Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Automatic Assessment of Timing and Rhythm in Electric Bass for Rock & Pop Repertoire

Colm Forkin

Supervisor: Vsevolod Eremenko

Co-supervisor: Xavier Serra

Aug 2021



Master thesis on Sound and Music Computing

Universitat Pompeu Fabra

Automatic Assessment of Timing and Rhythm in Electric Bass for Rock & Pop Repertoire

Colm Forkin

Supervisor: Vsevolod Eremenko

Co-supervisor: Xavier Serra

Aug 2021



Table of Contents

[1. Introduction 1](#_Toc78816461)

[2. State of the Art 3](#_Toc78816462)

[3. Datasets 4](#_Toc78816463)

[3.1. Numbers of second level headlines should not be indented 4](#_Toc78816464)

[3.1.1. The same holds for third level headline numbers 4](#_Toc78816465)

[4. Methodology 6](#_Toc78816466)

[5. Experiment 7](#_Toc78816467)

[6. Results 8](#_Toc78816468)

[6.1. Yellow 8](#_Toc78816469)

[6.2. Billie Jean 12](#_Toc78816470)

[6.3. Just Looking 18](#_Toc78816471)

[6.4. Brown Eyed Girl 21](#_Toc78816472)

[6.5. Roadrunner 26](#_Toc78816473)

[6.6. Walking on the Moon 30](#_Toc78816474)

[6.7. Technical Control Grades 32](#_Toc78816475)

[6.8. Teacher Comments 33](#_Toc78816476)

[7. Conclusion 36](#_Toc78816477)

[8. List of figures 41](#_Toc78816478)

[9. List of tables 42](#_Toc78816479)

[10. List of symbols 43](#_Toc78816480)

[11. Bibliography 44](#_Toc78816481)

[12. Appendices 45](#_Toc78816482)

[12.1. Appendix 1 45](#_Toc78816483)

[12.2. Appendix 2 45](#_Toc78816484)

Dedication

I would like to thank those near to me personal that made this whole Masters possible, my family in Barcelona. One of the music biographies, I read around 2016, U2 on U2 [2] was the beginning of a long journey which led me to choosing to do the SMC Masters full time. I was fascinated about the musical journey their careers took and hwo the producers played a big role. Fast forward to November 2020 and my thesis topic was chosen, I was know faced with the burden of getting this one chance to really unlock what makes up a great bass sound. In terms of heavy project workload, writing a Thesis is like recording an album. But I felt fortunate, I had Vsevelod and Xavier as supervisors in the same way Adam Clayton felt fortunate in having Brian Eno and Danny Lanois as producers for U2.

Along the way there were a lot of challenging assignments and the one that definitely stick out was the MPC exercise for P. Herrero. I learnt a lot about music and emotions and I found myself getting better at choosing music to listen to help my mood at different occasions ( I really hope that nobody noticed I had Phil Collins on Spotify when I was preparing my presentation)

Acknowledgments

(Optional, if used placed on a right page next to an empty left page)

Abstract

Music Education has undergone significant changes in the last twenty years, with a wide array of applications and online tools emerging to help students learn an instrument autonomously offering automatic feedback. Timing and rhythm are crucial in playing good quality electric bass and although tools exist that help measure their synchronization with the metronome there are some micro-timing improvements that can be made. Experience in preparing for electric bass music exams and the identification of shortcomings in performance assessment tools have been the motivation of this thesis.

Note length and note rests are two missing measurement criteria in state of the art tools. The algorithms and technology exist to do this, but their application has been in automatic music transcription where precision requirements are not as high as they are for music education. This thesis evaluates algorithms for onset and offset detection, offers some new suggestions and tests them on songs with different musical properties on the Rock and Pop repertoire

Keywords:

Audio Signal Processing, Automatic Music Transcription, Bass transcription, durations, electric bass guitar, expression style, expressive performance analysis, fretboard, Machine Learning, Music Assessment, Music Education, Music Information Retrieval, Music Performance Analysis, offset, onset, playing technique, plucking, position, rhythm, source separation, string detection, style, Automatic Music Transcription, ground -truth, Rhythm, Timing

1. Introduction

Using technology to assist in Music Performance Assessment in the context of Music Educations is the core subject matter of this thesis. Audio Signal Processing and Music Information Retrieval are the technologies used and Dittmar etc all [8] gives us a brief history of the role of MIR in Music Education. A key step forward was the transition to digital formats for both recorded and symbolic notation and hence the transition from CDs and score books to today’s smart phone apps. These apps[[1]](#footnote-1) offer performance assessment for learning help guide the student without an expert giving feedback on tuning note accuracy and metronome accuracy. However, although they engage the student well with attractive edutainment front ends (e.g. they include scoreboards for highest accuracies) there are other aspects musical endeavour not well covered such as note duration, articulation, good use of dynamics etc.

A music student learning in a formal context can now get daily feedback from an app and this can complement the practicing habits recommended by a teacher. The same apps offer more engagement with the student, “gamifying” the process with score and league tables of performance for songs. This thesis aims to bring the push the sound analysis technologies further to better support the strict educational requirements for professional music performance.

