Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

Objective descriptors for the assessment of student music performances

Amruta Vidwans[1], Siddharth Gururani[1], Chih-Wei Wu[1], Vinod Subramanian[1], Rupak Vignesh Swaminathan[1], and Alexander Lerch[1]

[1] *Center for Music Technology, Georgia Institute of Technology*

Correspondence should be addressed to Alexander Lerch (alexander.lerch@gatech.edu)

ABSTRACT

Assessment of students' music performances is a subjective task that requires the judgment of technical correctness as well as aesthetic properties. A computational model automatically evaluating music performance based on objective measurements could ensure consistent and reproducible assessments for, e.g., automatic music tutoring systems. In this study, we investigate the effectiveness of various audio descriptors for assessing performances. Specifically, three different sets of features, including a baseline set, score-independent features, and score-based features, are compared with respect to their efficiency in regression tasks. The results show that human assessments can be modeled to a certain degree, however, the generality of the model still needs further investigation.

1 Introduction

The qualitative assessment of music performance is an essential pedagogical component when learning a musical instrument. It requires the observation, quantification, and judgment of characteristics and properties of a music performance. This is inherently subjective — the teacher's assessment might be impacted by many contextual and even non-musical considerations. Wesolowski et al. point out that raters may vary significantly in terms of their severity, rating scale, and interpretation of rating categories [1]. In addition, the bias of the human raters and closely related rating categories could, according to Thompson and Williamon, adversely affect the discriminability and fairness of the assessment [2]. As a result, the objectivity and reproducibility of human assessment can be questioned. However, an overall assessment is still often desired or required, e.g., for rating a student in an audition. A computational approach to quantitatively assessing student music performance could provide objective, consistent, and repeatable feedback to the student. It can also enable qualitative feedback to the student in situations without a teacher such as in practice sessions.

The realization of automatic systems for music performance assessment generally requires knowledge from multiple disciplines such as digital signal processing, musicology, and music psychology. With recent advances in Music Information Retrieval (MIR) [3], which involves the study of the above mentioned fields, noticeable progress has been made in related research topics such as source separation [4] and music transcription [5]. Examples of MIR-approaches applied to music education have been summarized by Dittmar et al. [6]. In addition to academic research, commercial

systems such as Smart Music[1] and Yousician[2] are available. Despite these efforts, identifying a reliable and effective method for assessing music performances remains an unsolved topic and requires further research.

In this paper, we explore the effectiveness of various objective descriptors by comparing three sets of features extracted from the audio recording of a music performance, a baseline set with common low-level features, a score-independent set with designed performance features, and a score-based set with designed performance features. The goal is to identify a set of meaningful objective descriptors for the general assessment of student music performances.

This paper is structured as follows: in Sect. 2, the related work on objective music performance assessment is introduced. The methodology is mentioned in Sect. 3, and the dataset used in this work is described in Sect. 4. Sect. 5 includes the experiment setup and results. Finally, the discussion and conclusion are presented in Sects. 6 and 7, respectively.

2 Related work

Music performance analysis deals with the observation, extraction, description, interpretation, and modeling of music performances [7]. Even before the age of the computer, Seashore points out the value of scientific observation of performances for music education [8]. Automatic performance analysis was introduced to the classroom as early as 1971 when products like the IBM-1500 instructional system spearheaded computer-assisted (music) education [9].

