AUTOMATIC ASSESSMENT OF SIGHT-READING EXERCISES

Jiawen Huang
Center for Music Technology
Georgia Institute of Technology
Atlanta, Georgia
jhuang448@gatech.edu

Alexander Lerch
Center for Music Technology
Georgia Institute of Technology
Atlanta, Georgia
alexander.lerch@gatech.edu

ABSTRACT

Sight-reading requires a musician to decode, process, and perform a musical score quasi-instantaneously and without rehearsal. Due to the complexity of this task, it is difficult to assess the proficiency of a sight-reading performance, and it is even more challenging to model its human assessment. This study aims at evaluating and identifying effective features for automatic assessment of sight-reading performance. The evaluated set of features comprises task-specific, hand-crafted, and interpretable features designed to represent various aspect of sight-reading performance covering parameters such as intonation, timing, dynamics, and score continuity. The most relevant features are identified by Principal Component Analysis and forward feature selection. For context, the same features are also applied to the assessment of rehearsed student music performances and compared across different assessment categories. The results show potential of automatic assessment models for sight-reading and the relevancy of different features as well as the contribution of different feature groups to different assessment categories.

1. INTRODUCTION

Sight-reading, also known as *prima vista*, describes the task of reading and performing an unknown piece of music from its musical score with little or no preparation. It is a challenge to most students who are learning a musical instrument.

Sight-reading performance reflects the player's ability in different aspects including reading music, applying fingering and playing techniques, and interpreting music in a relatively short time. As an important skill for musicians, sight-reading is often part of school curricula as well as auditions for professional orchestras [6]. The assessment of sight-reading in auditions and teaching environments faces multiple difficulties. While there are efforts to make human assessments comparable and "less subjective," for example by using grading rubrics, the fairness of the assessment can be impacted by bias effects (gender, ethnicity, general appearance, etc.), fatigue effects after hours of listening and assessing, as well as individual preferences and tolerances for various error types. An automatic assessment system can potentially provide objective, repeatable, and unbiased assessments. Thus, it could be helpful both as a tool available to judges to inform their decisions as well as a tutoring system for students providing feedback in individual practice sessions. It can also help understand the important performance parameters of sight-reading assessment and how they compare to the assessment of general (student) music performances.

In this study, we create a prototype and investigate the feasibility of a sight-reading assessment system by designing interpretable features for the task and evaluating the system on a large database of professionally rated recordings. We also inspect commonalities and differences of feature sets for the assessment of sight-reading vs. prepared performances of sheet music. More specifically, we perform feature selection and detailed feature analysis on a score-aligned hand-crafted feature set, identify the most effective features for sight-reading assessment and observe the difference in the assessment ratings of sight-reading and a rehearsed performance.

The paper is structured as follows: the related work on sight-reading assessment is introduced in Sect. 2 and the evaluated features are presented in Sect. 3. Section 4 explains the experiments and discusses the results of the feature analysis. The final Sect. 5 gives concluding remarks and outlines future work.

2. RELATED WORK

2.1 Sight-reading skills and parameters

Sight-reading involves coordination of auditory, visual, spatial, and kinesthetic systems to produce an accurate and musical performance [11]. In sight-reading exercises, multiple layers of visual information are processed simultaneously when reading the score while playing the instrument. Besson et al. has demonstrated distinct processing between melodic and rhythmic information [2]. This indicates that pitch accuracy and rhythmic accuracy can be treated as two independent assessment categories. Elliott found a strong positive relationship between wind instrumentalists' general sight-reading ability and the ability to sight-read rhythm patterns [5]. This suggests that features containing rhythmic information are important for assessing a sight-

© Jiawen Huang, Alexander Lerch. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Jiawen Huang, Alexander Lerch. "Automatic Assessment of Sight-reading Exercises", 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 2019.