Typically for aspiring musicians starting out in a Rock and Pop Music, the informal context is where all the learning takes place. It was not uncommon for young people starting out to try form a band before they have even learnt their instruments. Neill McCormack [2] describes how U2 got together in the early days and difficulties they had in trying to get a good sound with friends and older siblings around giving them

feedback during rehearsal. Their initial success was enough to get them into a studio with producer Steve Lillywhite giving them the feedback they needed for professional quality sound recording on their first album.

But even after achieving success, Adam Clayton sought bass lessons from Patrick Pfeiffer in the mid 1990s, (author of “Bass for Dummies”) and gained new heights in performance of the instrument as a result.

Performance assessment of a particular instrument, in this case the bass guitar, has be placed in the context of the goals that a musician wants to achieve and although it maybe have a role in commercial success, the aims and goals are distinct.

The effectiveness of a learning program, i.e. the goals and purpose has to be considered [1] technology into the music classroom. The Trinity Rock and Pop Bass Syllabus [3] is the syllabus chose in this thesis, since it has a good reputation in preparing the musicians with the necessary studio, session, and live performance skills in Modern Music. It focuses particularly on micro-rhythmic skills, which can be measured objectively: plucking the string at the correct time, holding the note for the correct length, technical control of the instrument in order to produce good quality sound, managing the dynamics.

The SOTA will frequently refer to examples from the Datasets under study,but their formal introduction shall be in Chapter 3.

1. State of the Art

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

1. Datasets
   1. Numbers of second level headlines should not be indented
      1. The same holds for third level headline numbers

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

Andrzejak Supp Fig 3.tif

*Figure 1: This is an example of a figure and its caption.*

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

*Table 1: This is an example of a table and its caption.*

|  |  |  |
| --- | --- | --- |
|  | Feature 1 | Feature 2 |
| Experiment 1 | 25 | 23 |
| Experiment 2 | 26 | 25 |

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

1. Methodology

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

1. Experiment

This is an example paragraph. As you can see, the main text uses a font size of 12 pt and a line spacing of 1.5. Neither the paragraphs nor the first lines of paragraphs should be indented.

There is no very strict page limit. Your number of pages will be strongly influenced by the size and total number of your figures and tables. It is recommended staying within 30-50 pages. Do not try to fill as many pages as you can. Longer theses are not necessarily of higher quality and of more non-redundant content than shorter theses. Certainly, a master thesis of 15 pages is too short, and a master thesis of 100 pages is too long.

1. Results

This chapter dedicates a section to each song to discuss the specific musical properties of each song, the impact of choosing different inputs for predicting grades and the reasons for outliers in the Student grades. The final section discusses the Technical Control grade predictions followed by a review of the role played by the Teachers Comments in the grade prediction.

The student recordings were analyzed for their onsets and offset measurements and compared against the ground truth stems. The Machine Learning algorithms used to train the models is Linear Regression with multiple variables. The Cross Validation strategy is that 30% of the total cases are used as Test Cases. The X inputs are the Precision, Recall and F-Measure for both predicting both Onset and Duration Grades. The additional inputs are the Mean or Absolute Mean and Standard Deviation of the Onset/Offset Deviations.

*Table 2: Grades Actual vs Predicted*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ONSET | | DURATION | |
| **Song**. | Actual | Predicted | Actual | Predicted |
| yellow | 76.5  76.5  49.5 | 60.924392  69.760548 47.442291 | 90 63 76.5 | 79.76 78.3 66.13 |
| bjean | 76.500000  68.849998  49.500000  49.500000  76.500000 | 68.763970  55.128501  42.868460  63.165241  78.933811 | 76.5  81.0  63.0  36.0  76.5 | 64.393139  72.570609  67.022827  64.313710  74.541020 |
| just | 79.199997  72.000000  90.000000  56.700001 | 72.854928  73.866432  90.703671  79.140241 | 90.000000  79.199997  90.000000  79.199997 | 78.190844  78.908711  92.321572  83.463884 |
| Brown | ^? | ? | 90.000000  79.199997  90.000000  79.199997 | 78.190844  78.908711  92.321572  83.463884 |
| Road | 76.5  76.5  69.460953 | 71.418147  71.420463  63.0 | 76.5  90.0  76.5 | 65.420814  69.979021 72.623943 |
| Wotm | 63.0  76.5  49.5  63.0 | 83.797603 60.819976 64.732879 65.781555 | 76.5  76.5  49.5  63.0 | 99.302107 68.546260 74.259509 74.963361 |

*Table 3: Grade Prediction Errors*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Song. | # | Algo. | MAE Onset | RMS Error Onset | Extra Inputs | MAE  Duration | RMS Error Duration | Extra Inputs |
| yellow | 8 | SOP | 8.12 | 9.87 | Mean | 11.97 | 12.2 | ABS Mean |
| bjean | 15 | IEC/ SOP | 8.84 | 9.46 | - | 8.84 | 9.85 | ABS Mean Std. Dev |
| just | 11 | SOP | 7.84 | 11.7 |  | 4.67 | 6.38 | Mean |
| brown | 8 | IEC | ? | ? |  | 4.32 | 5.52 | - |
| road | 8 | IEC | 5.54 | 5.58 | - | 11.66 | 13.4 | - |
| wotm | 11 | SOP/IEC | 13.62 | 15.15 |  | 16.87 | 18.3 | Mean |

Table 17 gives an overview of how the different songs responded to predicting the main grades: Onset and Duration.