Performance analysis may or may not use the musical score in addition to the audio input. Approaches that do not require the score make sense in a setting where the score is not available, including improvisation or free practice. It can also be argued that humans can, at least to a certain degree, assess the proficiency of a music student without prior knowledge of the piece being played; a machine learning model should theoretically be able to do the same. Nakano presented an automatic system to evaluate user's singing skills without any score input [10], in which a singing performance is classified as good or poor using features such as pitch accuracy and vibrato length. Romani et al. developed a software tool that assesses the sound quality of a performer in real-time by analyzing the audio, note by note, in order to assess the stability and tonal richness of each individual note and reports an overall goodness score [11]. Isabel et al. present a score-independent algorithm to identify the technique that a violin performer is using such as pizzicato and vibrato using pitch and envelope features [12]. Musical expressions of four types (happy, sad, angry, and calm) were classified by Mion and De Poli [13]. They extracted instantaneous and event-based features such as spectral centroid, residual energy, and notes per second from violin, flute, and guitar performances. They argue that a known mapping of physical properties of sound to expressive properties of a performance can support effective querying in music retrieval systems. Han and Lee proposed an instrument specific approach to identify common mistakes of beginner flute players. The system was designed to detect incorrect assembly of the flute, poor blowing, and mis-fingering [14]. More recently, Wu et al. have proposed the automatic assessment of students' instrumental performances using score-independent audio features based on pitch, amplitude and rhythm histograms [15]. The results of a trained regression model showed reasonable correlation between model output and subjective assessments by human judges.

While the above approaches emphasize the use of score-independent features, it is common for beginner or intermediate students to practice on a well-known musical piece with readily available score. Therefore, many approaches take advantage of this additional score information. Abeßer et al. proposed a system that automatically assesses the quality of vocal and instrumental performances of 9th and 10th graders [16]. Score-based features like pitch, intonation and rhythmic correctness were designed to model the experts' ratings with a four-class classifier (rating scale: 1–4). They report the system to be able to classify the performances mostly correct with some confusion between adjacent ratings. A score-informed piano tutoring system has been presented by Fukuda et al. [17]. It applies automatic music transcription and audio-to-score alignment to detect mistakes in the performance. Schramm et al. use pitch deviations, onset and offset time deviation information annotated from student performances to create a model to classify correct or incorrect notes using a Bayesian classifier. Devaney et al. have created a performance analysis toolkit for ensemble singing by aligning the audio to the midi score and extracting pitch, timing and

[1] www.smartmusic.com Last Access: 2017/01/23
[2] www.yousician.com Last Access: 2017/01/23

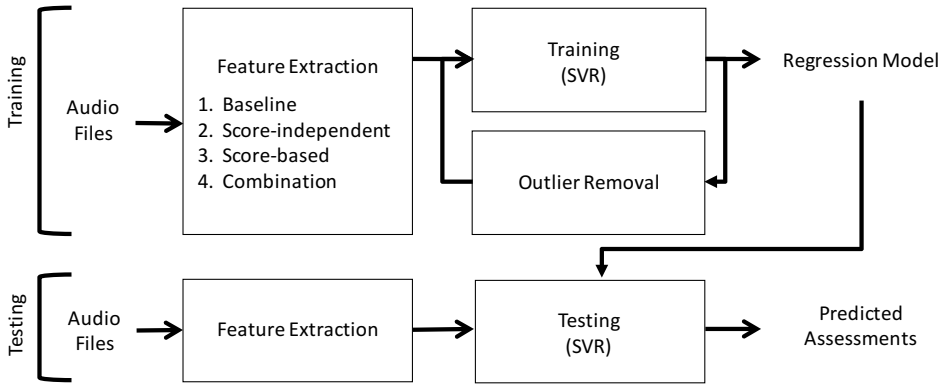

Fig. 1: Block diagram of the experimental setup

dynamics features [18]. The algorithm uses a Hidden Markov Model (HMM) model, trained to detect silence, transient and steady state, in addition to Dynamic Time Warping (DTW) to align the score to the pitch contour of the performance. This study reports a trend of the intonation change by the singers in 4 ensembles which can be further used to provide overall assessment of how well one ensemble performed with respect to the other. Mayor et al. have proposed a system for assessing a singer and providing feedback not only via a final evaluation of the performance but also through real-time feedback about expressivity, tuning and timing [19]. Their system makes use of a reference MIDI track which they align with the user's pitch contour. For expression, they define a set of audio features that uniquely identify each expression; an HMM is used to segment the performance into different expression regions. Tsai and Lee proposed a method for karaoke singing evaluation which provides ratings for users' singing performances on pitch, rhythm and loudness [20]. For pitch ratings, the DTW distance is computed between the pitch contour of user performance and reference audio after removing the background accompaniment using spectral subtraction. For rhythm ratings, the synchronicity between the singing and the accompaniment is measured. For volume ratings, the DTW distance between the short-term log-energy sequence of both audio is used.