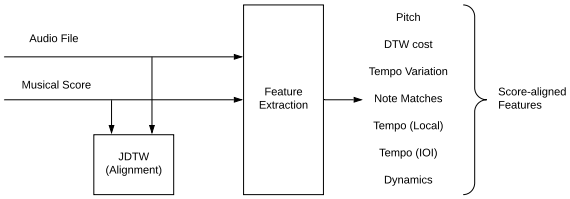

Figure 1. Feature extraction

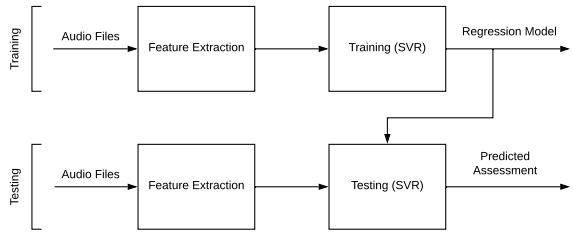

Figure 2. Flow chart of training and testing

reading performance. While intonation, rhythm, and tone quality are typical properties to be assessed, in many cases only an overall rating is given without details on individual properties [1, 4].

2.2 Automatic assessment

There is only a limited number of publications for the automatic assessment of sight-reading. Cheng et al. developed a real-time system for sight-reading evaluation of piano music [4]. The real-time system transcribes the polyphonic music and detects wrong notes. Commercial interest is shown by the existence of systems such as *Sight Reading Practice and Assessment* [1] and *SightReadPlus* [2], which aims at assessing a student playing and tracking the progress of sight-reading.

The automatic assessment of sight-reading has many similarities to the assessment of music performance in general. Therefore, we should expect similar features to be relevant for both tasks and take advantage of the broader spectrum of publications in general performance assessment. Abeßer et al. designed a feature set consisting of 138 features based on the pitch contour of students' vocal and instrumental performances, applied feature selection and used the selected features to train a Support Vector Machine (SVM) [1]. They found that features describing the similarity of score and audio, and the variability of note durations are the most impactful features. Fukuda et al. presented a piano tutoring system which applied non-negative matrix factorization for transcription and DTW for audio-to-score alignment [8]. They basically use, similar to Cheng et al. [4], the number of detected mistakes as core information for performance assessment. Wu et al. proposed assessing a performance independent of the musical score using features based on pitch, amplitude, and rhythm histograms [16]. Vidwans et al. extracted a set of pitch, dynamics, and tempo features after aligning the performance to the score with Dynamic Time Warping (DTW) [15]. Their work is followed by Gururani et al., who investigated the impact of hand-crafted descriptors for the assessment of student alto saxophone technical exercises by feature selection [9]. The results reveal that score-aligned features have a higher correlation with human assessments than score-independent features.

More recently, deep learning methods have been applied to automatic performance assessment [14]. Although deep learning might be a useful tool to achieve better performance for the prediction, the current success of such approaches is often impeded by the available dataset sizes which are often insufficient to train the models properly. A maybe even more important drawback of deep learning is that the interpretability is lost in the hidden layers, so that systems based on deep learning might not be able to give meaningful detailed feedback to a student. This is the main reason why we focus on hand-crafted, knowledge-based features in this study.

3. FEATURE EXTRACTION

3.1 Overview

The flow chart of feature extraction process is shown in Figure 1. [3] Given a recording of a student's sight-reading exercise, the pitch contour is extracted by pYIN [12] from the audio signal (sample rate 44.1 kHz, window and hop size 1024 and 256 samples, respectively). This pitch contour is then aligned to the score of that piece using a modified DTW algorithm which we refer to as Jump-enabled Dynamic Time Warping (JDTW), a DTW variant which can account for repeated score passages. After the alignment, features that capture pitch, rhythmic, and dynamics properties are extracted. The following sections will introduce JDTW, the extracted features, and the inference model.

3.2 Jump-enabled Dynamic Time Warping

Intuitively, we expect the main difference between sight-reading and the performance of a rehearsed piece of music, besides a higher likelihood of errors and more variability in tempo, to be in a higher probability of the student stopping and restarting from a preceding score position after a pause. The frequent occurrence of these jumps has been verified through informal dataset analysis. As standard alignment approaches such as DTW cannot properly handle such jumps, a modification of the DTW algorithm is necessary to properly align the audio sight-reading performance to the score (in our case in MIDI format). Therefore, we propose a Jump-enabled Dynamic Time Warping (JDTW) which is able to handle these repetitions in the students' sight-reading performance. The approach is inspired by Fre-

[1] `http://standardassessmentofsightreading.com`, Last access: 2019/04/10

[2] `http://mymusicta.com/products`, Last access: 2019/04/10

[3] Source code can be accessed at `https://github.com/jhuang448/FBA_code_2019`

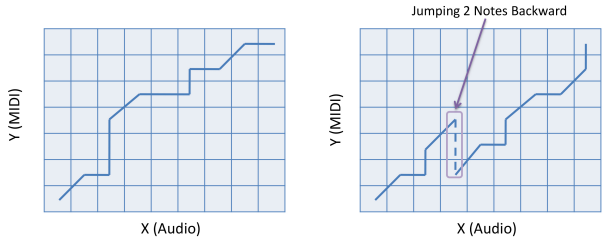

Figure 3. Illustration of paths of index pairs for a sequence X of length N = 9 and a sequence Y of length M = 7. Left: original DTW; Right: JDTW.

merey et al.'s jumpDTW [7] but uses different constraints in terms of potential jump positions and jump lengths.