Four songs used a single algorithm, while two of the songs, “bjean” and “wotm” used a blended algorithm, i.e., different algorithms were applied to different song segments. The criteria for algorithm, selection was the “muteness” of the string after plucking. The muted property is also annotated in the Three-Column “rhythm” csv-format annotation file: [Onset, Muted, Offset] using the IEC algorithm.

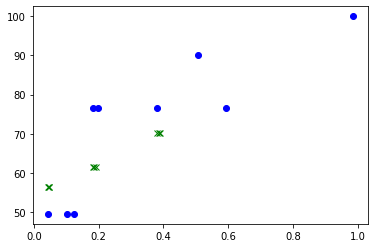
As table 17 shows, Sometimes the addition of these additional inputs did note reduce the Mean Absolute Error. One explanation for this might be the following:  
a student can have a low PRF but it also has a lower Duration Mean and Duration Absolute Mean. The reason this happens is that less onset/offset pairs are considered (through higher missed notes) for calculating the deviations, thus the low number of deviations values for low PRF scoring.

* 1. Yellow

Yellow, being a Grade 0 song is the most basic: there are no rest notes, just repeated notes in the verse with some syncopation and tied notes in the bridge, thus with very short inter-note gaps it was more suitable for SOP than IEC. While the PRF score of the Ground truth was close to 100% the highest graded Student only scored a 50% precision while 2nd highest graded Student 4 for a P.R.F score with a 59% precision.

*Table 17:Best two “yellow” students*

|  |  |
| --- | --- |
|  |  |
| Onset Grade = 90.0 Duration Grade = 90.0 Onset ABS Mean: 0.010412,Onset Mean: 0.002990, Dev. from 0: 0.011992 Offset Mean: 0.014845, Dev. from 0: 0.055124 Articulation Grade = 90.0 Sound Control Grade = 90.0  Volume Control Grade = 90.0 Final Mark = 4.5 Precision = 0.508 | Onset Grade = 76.5 Duration Grade = 76.5 Onset ABS Mean: 0.009035 Onset Mean: -0.000789, Dev. from 0: 0.010740 Offset Mean: 0.000439, Dev. from 0: 0.069981 Articulation Grade = 76.5 Sound Control Grade = 76.5  Volume Control Grade = 49.5 Final Mark = 1.8 Precision = 0.594 |



*Figure 32: Plot of Onset Grades vs Precision*

plt.plot(X\_test,y\_pred, 'x', color='green')  
plt.plot(onset\_precision\_list,onset\_grade\_list, 'o',color='blue')

One of this significant observations in experimenting with this song was to set all notes in the “muted” column in the ”rhythm csv files” to “N”. This means all onset took the next onset

Student 1 is an outlier. Apart from a slight lack of consistency in hitting the notes, when you listen to the Audio you can hear some clicks which cannot be located on the audio waveform. These noise sources were probably due to the settings or the environment of the Sound Card. For Student1, it was noticed that the IEC algorithm returned a 12 % rather than a 4% precision from SOP. It may be that the sound island approach is less sensitive to noise. Usually when doubts like these occur, the best solution is to take more recordings. Three additional ungraded recordings Student9,10,11 have been made for further validation checks.

* 1. Billie Jean

The song was divided into three sections: 1st verse (Muted), Bridge (Non-Muted), Chorus (Muted), so the duration accuracy metric is applied to the Bridge Section. The

The M.A.E increased to 11% when including 'Onset Mean', 21%, when including ‘Absolute Mean’ and 11% and when including Onset Std. Deviation. Generally speaking, overfitting can be a problem with 5 or more X variables. In this particular case, the deviations for the poor grades may not add value because there are bad onsets that did not make it through the minimum window filter. Therefore, the deviations of these “bad onsets” are not considered. If you have two student recordings with very similar Precision, Recall , F Measure Values, then you can do a fair comparison the Statistics

The interesting thing about the Table 17 is that it’s the only result that produces the minimum error when the Standard Deviation of the Duration Deviation is included Although relatively speaking, the Mean Absolute Errors, compare well against the other songs, there are huge prediction errors for the low grades. For Test 3 in table 17, the predicted grade is the same as “Test 0”, i.e. 64.3% but the actual grades are widely different: 36% and 76.5%. The failed grades of 36 are from Students 9 and 10. However, if these two rows are removed and the Linear Prediction is applied on the reduced set , the MAE of 63%. Removing Std Dev. brings is back down to 14%

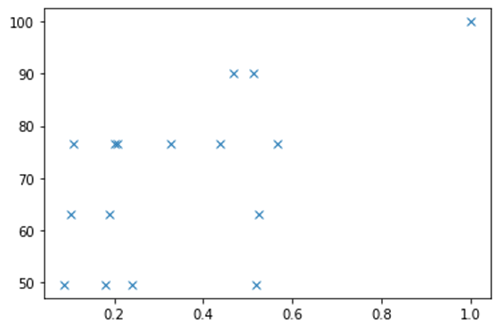
Grading policy may explain small or medium errors, but large outliers like this bring into question whether a policy to set a minimum PRF score for inclusion. The main observation her is that Billie Jean is the song with most Student recordings, however they are more divergent in quality than other songs.

