recall results from various onset features can also be examined in Table 1.

For the linear grade estimator, the performance analysis consisted of measuring the errors observed between the predicted grades and the expected values. Our estimator was compared with baseline naive versions whose projected evaluations are modeled using uniform and random distributions. Results are summarized in Table 3 and indicate that our approach provides better predictions (yet with a small margin) than the naive estimators regardless of the trained feature, with HFC once again delivering the best performance. As for the influence of specific features over the decision function, the coefficients suggested by the linear regression (Table 2) encourage us to infer that Jaro, Jaro-Winkler and Hamming distances are the features whose influences are higher over grading predictions.

Our final evaluation is carried out by crossing data from the two proposed datasets. Models trained with sampled data from the re-annotated dataset were provided with all performances coming from the complete dataset in order to verify how well the trained assessment systems would behave when asked to predict evaluations for new unseen data. All the 1040 performances comprised by the full dataset were submitted to the same pre-processing steps described in Section 4 and the resulting similarity measures were tested against both the logistic classifier and linear estimator trained with HFC features (since it's the extraction method that delivers the best performance). Here accuracy is calculated according to the number of predic-

[2] https://scikit-learn.org

[3] https://zenodo.org/record/2620357

[4] https://zenodo.org/record/2619499

[5] https://github.com/MTG/mast-rhythm-analysis

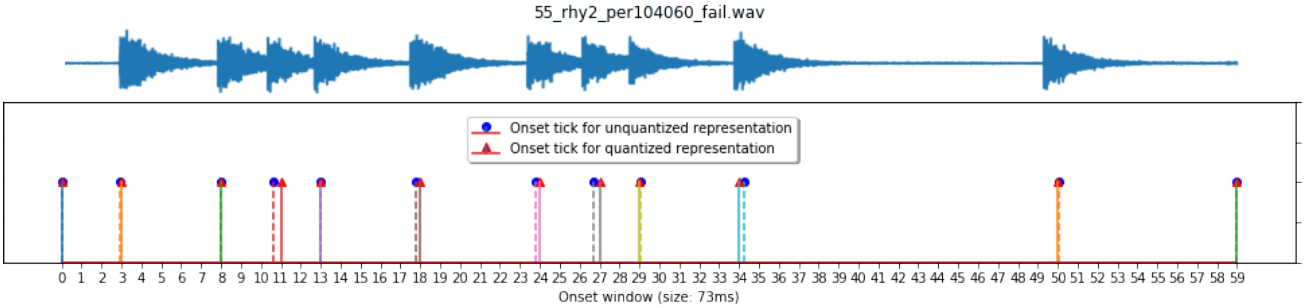

Figure 3. Example of waveform and onset information (detection method: HFC) before and after quantizing procedures. The stems in both ends relate with the fact that audios were prior cropped to range from first to last onsets

Feature	Coefficient	Intercept
Beat Difference	-44.31	
Rhythmic Similarity	72.71	
Levenshtein	-8.94	
Damerau-Levenshtein	11.17	-54.85
Jaro	**-8229.71**	
Jaro-Winkler	**6894.59**	
Hamming	**-988.81**	
Yule Similarity	188.32	

Table 2. Feature coefficients and intercept for the linear regression model trained with HFC onsets. Features with higher influence over predictions are highlighted

Estimator	MAE	MSE	R^2
Fixed grading to 2	1.12	1.71	-1.52
Fixed grading to 3	0.71	0.76	-0.04
Random grading	0.99	1.52	-1.53
Linear Regression (FLUX)	0.69	0.69	0.01
Linear Regression (COMPLEX)	0.59	0.57	0.23
Linear Regression (HFC)	**0.59**	**0.50**	**0.21**

Table 3. Comparison between error measures (Mean Average Error, Mean Squared Error and R-squared) observed in predictions for naive estimators and proposed model trained with different features

tions matching the jury evaluations (*pass/fail*). Final results report a matching rate of 65% for the binary predictions while the grade estimator guessed the right class for 70% of the unseen data.

7. CONCLUSIONS

The present study is an attempt to address the lack of data sources designed for automatic rhythmic assessment. Student performances for a set of rhythmic patterns were recorded and evaluated by a jury during entrance exams conducted in a music conservatory. An additional data annotation step was also carried out with seven annotators, this time grading a subset of these performances with grades ranging from one to four. The re-annotated subset trained an automatic assessment system able to predict

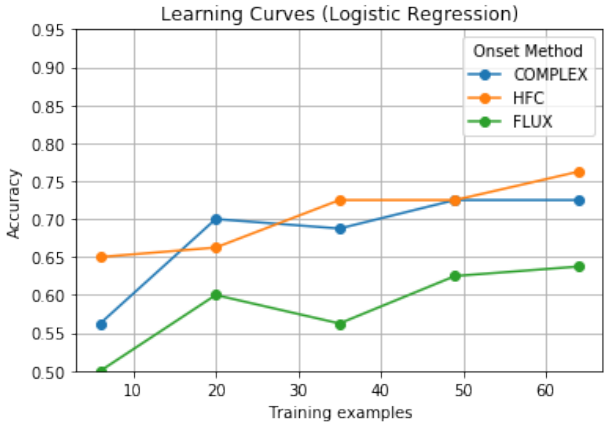

Figure 4. Learning curve for the *pass/fail* logistic classifier

students evaluations for rhythmic tasks. Models delivered a maximum accuracy of 76% for a binary (*pass/fail*) classifier and presented a minimum mean average error of 0.59 when predicting grades in a linear fashion, also pointing to Jaro, Jaro-Winkler and Hamming distances as the best model predictors. When compared with the data remained from the re-annotation process, the baseline assessment system predictions matched the jury-labeled data in about 70% of the cases.