3 Method

A block diagram of the method is shown in Fig. 1.[3] A pre-processing step involves downmixing and normalization of the audio signal.

3.1 Feature extraction

The recording will be represented by three sets of features: (i) baseline: a set of low-level features commonly used in MIR tasks [21, 7], (ii) score-independent: a set of designed features working with the audio signal without knowledge of the musical score, and (iii) score-based: a set of designed features extracted after aligning the audio with the musical score. The pitch contour of the recordings, required for the designed features, is extracted using a simple autocorrelation-based pitch-tracking method.

3.1.1 Baseline features

The baseline feature set consists of 13 Mel Frequency Cepstral Coefficients (MFCCs), zero-crossing rate, spectral centroid, spectral rolloff, and spectral flux. The implementation of these common features follow the definitions in [7] (see also the online repository[4]). To represent each recording with one feature vector, a two-stage feature aggregation process is applied. In the first stage, the block-wise features are aggregated and represented by their mean and standard deviation within a 250 ms texture window. In the second stage, these texture window level features are aggregated over the entire audio file and represented by their mean and standard deviation. This results in a single feature vector with a dimensionality of $d_{\mathrm{B}} = 68$ per recording.

[3] The corresponding source code is available online at `www.github.com/GTCMT/FBA2013`

[4] `www.github.com/alexanderlerch/ACA-Code`

3.1.2 Score-independent features

The score-independent feature set is designed to represent the performance accuracy with respect to pitch, dynamics, and rhythm. If not otherwise mentioned, the features are extracted at the note-level and then aggregated across all the notes. In order to compute note-level features, the pitch contour is segmented into notes by using the edges between the adjacent notes as the onsets.

Pitch The pitch features are extracted from the pitch contour. The features are:

- *note steadiness* ($d_{\mathrm{p1}} = 2$): For each note, the standard deviation of pitch values and the percentage of pitch values deviating from the mean by more than one standard deviation are computed. These two features are designed to represent fluctuations in the pitch of a note.
- *average pitch accuracy* ($d_{\mathrm{p2}} = 1$): The histogram of the pitch deviation from the closest equally tempered pitch is extracted with a 10 cent resolution. The feature is the area around the bin with highest count (width: 30 cent) of this histogram. This feature characterizes the pitch deviation of the notes played.
- *percentage of in-tune notes* ($d_{\mathrm{p3}} = 1$): Each note is labeled either in-tune or detuned, and the percentage of correct notes across the entire exercise is computed as the feature. A note is labeled correct if the percentage of pitch values with a deviation from the mean pitch is lower than a pre-defined threshold.

Dynamics Similar to the pitch features, these features use the note segmentation in order to compute per note features that can then be aggregated.

- *amplitude deviation* ($d_{\mathrm{a1}} = 1$): This feature aims to find the uniformity of the Root Mean Square (RMS) per note. For each note, the standard deviation of the RMS is computed.
- *amplitude envelope spikes* ($d_{\mathrm{a2}} = 1$): This feature describes the spikiness of the note amplitude over time. The number of local maxima of the smoothed derivative of the RMS is computed per note.

Rhythm The rhythm features are computed from the Inter-Onset-Interval (IOI) histogram (with 50 bins) of the note onsets.

- *timing accuracy* ($d_{\mathrm{r}} = 6$): The standard statistical measures of crest, skewness, kurtosis, rolloff, tonal power ratio, and the histogram resolution are extracted from the histogram.