Dynamic Time Warping (DTW) is a commonly used path finding technique based on dynamic programming to find an optimal alignment between two time series through a pair-wise distance matrix [13]. It has been widely used in speech recognition and musical information retrieval. It only allows sequential alignment, which means that we can neither walk back in a sequence nor jump in time. Given the two sequences $X := (x_1, x_2, ...x_N)$ (audio) of length $N \in \mathbb{N}$ and $Y := (y_1, y_2, ...y_M)$ (midi) of length $M \in \mathbb{N}$, the recursion formula of the accumulated cost matrix D of the classical DTW is as follows:

$$D(n,m) = \min\{D(n-1,m-1), D(n-1,m), D(n,m-1)\} + c(x_n, y_m) \quad (1)$$

for $1 < n \leq N$ and $1 < m \leq M$; $c(x_n, y_m)$ is a measure of distance between x_n and y_m.

The modified accumulated cost matrix D_{J} for JDTW introduces an additional cost term $J(n,m)$ as follows:

$$D_{\mathrm{J}}(n,m) = \min\{D_{\mathrm{J}}(n-1,m-1), D_{\mathrm{J}}(n-1,m), D_{\mathrm{J}}(n,m-1), J(n,m)\} + c(x_n, y_m) \quad (2)$$

in which $J(n,m)$ is the minimum accumulated cost for a path jumping to point (n,m):

$$J(n,m) = \begin{cases} \min\limits_{i \leq I}\{D_{\mathrm{J}}(n-1, m+i) + p\}, & \text{pause before n} \\ \infty, & \text{otherwise} \end{cases} \quad (3)$$

for $1 < n \leq N$ and $1 < m \leq M$, where I is the largest distance in notes allowed for a jump and p is the penalty for jumps. Figure 3 illustrates the paths of the original DTW and the JDTW for an example.

3.2.1 Parametrization and implementation

The adjustment of two JDTW parameters is essential: I, the maximum length of a jump in notes, and p, the penalty of the jump itself. The parametrization with the lowest accumulated cost is found empirically from a simulated validation set of 120 synthesized sound files, leading to the values of $I = 3$ and $p = 3 \cdot \mathrm{mean}(C)$, in which C is the cost matrix between X and Y (meaning that the penalty depends on the average cost).All other DTW-related parametrizations follow standard settings.

Two details are noteworthy in the context of the current implementation: (i) after obtaining the pitch contour from the audio and before computing the alignment, silent frames are temporarily removed from both pitch contour and MIDI sequence, and (ii) the distance between pitch contour x_n and MIDI pitch y_m is computed after tuning frequency adjustment as the octave-independent *wrapped* distance to eliminate pYin's frequent octave errors, however, a small penalty of 1 is added for distances equal or higher than 12 to account for possible octave jumps in the score. After successful application of JDTW, each audio frame is aligned to a note in the MIDI sequence.

Index	Group	Description
1–8	Pitch	Mean and std of pitch dev. (mean, std, max, min)
9–11	DTW cost	Cost of whole path, jumped path and correct path
12–14	Tempo var.	Slope dev., number and distance of jumps
15–16	NIR, NDR	% of silence inserted % of short notes
17–18	Tempo (local)	Inversed tempo per note (mean, std)
19–24	Tempo (IOI)	Crest, bin resolution, skewness, kurtosis, roll-off, power ratio of the IOI histogram
25–32	Dynamics	amplitude envelope and amplitude spikes (mean, std, max, min)

Table 1. Overview of extracted features.

3.3 Feature set

The evaluated feature set can be divided into seven categories: pitch, DTW cost, tempo variation (DTW-based), note matches, tempo (local), tempo (Inter-Onset-Interval-based), and dynamics. Table 1 lists all 32 features explained below with their feature indices.