*Table 18:Billie Jean: Actual vs Predicted Grades Overall*

X=dataset[['precision','recall','f\_measure\_value','Onset ABS Mean','Duration ABS Mean']]

|  |  |  |
| --- | --- | --- |
|  | Actual Grade | Predicted Grade |
| 0 | 3.600 | 2.458553 |
| 1 | 3.645 | 1.818748 |
| 2 | 2.700 | 1.369066 |
| 3 | 0.900 | 4.263822 |
| 4 | 3.150 | 3.510796 |
| Mean Absolute Error (M.A.E) | 1.6% | |
| Root Mean Squared Error (RMS Error) | 1.89% | |

The low MAE for the overall grade does not fit well with the prediction for test 3. Figure 33 shows 3 over estimated outliers and two underestimated outliers.



*Figure 33: Plot of Precision (X) vs Onset Grades (Y) Billie Jean*

Lets examine the 50% grades with precision value above 0.5. This could be the result of recording that was so out of time with the onset that it fell behind or ahead into the window of the next onset. It is actually very difficult to play this badly. This recording needs to be identified examined for other metrics to see its properties.

Student 6 has a much higher onset grade than Student 8 despite having the same P value. The reason for this is that Student 6 did not follow the score in the A chord parts, and played no rest notes. The teacher missed this and gave a high grade, so this grade was further scaled down by 90% however this is not enough to reduce its outlier position

You can see that the best (Student 14) and worst (Student 10) student performance show similar statistics, partly because of the “filter effect” of the minimum window size.

*Table 17:Best vs Worst wotm students*

|  |  |
| --- | --- |
| Best Grade | Worst Grade |
|  |  |

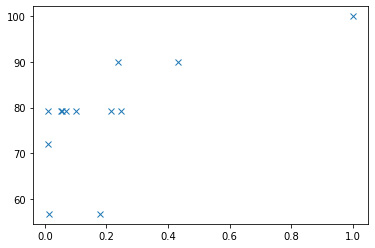
*Table 17:Best two “yellow” students*

* 1. Just Looking

Just Looking had the lowest precision for the Energy Checker: 0.682. This very low precision means this song cannot use Energy Checker, so the SOP algorithm was be applied

Student 4 had best accuracy with a precision of around 0.43, suggesting high sensitivity of algorithm or an overall distinct performance characteristic from ground truth. This student was also the highest graded student summing all the grades together. Student 6 had very poor PRF 0.1) despite having the second best total grade, so these two students will be examined. The best predictions for the Onset and Offset grades came from only considering the PRF inputs.

The modified SOP algorithm is used to calculate this for these Non-Muted notes.

*Figure 36: Just Looking: Grades (Y) Precision (X)*

The left plot shows the Onset grade curve against precision. The right plot shows the Duration grade against precision.

* 1. Brown Eyed Girl

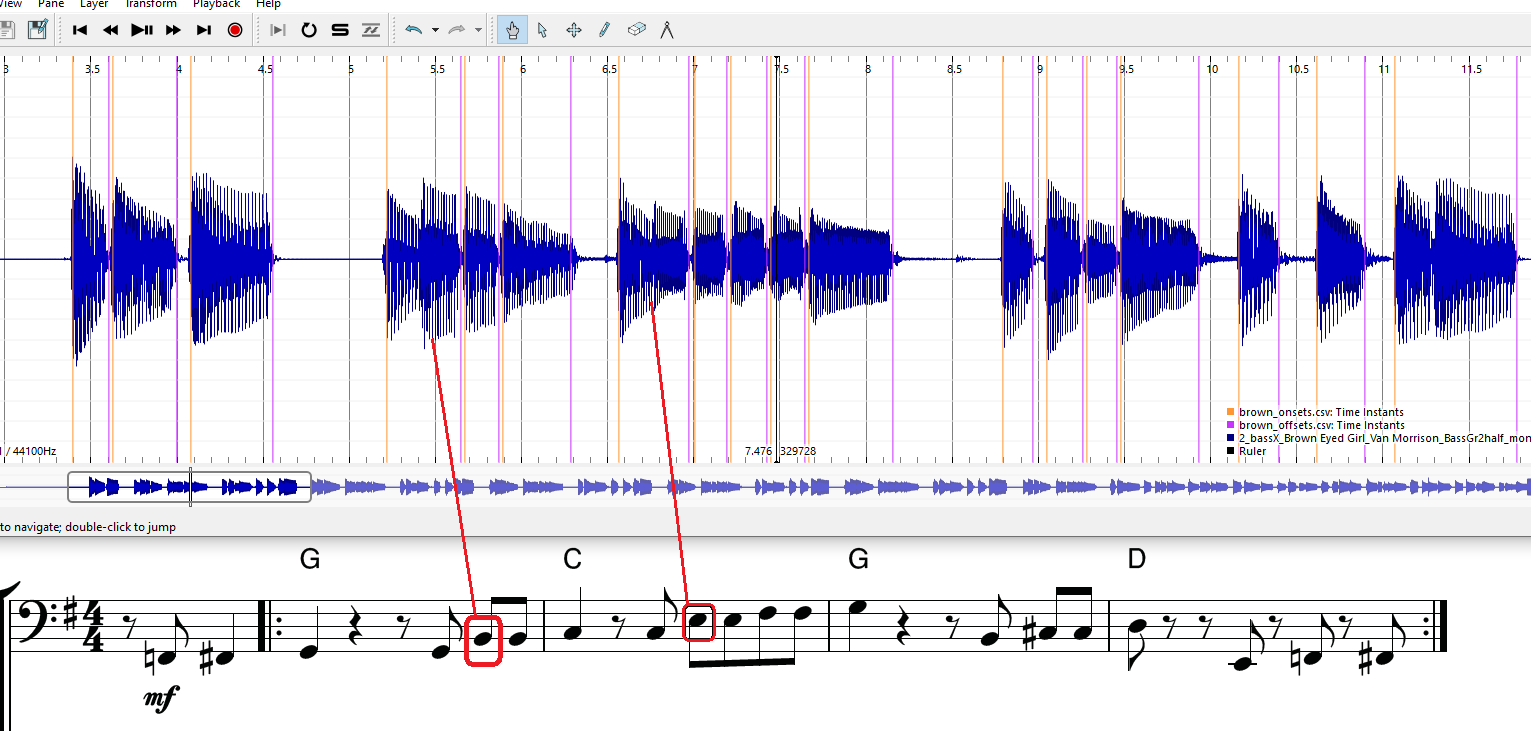
This track had a better onset performance with Energy Checker so we can discard the SOP method as a onset improvement strategy. One way to reduce this could be to partition the song to so that no repeated errors are overly accumulated. Even still this song starts to test the limitations of the algorithms we have and also the MIR evaluation window of 20 ms, since this song has different notes. For example there are six different onsets for the note sequence: C,E,E F# , F#, G