For all note level features, the mean, maximum, minimum, and standard deviation is computed across all notes to represent the recording. This results in an overall number of features of

$$d_{\mathrm{SI}} = 4 \cdot d_{\mathrm{p1}} + d_{\mathrm{p2}} + d_{\mathrm{p3}} + 4 \cdot d_{\mathrm{a1}} + 4 \cdot d_{\mathrm{a2}} + d_{\mathrm{r}} = 24$$

3.1.3 Score-based features

The set of score-based features is extracted utilizing score information by aligning the extracted pitch contour to the sequence of pitches from the score with DTW. Before aligning the pitch contour, the tuning frequency is estimated using the mode of the pitch histogram. The pitch contour is subsequently shifted by the tuning frequency estimate. The output of the DTW is an accurate segmentation into notes, combined with the knowledge of the actual note length in beats from the score. Some of the presented features are similar to the score-independent features, with the notable difference that in this case, the reference is the actual score value rather than, e.g., the closest pitch on the equally tempered scale.

- *note steadiness* ($d_{\mathrm{n}} = 12$): The mean, standard deviation and the percentage of pitch values deviating more than one standard deviation from the expected midi pitch are computed (compare: d_{p1}. Of these three features, aggregate values over all the notes in the performance are computed in the form of mean, standard deviation, maximum, and minimum value. These features are designed to capture the accuracy of the student's intonation.
- *duration histogram features* ($d_{\mathrm{d}} = 6$): This feature uses the distribution of note lengths played by the students for the one most frequently occurring note length in the score (e.g., quarter note). We compute the histogram (50 bins) of the durations for these notes as played by the student. The same standard statistical measures as introduced for the score-independent timing accuracy features are extracted.

- *DTW based features* ($d_{\text{dtw}} = 2$): The DTW alignment cost normalized by the DTW path length and the slope deviation of DTW path from a straight line are used to capture how close the pitch contour fits the MIDI pitches from the score.
- *note insertion ratio* ($d_{\text{nir}} = 1$): The note insertion happens when an intended note in the score is separated into multiple segments by silences due to student's playing. The duration ratio of total silences to the total pitched region across all the notes is used as a feature.
- *note deletion ratio* ($d_{\text{ndr}} = 1$): Note deletions are found by by detecting notes with duration less than 17ms (3 frames) in student's playing. The duration ratio of these notes to the total pitched region in the student's performance is used as a feature.

The overall number of score-based features is

$$d_{\text{SB}} = d_{\text{n}} + d_{\text{d}} + d_{\text{dtw}} + d_{\text{nir}} + d_{\text{ndr}} = 22$$

3.2 Regression

Using the extracted features from the audio signals, a Support Vector Regression (SVR) model with a linear kernel function is trained to predict the human expert ratings. The libsvm [22] implementation of this model is used with default parameter settings. A Leave One Out cross-validation scheme is adopted along with 5% outlier removal to train a model with 2 years of data and test it on the remaining year. Thus, there are 3 combinations of train and test sets. We report the average test evaluation values over each year as the test year. Predicted values that exceed the range of the allowed scores are truncated to 0 or 1.

4 Dataset

The dataset used for this study is provided by the Florida Bandmasters Association (FBA). The dataset has audio recordings of students and accompanying assessments from expert judges of the Florida all-state auditions for three years (2013–2015). There are three groups of students: middle school (7th & 8th grade), concert band (9th & 10th grade), and symphonic band (11th & 12th grade). Auditions are conducted for 19 types of instruments. The pitched instrument audition includes 5 different exercises, namely lyrical etude, technical etude, chromatic scale, 12 major scales, and sight-reading. The musical score of the technical exercise is announced by the FBA. For each exercise, the judges use assessment categories such as *musicality*, *note accuracy*, *rhythmic accuracy*, *tone quality*, *artistry*, and *articulation*. The maximum score given by the judges for each of the exercises varies from 5 to 40. In our experiments, all of the ratings are normalized to a range between 0 and 1, with 0 being the minimum and 1 being the maximum allowed score. The audio recordings have a sampling rate of 44100 Hz and are encoded with MPEG-1 Layer 3.