- **Pitch** ($d = 8$)**:** For each note, the mean and the standard deviation of the pitch deviation from the MIDI pitch is computed. Then, these features are aggregated over the whole performance using mean, standard deviation, maximum, and minimum of each series. The resulting eight features are used to capture intonation accuracy.
- **DTW cost** ($d = 3$)**:** As a result of the alignment, we can compute three cost metrics from the path. The first one is the overall cost of the whole JDTW path. The second cost is the cost of the discarded parts, i.e., the accumulated cost of all the repeated parts except the last run. The third cost is the overall cost of the path ignoring these discarded parts. These three cost features are normalized by the length of the overall

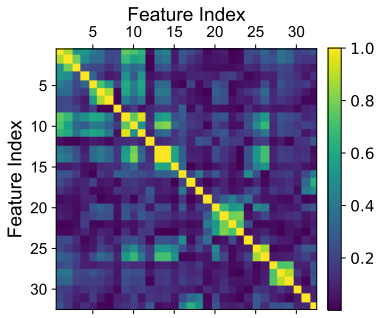

Figure 4. The covariance matrix among features.

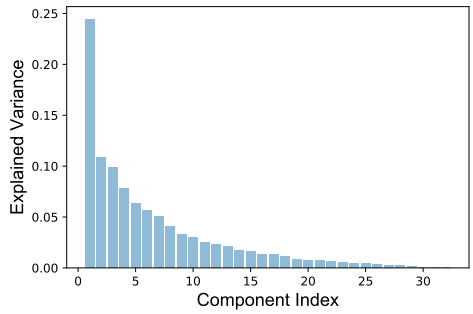

Figure 5. Explained variance for each principle component.

path, and are a measure of pitch similarity between the two sequences.

- **Tempo variation (DTW-based)** ($d = 3$)**:** In addition to the cost-based features, additional features can be extracted from the alignment path. We extract the deviation of the path slope from the diagonal of the matrix, the number of jumps, and the total accumulated distance of jumps.
- **Note matches** ($d = 2$)**:** The Note Insertion Ratio (NIR) is a feature representing additional notes in the student performance, and the Note Deletion Ratio (NDR) represents the missing notes in the performance. As the alignment is performed on pitch contour after removing all the silent frames, these frame have to be inserted back. It is possible that a note is split into multiple notes and that very short (less than 3 frames) notes occur. The NIR is the duration ratio of the inserted silence to the total duration of pitched region. The NDR is the duration ratio of very short notes to the duration of pitched region.
- **Tempo (local)** ($d = 2$)**:** The mean and the standard deviation of the inverse of the tempo per note is an estimate of the overall (inverse) tempo and its variability. For example, an eighth note lasting 1 s results in an inverse local tempo of $\frac{8\,\text{notes}}{1\,\text{s}}$.
- **Tempo (IOI-based)** ($d = 6$)**:** From the histogram of Inter-Onset-Intervals, the crest factor, bin resolution, skewness, kurtosis, roll-off, and the peak power ratio (ratio of the sum of the peak values to the sum of all histogram values) are extracted. These features describe general tempo characteristics.
- **Dynamics** ($d = 8$)**:** For every note, the standard deviation of the envelope as well as amplitude spikes (number of sharp amplitude changes within a note) is computed. Similar to the pitch features, the mean, standard deviation, maximum, and minimum are aggregated over all notes. The resulting eight features are used to capture the dynamic properties of the performance.

4. METHODOLOGY

Our assessment system follows a general machine learning setup as visualized in Figure 2. Our evaluation aims at not only investigating the general feasibility of assessing sight-reading automatically, but also an analysis of which features are most relevant. Furthermore, a general music performance assessment is compared with sight-reading assessment in order to identify similarities and differences between the two tasks.

This section first introduces the dataset used. Then, feature analysis is performed with Principal Component Analysis and forward feature selection. Finally, the performance and features of sight-reading assessment and general music performance assessment is studied.