28.223, 28.447, 28.665, 28.885, 29.106, 29.326

So even if the Sound Island approach had better statistics than the SOP, the onsets are simply two close together to reliably separate into sound island, so the following rule: offset = next onset rule should be applied.

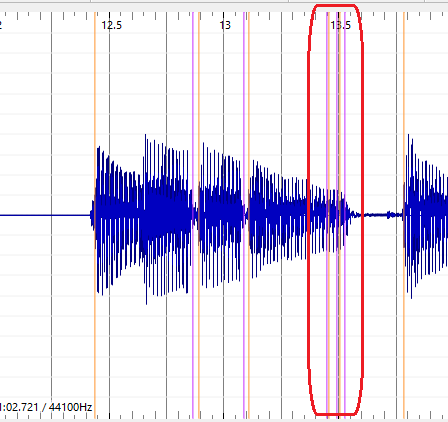
Notwithstanding these limitations, the same procedure of evaluating students was applied. It would be useful to have “confidence formula” that could be applied to the student grade predictions, that could be derived from the accuracy of the ground truth measurements.

The song B.E.G from grade 2 is a good example in highlighting the limits of all the state of the two main algorithms shortlisted and the discussion focuses on how the Recall metric for the IEC methods was only 82 % and the SOP not much better at 85%. There are less false alarms with the IEC but because of the closeness of the notes, there are more Archipelagos instead of sound islands



*Figure 37: Brown: Missing Onsets*

In the first five bars you can see that the closeness of the eight notes result in two Archipelagos. In bar 4 the four notes B, C#, C# and D are all detected as 4 sound islands. In another occurrence of the exact same note sequence of bar 2, not only is there a missing onset of the B note but there are two false alarm sound islands (13.4 seconds). This particular note pattern tends to throw false alarms, thus reducing precision below and acceptable amount.



*Figure 38: Brown: False Onsets*

False Alarms:However, the student performances were well distributed and the analysis

One major outlier is that the best graded student did not correspond with the best PRF results, and they were very low in comparison with the stem. This highlighted the mistake in not letting the teacher consider the stem as a “blind” student performance.  
“brown” clearly has an anomaly with Student 4. This requires a waveform comparison.

|  |  |
| --- | --- |
| *Onset Grades (Y)  Precision (X)* | *Duration Grades (Y)*  *Duration Mean Deviation (X)* |
|  |  |

*Figure 39: Brown: Grade Plots*

* 1. Roadrunner

The tenuto notes for the song Roadrunner, were too weak to be measured and this ambiguity could explain the overall low PRF of the IEC and SOP methods. The second half of the song had no tenuto from 77 seconds to the end. Truncating the second half of the performance would not yield better accuracy and since the student recordings were made of the full song, the full recording was considered for analysis.

The Student grades were in a very range, which suggest a high difficulty with the song. Removing the tenuto part would mean truncating at 34.5 seconds. This resulted in getting a better PRF score using the PRF algorithm. The Grade Prediction for Onset and Duration worked optimally when considering the Mean Onset and Mean Duration respectively alongside the PRF scores

|  |  |
| --- | --- |
| *Onset Grades (Y) Precision (X)* | *Duration Grades (Y) Duration Mean Deviation (X)* |
|  |  |

*Figure 8: Roadrunner: Onset & Duration Grades*

Regarding the Tenuto scenario, a starting point for looking into it would be to examine the two extreme cases: Max Case: All notes equal , Min Case the third note is absent.  
Between these two extremes: an ideal energy statistic should be chosen for tenuto, considering the limit of human perception.