Table 1: Per year statistics of the used audio recordings

Year	Total Duration (s)	Average Duration (s)	#Students
2013	3997	32.7	122
2014	4605	30.9	149
2015	2991	24.3	123

To narrow the scope of this study, only a small subset of this dataset is used. We are focusing on the technical exercise played by the middle school student performers for the instrument Alto Saxophone. This instrument was selected because it has comparably high number of students. The judges' assess the categories musicality, note accuracy, rhythmic accuracy, and tone quality. There are a total 394 students performing with an average performance length of approx. 30 s. Table 1 shows additional details of the used part of the dataset.

5 Experiment

The suitability of the 3 feature sets is investigated by comparing the regression model outputs with the ground truth expert assessments for all categories: musicality (L1), note accuracy (L2), rhythmic accuracy (L3), and tone quality (L4).

5.1 Experimental setup

We conduct 5 experiments:

- E1: baseline features,
- E2: score-independent features,
- E3: score-based features,
- E4: score-independent plus score-based features,
- E5: score-independent plus score-based features evaluated on the combined dataset.

We did not perform experiments with the combination of all feature sets due to the high dimensionality of the combined set. Each experiment is carried out with 3-fold cross validation. In the first four experiments (E1–E4) the regression model is trained on two years and tested on the remaining year. The average performance over the three years is reported as final result. In the E5 experiment set, the 3 folds contain approximately the same amount of data from each year. An outlier removal process is included in the training. This process removes the training data with the highest prediction residual (prediction minus actual rating); it is repeated until 5% of the data are eliminated. By removing the outliers, the regression models should be able to better capture the underlying patterns in the data.

5.2 Evaluation metrics

The performance of the models is investigated using the following standard statistical metrics: the Pearson correlation coefficient r and the R^2 value. These metrics are commonly used to evaluate the strength of the relationship between the regression predictions and ground truth. Details of the mathematical formulations can be found in [23].

6 Results & Discussion

The results of experiments E1 to E5 are presented in Table 2 using the metrics introduced above. All correlation results, except E1 for labels L1, L2, L3 and E2 for label L2, are significant ($p < 0.05$). All results have a standard error less than 0.2.

As expected, the results show that the baseline features (E1) are clearly outperformed by the other feature sets with designed features (E2–E5). The baseline features are essentially unable to capture useful information for the assessment of student performances. Baseline features are seen to show some correlation with L4, suggesting that some limited meaning with respect to tone quality can be captured.

The score-based features (E3) show generally higher correlation coefficients than the score-independent features (E2) in all the assessment categories. This is expected as the score-based features should be able to model the assessments better due to the additional score information.

Table 2: Result table to compare the experiments. Labels L1, L2, L3, L4 correspond to musicality, note accuracy, rhythmic accuracy and tone quality

	Label	L1	L2	L3	L4
E1	r	0.19	0.07	0.14	0.21
	Rsq	-0.51	-0.15	-0.48	-0.43
E2	r	0.49	0.25	0.34	0.31
	Rsq	0.04	-0.16	-0.25	-0.30
E3	r	0.56	**0.42**	0.39	0.39
	Rsq	0.13	**0.13**	-0.08	-0.09
E4	r	0.58	0.37	0.47	0.42
	Rsq	0.05	-0.03	-0.02	-0.14
E5	r	**0.64**	0.37	**0.60**	**0.46**
	Rsq	**0.33**	0.05	**0.34**	**0.13**

The correlation coefficient increases for rhythmic accuracy (L3) when score-based and score-independent features are combined (E4). Interestingly, this is not true for the category note accuracy (L2) and only marginally true for musicality and tone quality. Investigating this result, we found that the results for the year 2014 are responsible for the drop: It turns out that the regression output is unreliable because of different feature ranges between training set (2013 and 2015) and test set (2014) in this case. This indicates that this training set might not be representative enough; possibly, the different musical pieces impact the score-dependent features more significantly than expected. Other possible reasons include the designed features being unable to model the L2 category or the ground truth somehow being unreliable for this year. In addition, not much improvement is seen in E4 for the musicality label. The minimal increase in E4 for musicality (L1) and tone quality (L4) could hint at redundancies between features sets, incomplete feature sets (missing features to model important characteristics of the performance), varying sound quality of the recordings, or disagreement on the definition and assessment of broad categories such as musicality and tone quality.