4.1 Dataset

The dataset used for this study is provided by the Florida Bandmasters Association (FBA). It consists of audio recordings of Florida All-State auditions of middle and high school students in the three years 2013, 2014, and 2015. Each recording consists of exercises such as etudes, scales, and sight reading and provides one expert assessments per exercise in four categories: *musicality*, *note accuracy*, *rhythmic accuracy*, and *tone quality*. For this study we focus on the first three categories. Only a subset of this dataset is used: we are focusing on the sight-reading exercise played by middle school student performers for the instrument Alto Saxophone. The recordings of technical exercise are used to compare sight-reading assessment with the assessment of prepared and rehearsed performances. There are a total of 391 students' audition recordings in the 3 years. Each recording contains technical exercise, sight-reading exercise, and other sections. The total lengths of technical and sight-reading exercise recordings are 192 minutes and 344 minutes, respectively. As the rating scales differ over years and categories (most of the ratings are given within 0–10, others have the ranges 0–5, 0–15, and 0–20), they are all linearly mapped to our target range $[0, 1]$.

The musical score of the sight-reading exercise has been transcribed manually after reviewing multiple highly-rated performances from the three years.

4.2 Principal Component Analysis

Principal Components Analysis is a method to linearly transform a set of possibly correlated variables into a set of uncorrelated variables (components). For the presented analysis, we will use both the covariance matrix of the features and the PCA loading matrix.

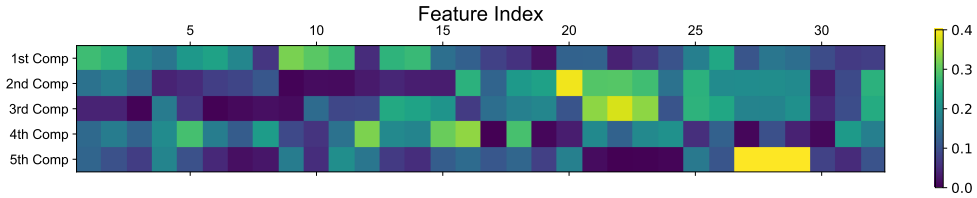

Figure 6. The PCA loading matrix.

The covariance matrix of the features is shown in Figure 4. It can be observed that —unsurprisingly— features are correlated with each other within groups. This is true for pitch features (1–3) and (5–7), DTW cost features (9–11), Tempo variation features (13,14), Tempo (IOI) features (20–23), as well as Dynamics features (25,26) and (27–29). This expected result shows that features within one group carry similar information and verifies that the proposed feature grouping is reasonable.

In addition to high correlation within each feature group, some high correlation is observed across groups. The pitch features (1–7), for instance, are highly correlated with DTW cost features (9–11). This is the case because the DTW-cost features are the accumulated difference between the pitch being played and the reference pitch. The cost of the jumped path (10) is highly correlated with number and distance of jumps (13,14). All of these three features are a measure of the amount of jumps in the performance. The std of the amplitude envelope (26) is also correlated with the jump features. One possible reason for this is that a high number of pauses and jumps might significantly impact the amplitude variation. Other feature correlations are less interpretable; for example, the correlation between min of amplitude spikes (32) and mean of the inverse local tempo (17) is not easily explained.

Figure 5 displays the explained variance by principal components. The eigenvalue of the first component is considerably higher than that of the following components. The first five components explain 60% of the total variance. The loading matrix, shown for these first five components in Figure 6, indicates that the first component is mostly a combination of pitch features (1–7) and the pitch-related DTW cost features (9–11). Both the second and the third component are combinations of rhythmic IOI features with the second focusing on tempo (20) and the third component mostly describing tempo variation (21–23). While the interpretation of the fourth component is difficult, the fifth component clearly represents dynamics (27–29).

4.3 Inference

A SVM Regression model is trained using the extracted features. As a linear kernel gave comparable results to an RBF kernel, the linear kernel was chosen for sake of simplicity. Libsvm [3] is used as implementation.

4.4 Forward Feature Selection

While the PCA gives us insights into feature correlation and which features contribute most to explaining the variance in the feature set, it is of limited use in deciding which features contribute most to the assessment task. In order to identify these, we apply forward feature selection [10]. As this selection approach 'wraps' the target regression algorithm, the selected features will be task-relevant. Forward feature selection is performed on the SVR model with 5-fold cross-validation. The used metric to evaluate success is the R-squared value, which is a common metric for the evaluation of regression systems.

The result of the selection process is a list of features ordered according to their relevance for the task. The indices of the first 10 selected features for each assessment category are listed in Table 2. This table also compares the selected feature sets for sight reading with the sets for a rehearsed student performance. The R-squared results depending on the number of selected features, comparing the sight-reading exercise with the technical exercise, are shown in Figure 7.