* 1. Walking on the Moon

The PRF before applying this muting classification for WOTM was 0.701 0.932 0.8.

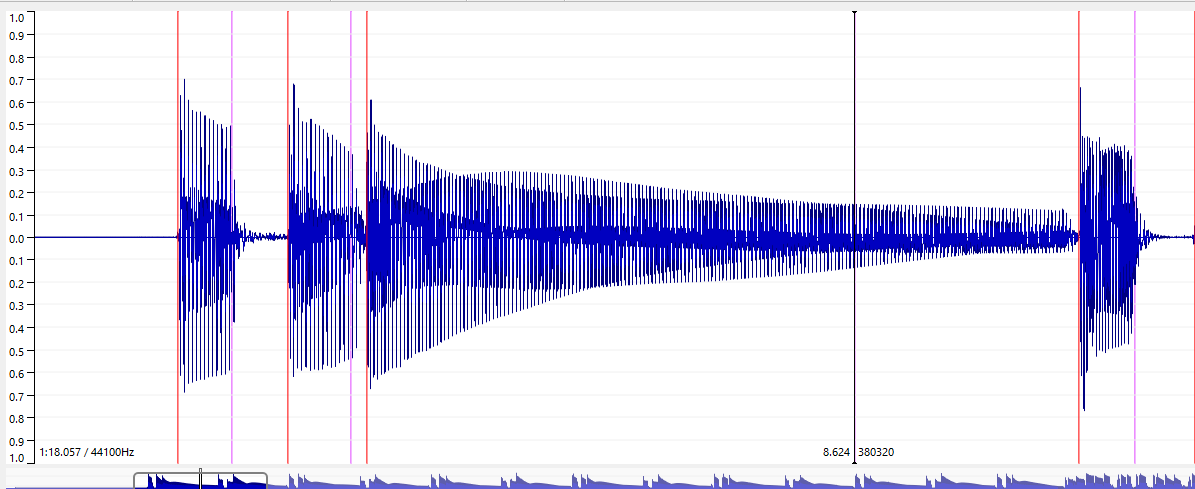
The precision was low for the stem and the range of student precisions was also very low in comparison.

After segmenting the verse into the SOP algorithm and the bridge into the IEC algorithm, the new PRF were as follows;

|  |  |  |
| --- | --- | --- |
| 0.983193 | 0.991525 | 0.987342 |

It is important to point out a trade off made in using SOP for the first part. The very first note of WOTM has a subtle gap, which is ignored if you use the SOP, “offset=next onset” algorithm. How would you determine if this subtle gap is respected in a Student recording? Checking the Energy would be one way, but that requires a lot of effort in determining what is the perceptual drop off point. A better approach is to use an existing library function from Essentia that has already researched the topic [x]

If you take a measurement of the “Effective Duration” using a Threshold setting of 0.05 of the first three notes WOTM you will obtain offset points which approximately align with the human hearing threshold



*Figure 43: WOTM: Offsets using Effective Duration*

The orange line are the onsets and the purple lines are the offset (the black cursor aligns with 3rd offset at 8.624 secs). Any future improvement in automating or semi-automating the assessment of duration should consider these adjusted offsets. This should improve the value that the Mean Duration Deviation has for predicting Duration Grades.

* 1. Technical Control Grades

Bjean having the most recordings is discussed here under the two technical control2% was achieved elements and sound quality. A Mean Absolute Error of 10.6 % in predicting the Articulation Grades using the PRF and Mean Duration Deviation as inputs. The smallest MAE error on predicting the Volume Control Technical Control grades was 7.9% and this was achieved with considering PRF only. For the sound quality grade, no grade prediction was attempted the audio features that the algorithms extract don’t relate directly to sound quality. This requires other audio features to be extracted, such as background noise level, stability of the dynamics which are not the subject of this thesis. Apart from collecting valuable data for future research, having the recordings labelled with Sound Quality in trying understanding how robust the osnet detection is in the presence of noise

*Table 2: Tech Control Grades Actual vs Predicted*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ONSET | | DURATION | |
| **Song**. | Actual | Predicted | Actual | Predicted |
| yellow |  |  |  |  |
| bjean |  |  |  |  |
| just |  |  |  |  |
| Brown |  | ? |  |  |
| Road |  |  |  |  |
| Wotm |  |  |  |  |

*Table 3: Technical Grade Prediction Errors*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Song. | # | Algo. | MAE Onset | RMS Error Onset | Extra Inputs | MAE  Duration | RMS Error Duration | Extra Inputs |
| yellow |  |  |  |  |  |  |  | ABS Mean |
| bjean |  |  |  |  |  |  |  |  |
| just | 11 | SOP |  |  |  |  |  |  |
| brown | 8 | IEC |  |  |  |  |  | - |
| road | 8 | IEC |  |  | - |  |  | - |
| wotm | 11 | SOP/IEC |  |  |  |  |  |  |

* 1. Teacher Comments

This part introduces the topic of text analysis without going into real depth. The intention is to uncover patterns in the text that can help towards predicting scores.  
After Translating the comments using deepl online translator.