The experiment E5 shows improved Rsq and correlation values for L1, L3, L4. These results clearly indicate that a large and representative training set is necessary and helpful. There is no difference in correlation for note accuracy, suggesting the need to look into feature normalization or other possible issues with the data for the year 2014.

7 Summary & Conclusion

The goal of this study is to investigate the power of custom-designed features for the assessment of student music performances. More specifically, we compare a baseline feature set (low-level instantaneous features) with both score-independent and score-based features. The data used in this study covers Alto Saxophone recordings of three years of student auditions rated by experts in the assessment categories of musicality, note accuracy, rhythmic accuracy, and tone quality.

As expected, the baseline features are not able to capture any qualitative aspects of the music performance so that the regression model mostly fails to predict the expert assessments in all categories (except, to a limited degree, for tone quality). Score-based features are shown to represent the data generally better than score-independent features in all categories. The combination of score-independent and score-based features show some trend to improve results, but the gain remains small, hinting at redundancies between the feature sets. The tone quality category seems to require additional features to be properly modeled; possible candidates include note-based timbre features.

Overall, the best results for all categories (except note accuracy, see above) were obtained using score-independent and score-based features combined and a training set including recordings from all three years. The results indicate the general effectiveness of the features and are generally encouraging. However, they are still not in a range that would allow for reliable automatic assessment.

There are aspects of the student performances that cannot be represented with the current feature set. For example, a student may stop playing after a mistake in her performance and start over again (or not continue at all). In rare cases, sounds of adjacent student auditions were interfering with the recording. While an approach such as feature learning would be more "modern" than designing features with expert knowledge, it is the belief of the authors that it will be hard to learn such high level features from the data without expert interaction. However, with the data set hopefully expanding each year, feature learning becomes a viable option. For instance, sparse coding and Restricted Boltzmann Machines were reported to be effective for feature learning to predict note intensities of performances [24]. Thickstun et al. report neural networks outperforming handcrafted spectrogram-based features in predicting notes in a performance [25]. Given these examples, feature learning is a direction that we intend to look into in the future.

8 Acknowledgment

The authors would like to thank the Florida Bandmasters Association for providing the dataset used in this study.

References

[1] Wesolowski, B. C., Wind, S. A., and Engelhard Jr., G., "Examining Rater Percision in Music Performance Assessment: An Analysis of Rating Scale Structure using the Multifaceted Rasch Partial Credit Model," *Music Perception*, 33(5), pp. 662–678, 2016, ISSN 15338312, doi:10.1525/MP.2016.33.5.662.

[2] Thompson, S. and Williamon, A., "Evaluating evaluation: Musical performance assessment as a research tool," *Music Perception*, 21(1), pp. 21–41, 2003.

[3] Schedl, M., Gómez, E., and Urbano, J., "Music Information Retrieval: Recent Developments and Applications," *Foundations and Trends® in Information Retrieval*, 8(2–3), pp. 127–261, 2014, ISSN 1554-0669, doi:10.1561/1500000042.

[4] Ewert, S., Pardo, B., Mueller, M., and Plumbley, M. D., "Score-informed source separation for musical audio recordings: an overview," *IEEE Signal Processing Magazine*, 31(April), pp. 116–124, 2014, ISSN 1053-5888, doi:10.1109/MSP.2013.2296076.

[5] Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., and Klapuri, A., "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, 2013, ISSN 0925-9902, doi:10.1007/s10844-013-0258-3.