4.4.1 Discussion

Figure 7 shows that the R-squared value starts to converge after about 10 iterations of the feature selection. The highest R-squared for musicality and rhythmic accuracy is higher for the practiced performance, while that for note accuracy is higher for sight-reading.

Of the selected features listed in Table 2, two of the dynamics features (25,26) rank high for both practiced performance and sight-reading for all three assessment categories. These two features are the mean and std of the amplitude standard deviation per note. Apparently, the steadiness of loudness plays an important role in assessing the performance.

Looking closer into the Note Accuracy row of Table 2,

Category	Practiced	Sight-reading
Musicality	25,16,21,15,5, 26,11,1,6,18	25,6,32,15,29, 7,26,5,24,23
Note Acc.	9,20,17,3,28, 14,25,21,26,23	26,6,32,15,7, 23,2,9,3,11
Rhythm Acc.	25,21,16,13,26, 18,5,11,2,1	25,6,32,23,15, 29,31,8,20,1

Table 2. The first 10 selected features in forward feature selection. Colors represent different feature groups: **cyan** for pitch, **purple** for DTW cost, **grey** for tempo variance, **apricot** for note matches, **orange** for tempo (local), **pink** for tempo (IOI) and **green** for dynamics.

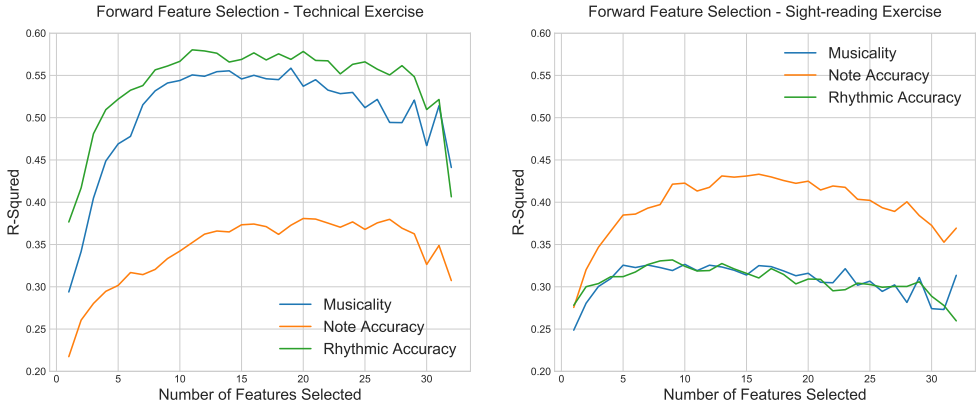

Figure 7. The R-squared curves in feature selection.

we can observe that six of the ten selected features (2,3,6,7,9,11) for sight-reading are features which contribute highly to the first (pitch-related) PCA component. This is not the case for technical exercise, indicating that the pitch features contain more relevant information for sight-reading than for practiced performance. This is also indicated by feature 6 ranking highly in all three assessment categories for sight-reading but not for practiced performance. This feature is one of the aggregated features (standard deviation of absolute differences between played pitch and reference pitch) and is thus a measure of pitch steadiness.

For the assessment categories Musicality and Rhythmic Accuracy, more dynamics features are selected for sight-reading exercise than for the practiced performance. The reason for this might be a different expectation for the two exercises. It might be that, either due to the low complexity of the score or little time for preparation, a dynamically steady performance is preferred by the judges.

During feature selection, the R-squared curve reaches its maximum at about 10–20 iterations and drops dramatically when nearly all the features are selected. This is unexpected behavior for an SVM. A possible reason may be that the dataset is not large enough to train an SVR with all the features or that there might be some 'misleading' features in the feature set.

According to the results above, the automatic assessment of sight-reading is even more challenging than assessing a practiced performance, which performs in the range that we expect (compare [9]) but not so well that it could be considered solved. The higher R-squared for Note Accuracy indicates that our features, especially the intonation features, model this category better for sight-reading than for technical exercise. The low R-squared values for Musicality and Rhythmic Accuracy indicate that we essentially cannot model the human assessments either due to irrelevant features or noisy ground truths. It means that the judges assess the two kinds of exercises differently for these categories and that our regression model fails to capture the information important for sight-reading.