A simple correlation of word count with the grades showed that poorer grades resulted in more words in the case of “brown”.

https://deepai.org/machine-learning-model/sentiment-analysis

Some real example of teacher text was used for testing purposes.

https://deepai.org/machine-learning-model/sentiment-analysis

An analysis of the text can be performed on the 48 student performances by correlating the comments with the grades given. A classification algorithm can then be applied to help predict a grade with a given set of comments. We don’t aim to generate text for a given audio recording. Personal experience has shown that low grades (e.g. 5/8 for Day Tripper Grade 5 Exam, Appendix C) tend to be a bit longer on words. A good insight into the teacher comments can allow micro adjustments of the final grade that was given.

A new set of plots can be made to then check if this improves the correlation between audio measurements and the grades given.

Here we present the Machine Learning algorithms used to train the models

Classification for handling the text processing of the examiners report.

1. Conclusion

Despite having the parameter for the minimum window size included, there were some outliers in the duration measurements. There is room for improvement in capturing accurate onset by blending different algorithms together.

Possible Improvements to Duration Accuracy

The main challenge was to do a score fidelity check on long notes, typically compose of tied quarter, eight notes. If a particular song yielded better results for the full extract using SOP, then this meant that there was effectively no strict validity check on whether a long note was properly held. The scope of the project was limited to blending two algorithms. There are other library functions that could be incorporated into the mix to identify the duration of an audible segment in a frame. EffectiveDuration() in Essentia returns the audible segment within a parameterizable threshold. Let us propose that a Student recording could obtain 100% Onset accuracy and that the only imperfection to be measured was the offsets. Let us also consider that this is the non muted scenario where all offsets are the next onset. In this scenario we have a student recording perfectly aligned with the stem onset and offset. In these conditions you could measure the EffectiveDuration of each segment and compare for similarity.

If we measure the opening riff of WOTM the first Effective Duration of note 3 is is xx sec. The Table below captures the manually calculated effective Durations of note 3 of all eight recordings an manually compares them to the grade.

Optimising Mean Absolute Error

We found that in some cases, adding the deviation statistics information did not help in improving the predicted grade. In some cases there were improvements with adding the absolute mean rather than the mean. In all cases adding the Standard Deviation statistic did not improve the prediction.

Logic would dictate, that if the PRF of the students was generally high, then adding the deviations statistics would improve the predictions. This observation was tested in the following table summary:

Choosing Mean over ABS Mean in the statistics, suggest that the grade is more sensitive to whether a note is late or early or too long or too short. Ti make a more in depth analysis to this it would be required to have significantly large number of recordings from diverse students so that separate statistics on early and late onset deviations could be studied.+ The Same could be said of separating the duration deviations into the shorter durations and the longer durations. Musically speaking , the particular song style would dictate the teacher tolerance is to lateness/vs earliness and note longitude.

We found that the teacher may overlook some “imposter notes” and this causes mismatch between mean absolute error and associated grade. A typical audience member may not detect these errors, but a TCL examiner should.

Another significant Teacher characteristic is their ability to hear the required bass duration in the presence of the mix. The first experiment did not have enough bass boost to help the teacher discriminate and this may help explain some of the outliers. This brings us back to the perception topic by Kopp-Scheinpflug [10], discussed in the State of the Art.

Another weakness in the grading is that the teacher used the “go through the students once”, so grade was given an absolute scale rather than a relative scale. To ameliorate this the previous grades allocated by the Bass Teacher were downgraded 90% and the additional grades to bring the Students up to 12 were maintained. The downgrading is built into the code for the specific students 1-8 for each song.

The method for classifying notes can improve the deviation statistics and it opens the door for more sophisticated methods to consider technical control parameters. The Duration grade only has meaning depending on the context. For example, it is only worth measuring duration in the bridge in Yellow. The repeated quavers in the verse leaves no gaps for duration. Theoretically it would be possible to play shorter notes than quavers, but that is actually more difficult to do, so its not worth checking for a minimum note length. Overall properties such as detecting adequate energy levels after a certain time for long notes, is more effective than trying to accurately determine the exact offset point. The testing of the algorithms with of section of the songs taken from midi translations of score highlight the importance of considering articulation as a parameter to be grades.

I made a second iteration of the g

Summary of remaining actions

* Re-check the prediction accuracy by after adjusting final grade from comment analysis
* Improve accuracy of the algorithms with hybrid methods and “next onset” approach
* Perform annotation for the remaining 4 songs (Yellow, Just Looking, Brown Eyed Girl, Roadrunner)
* Record more with deliberate errors (like in the Police song) for better curve alignment
* Perform the grades objectively against the actual score and mark down missing or extra notes
* Perform grading only on Onset and Offset and give a mark between 1-8 and just these two metrics and make song sections shorter.

The last action is the key to opening the door to getting more recordings graded.

Suggested future paths for scaling up the experiment as a follow on to the thesis.

* A pilot project that would involve a selection of up to 20 students who are studying bass guitar in private schools and conservatories.
* A custom portal that would allow them
* Continually annotating recorded data for future training
* Source Separation to obtain more annotated data.