[6] Dittmar, C., Cano, E., Abeßer, J., and Grollmisch, S., "Music Information Retrieval Meets Music Education." in *Multimodal Music Processing*, volume 3, pp. 95–120, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, ISBN 9783939897378, doi:10.4230/DFU.Vol3.11041.95.

[7] Lerch, A., *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, Wiley-IEEE Press, Hoboken, 2012, ISBN 978-1-118-26682-3.

[8] Seashore, C. E., *Psychology of Music*, McGraw-Hill, New York, 1938.

[9] Allvin, R. L., "Computer-Assisted Music Instruction: A Look at the Potential," *Journal of Research in Music Education*, 19(2), 1971.

[10] Nakano, T., Goto, M., and Hiraga, Y., "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 12, pp. 1706–1709, 2006.

[11] Romani Picas, O., Parra Rodriguez, H., Dabiri, D., Tokuda, H., Hariya, W., Oishi, K., and Serra, X., "A Real-Time System for Measuring Sound Goodness in Instrumental Sounds," in *Proc. of the 138th Audio Engineering Society Convention*, 2015.

[12] Barbancho, I., de la Bandera, C., Barbancho, A. M., and Tardon, L. J., "Transcription and expressiveness detection system for violin music," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 189–192, IEEE, 2009.

[13] Mion, L. and De Poli, G., "Score-independent audio features for description of music expression," *IEEE Trans. on Audio, Speech, and Language Processing*, 16(2), pp. 458–466, 2008.

[14] Han, Y. and Lee, K., "Hierarchical Approach to Detect Common Mistakes of Beginner Flute Players." in *International Society for Music Information Retrieval (ISMIR)*, pp. 77–82, 2014.

[15] Wu, C.-W., Gururani, S., Laguna, C., Pati, A., Vidwans, A., and Lerch, A., "Towards the Objective Assessment of Music Performances," in *Proc. of the International Conference on Music Perception and Cognition (ICMPC)*, pp. 99–103, San Francisco, 2016, ISBN 1-879346-65-5.

[16] Abeßer, J., Hasselhorn, J., Dittmar, C., Lehmann, A., and Grollmisch, S., "Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils," in *Proc. of the 10th International Symposium on Computer Music Modelling and Retrieval (CMMR)*, 2013.

[17] Fukuda, T., Ikemiya, Y., Itoyama, K., and Yoshii, K., "A Score-Informed Piano Tutoring System With Mistake Detection And Score Simplification," *Proc. of the Sound and Music Computing Conference (SMC)*, 2015.

[18] Devaney, J., Mandel, M. I., and Fujinaga, I., "A Study of Intonation in Three-Part Singing using the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT)," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, ISMIR, 2012.

[19] Mayor, O., Bonada, J., and Loscos, A., "Performance analysis and scoring of the singing voice," in *Proc. of the 35th AES Conference on Audio for Games*, pp. 1–7, 2009.

[20] Tsai, W.-H. and Lee, H.-C., "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Trans. on Audio, Speech, and Language Processing*, 20(4), pp. 1233–1243, 2012.

[21] Tzanetakis, G. and Cook, P., "Musical genre classification of audio signals," *IEEE Trans. on Audio, Speech and Language Processing*, 10(5), pp. 293–302, 2002.

[22] Chang, C.-C. and Lin, C.-J., "LIBSVM: a library for support vector machines," *ACM Trans. on Intelligent Systems and Technology (TIST)*, 2(3), p. 27, 2011.

[23] McClave, J. T. and Sincich, T., *Statistics*, Prentice Hall, Upper Saddle River, NJ, 2003.

[24] Grachten, M. and Krebs, F., "An assessment of learned score features for modeling expressive dynamics in music," *IEEE Trans. on Multimedia*, 16(5), pp. 1211–1218, 2014.

[25] Thickstun, J., Harchaoui, Z., and Kakade, S., "Learning Features of Music from Scratch," *arXiv preprint arXiv:1611.09827*, 2016.