5. CONCLUSION

We presented a feature set of 32 hand-crafted features for the assessment of sight-reading and evaluated them for middle school alto saxophone performances. The feature analysis included PCA and forward feature selection based on the R-squared of the output from an SVR. We can identify the relevant assessment dimensions in the first few principal components and find that the assessment of sight-reading in general is highly influenced by dynamics, and that the assessment of Note Accuracy is mostly focused on pitch-related features. Judging from the absolute results, we can see that the automatic assessment of sight-reading is still an unsolved problem and that the presented features can model a human assessment only imperfectly. In order to be usable in a realistic scenario, we need to either identify additional, more relevant features or move towards state-of-the-art, uninterpretable feature learning solutions. As compared to a practiced and prepared performance, we can identify some commonalities and some differences in the set of relevant features, but the most striking difference is the gap of model performance between assessment categories. Further work is needed to identify where the cause for this gap can be found.

It is likely that rehearsed and sight-reading exercises do not share the same assessment criteria even if the categories are named identically. The performance, as imperfect as it might be, is not assessed by score deviations alone, so that our feature might not represent all critical factors. An additional complication is that in our dataset, we only have the assessment from one judge for each performance. The effect of possible subjectivity and uncertainty makes are complicated task even more challenging. More effort is needed to be able to explain the logic behind the assessment given by judges with quantitative and interpretable indicators before they can be used in music education.

6. ACKNOWLEDGEMENT

We would like to thank the Florida Bandmasters Association (FBA) for providing the dataset used in this work.

7. REFERENCES

[1] J. Abeßer, J. Hasselhorn, A. Lehmann C. Dittmar, and S. Grollmisch. Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils. In *Proc. of the International Symp. on Computer Music Multidisciplinary Research (CMMR)*, Marseille, 2013.

[2] M. Besson and F. Faïta. An event-related potential (erp) study of musical expectancy: Comparison of musicians with nonmusicians. *Journal of Experimental Psychology: Human Perception and Performance*, 21(6):1278, 1995.

[3] C.C. Chang and C.J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 2011.

[4] C.C. Cheng, D.J. Hu, and L.K. Saul. Nonnegative matrix factorization for real time musical analysis and sight-reading evaluation. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, 2008. IEEE.

[5] C.A. Elliott. The relationships among instrumental sight-reading ability and seven selected predictor variables. *Journal of Research in Music Education*, 30(1):5–14, 1982.

[6] A.L.P. Farley. *The Relationship Between Musicians' Internal Pulse and Rhythmic Sight-Reading*. PhD thesis, University of Washington, 2014.

[7] C. Fremerey, M. Müller, and M. Clausen. Handling repeats and jumps in score-performance synchronization. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, 2010.

[8] T. Fukuda, Y. Ikemiya, K. Itoyama, and K. Yoshii. A score-informed piano tutoring system with mistake detection and score simplification. In *Proc. of the Sound and Music Computing Conference (SMC)*, Maynooth, 2015.

[9] S. Gururani, A. Pati, C.W. Wu, and A. Lerch. Analysis of Objective Descriptors for Music Performance Assessment. In *International Conference on Music Perception and Cognition (ICMPC)*, Toronto, 2018.

[10] I. Guyon and A. Elisseefi. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(7-8):1157–1182, October 2003.

[11] C.M. Hayward and J. Eastlund Gromko. Relationships among music sight-reading and technical proficiency, spatial visualization, and aural discrimination. *Journal of Research in Music Education*, 57(1):26–36, 2009.

[12] M. Mauch and S. Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, Florence, 2014. IEEE.

[13] M. Müller. Dynamic Time Warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer, Berlin, Heidelberg, 2007.

[14] A. Pati, S. Gururani, and A. Lerch. Assessment of Student Music Performances Using Deep Neural Networks. *Applied Sciences*, 8(4):507, March 2018.

[15] A. Vidwans, S. Gururani, C.W. Wu, V. Subramanian, R.V. Swaminathan, and A. Lerch. Objective descriptors for the assessment of student music performances. In *Proc. of the AES Conference on Semantic Audio*, Erlangen, 2017. AES.

[16] C.W. Wu, S. Gururani, C. Laguna, A. Pati, A. Vidwans, and A. Lerch. Towards the Objective Assessment of Music Performances. In *Proc. of the International Conference on Music Perception and Cognition (ICMPC)*, pages 99–103, San Francisco, 2016.