All the aforementioned artifacts are now made public for further investigation. Both the full dataset and the re-annotated subset can be freely accessed and used to support assessment systems that builds on more sophisticated techniques to predict student grades for rhythmic lessons. Besides, the proposed implementation is also made available as Jupyter notebooks that can be examined and used as baseline in comparative studies.

8. ACKNOWLEDGEMENTS

This work was conducted during a visiting scholar period at Universitat Pompeu Fabra, sponsored by the Capes Foundation within the Ministry of Education, Brazil (grant n. 88881.189929/2018-01). The dataset was collected during a project funded by the Scientific and Technological Research Council of Turkey, TUBITAK, Grant [215K017].

9. REFERENCES

[1] Ozan Baysal, Baris Bozkurt, Turan Sager, and Nilgun Dogrusoz. Towards a hybrid assessment model for music conservatory entrance exams. In *Proceedings of Education Studies '18 - II.International Conference on Education and Learning Conference, Istanbul*, pages 81–96. DAKAM, 2018.

[2] Juan Pablo Bello, Chris Duxbury, Mike Davies, and Mark Sandler. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters*, 11(6):553–556, 2004.

[3] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Perfecto Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, Justin Salamon, José Ricardo Zapata González, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR), 2013.

[4] Baris Bozkurt, Ozan Baysal, and D Yuret. A dataset and baseline system for singing voice assessment. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR), Matosinhos, Portugal*, pages 25–28, 2017.

[5] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

[6] Norberto Degara, Antonio Pena, Matthew EP Davies, and Mark D Plumbley. Note onset detection using rhythmic structure. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5526–5529. IEEE, 2010.

[7] Simon Dixon, Fabien Gouyon, Gerhard Widmer, et al. Towards characterisation of music via rhythmic patterns. In *ISMIR*. Citeseer, 2004.

[8] Jonathan Foote and Shingo Uchihashi. The beat spectrum: A new approach to rhythm analysis. page 224. IEEE, 2001.

[9] M Goto and T Nishimura. Aist humming database: Music database for singing research. *IPSJ SIG Notes (Technical Report)(Japanese edition)*, 2005(82):7–12, 2005.

[10] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535, 2019.

[11] Paul Masri and Andrew Bateman. Improved modelling of attack transients in music analysis-resynthesis. In *ICMC*, 1996.

[12] Emilio Molina, Isabel Barbancho, Emilia Gómez, Ana Maria Barbancho, and Lorenzo J Tardón. Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 744–748. IEEE, 2013.

[13] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Ninth International Conference on Spoken Language Processing*, 2006.

[14] Kumar Pati, Siddharth Gururani, and Alexander Lerch. Assessment of student music performances using deep neural networks. *Applied Sciences*, 8(4):507, 2018.

[15] Graham Keith Percival. *Computer-assisted musical instrument tutoring with targeted exercises*. PhD thesis, 2008.

[16] Rodrigo Schramm, Helena de Souza Nunes, and Cláudio Rosito Jung. Automatic solfège assessment. In *ISMIR*, pages 183–189, 2015.

[17] Sam Thompson and Aaron Williamon. Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception: An Interdisciplinary Journal*, 21(1):21–41, 2003.

[18] Godfried T Toussaint. A comparison of rhythmic similarity measures. In *ISMIR*, 2004.

[19] George Tzanetakis and Perry Cook. Multifeature audio segmentation for browsing and annotation. In *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No. 99TH8452)*, pages 103–106. IEEE, 1999.

[20] Burak Uyar and Baris Bozkurt. An interactive rhythm training tool for usuls of turkish makam music. In *5th Int. Workshop on Folk Music Analysis (FMA)*, Paris, France, 2015.

[21] Amruta Vidwans, Siddharth Gururani, Chih-Wei Wu, Vinod Subramanian, Rupak Vignesh Swaminathan, and Alexander Lerch. Objective descriptors for the assessment of student music performances. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[22] Brian C Wesolowski, Stefanie A Wind, and George Engelhard. Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted rasch partial credit model. *Music Perception: An Interdisciplinary Journal*, 33(5):662–678, 2016.

[23] Chih-Wei Wu, Siddharth Gururani, Christopher Laguna, Ashis Pati, Amruta Vidwans, and Alexander Lerch. Towards the objective assessment of music performances. In *Proceedings of International Conference on Music Perception and Cognition (ICMPC)*, pages 99–102, 2016.

[24] Chih-Wei Wu and Alexander Lerch. Learned features for the assessment of percussive music performances. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 93–99. IEEE, 2018.