A per-song solution has been found as a work around to the limits encountered in overlapping onsets, offsets. For “Yellow, a song with eight nots ats xxx bpm

Adding a Tenuto marking in the Rhythm Files, could be an extension to the existing methodology to customize algorithms for certain technical focus elements,

1. List of figures

(Optional)

[Figure 1: This is an example of a figure and its caption. 3](#_Toc471218465)

1. List of tables

(Optional)

[Table 1: This is an example of a table and its caption. 4](#_Toc471218472)

1. List of symbols

(Optional)

1. Bibliography

You should use the style used by the journal ‘Nature’ to format your references (see http://www.nature.com/nature/authors/gta/#a5.4). For example a journal article is cited as1

1. Stevenson, G. N., Collins, S. L., Ding, J., Impey, L. & Noble, J. A. 3-D Ultrasound Segmentation of the Placenta Using the Random Walker Algorithm: Reliability and Agreement. *Ultrasound Med. Biol.* **41,** 3182–3193 (2015).

1. Appendices
   1. Appendix 1

*WOTM; This account opened with clear rhythmic drive and placement in the first lines, as the style was portrayed effectively and consistently. A few moments missed complete rounded tone or proficient legato but movement across the instrument with clear picking was achieved. Musical details were observed very well and some confident moments of execution rendered a pleasing musical flow overall.*

*Marks: 8, 7, 9 = 24*

[

"Positive",

"Negative",

"Positive"

]

Brown Eyed Girl The rhythms were accurate in this, and the rests were well counted. Most of the note-lengths were correct apart from the doted-crotchet C in the second bar of the Chorus, which was played short. The notes were mostly fine, and the few alterations in the Pre-Chorus were all fine. For some reason the dynamics were ignored; the crescendo in the 2nd time bar at 28 was absent, and the f chorus was no louder than the mf verse. It was all good otherwise—it just needed more shape. Marks: 7/10/8 = 25

[

"Negative",

"Negative",

"Negative",

"Negative",

"Positive"

]

Yellow Repeated notes were broadly steady, although underlying pules wavered at times, and the dynamic drop in bar 13 was effective. The odd placement error affected flow a little and attack in the chorus was not fully controlled, but syncopation was handled well on the whole. Marks: 7/9/8 = 24

[

"Positive",

"Negative"

]

* 1. Appendix 2

1. Technical Focus Grades are not included in tables 14 and 15.
2. *Table 14:* Inputs (yellow) vs outputs (green) for Billie Jean Onsets

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Stud. | P | R | F | A. Mean | Mean | Std. D | ONSET | OVERALL |
| **0** | **1** | **0.99** | **0.995** | **0** | **0** | **0** | **100** | **5** |
| 1 | 0.328 | 0.315 | 0.321 | 0.008 | 0.002 | 0.009 | 76.5 | 3.6 |
| 2 | 0.519 | 0.531 | 0.525 | 0.006 | 0 | 0.009 | 49.5 | 2.7 |
| 3 | 0.189 | 0.185 | 0.187 | 0.008 | -0.004 | 0.009 | 63 | 3.6 |
| 4 | 0.102 | 0.098 | 0.1 | 0.009 | 0.002 | 0.01 | 63 | 3.6 |
| 5 | 0.206 | 0.21 | 0.208 | 0.007 | -0.001 | 0.009 | 76.5 | 2.7 |
| 6 | 0.201 | 0.206 | 0.203 | 0.009 | 0.006 | 0.01 | 68.85 | 3.645 |
| 7 | 0.107 | 0.108 | 0.108 | 0.009 | -0.004 | 0.011 | 76.5 | 1.8 |
| 8 | 0.239 | 0.259 | 0.248 | 0.008 | 0 | 0.01 | 49.5 | 2.7 |
| 9 | 0.088 | 0.091 | 0.09 | 0.005 | -0.001 | 0.008 | 49.5 | 0.9 |
| 10 | 0.18 | 0.154 | 0.166 | 0.009 | 0 | 0.01 | 49.5 | 0 |
| 11 | 0.568 | 0.455 | 0.505 | 0.007 | -0.001 | 0.009 | 76.5 | 3.6 |
| 12 | 0.525 | 0.469 | 0.495 | 0.007 | -0.002 | 0.009 | 63 | 2.7 |
| 13 | 0.437 | 0.374 | 0.403 | 0.007 | -0.001 | 0.009 | 76.5 | 3.15 |
| 14 | 0.512 | 0.448 | 0.478 | 0.008 | -0.003 | 0.01 | 90 | 4.5 |
| 15 | 0.468 | 0.43 | 0.448 | 0.008 | -0.003 | 0.01 | 90 | 1.98 |

1. *Table 15:* Inputs (yellow) vs outputs (green) for Billie Jean Offset

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Stud. | P | R | F | A. Mean | Mean | Std. D | Acc. | DUR | OVERALL |
| **0** | **1** | **0.99** | **0.995** | **0** | **0** | **0** | **1** | **100** | **5** |
| 1 | 0.328 | 0.315 | 0.321 | 0.046 | -0.017 | 0.14 | 0.48 | 76.5 | 3.6 |
| 2 | 0.519 | 0.531 | 0.525 | 0.023 | -0.009 | 0.036 | 0.57 | 63 | 2.7 |
| 3 | 0.189 | 0.185 | 0.187 | 0.046 | -0.031 | 0.124 | 0.48 | 76.5 | 3.6 |
| 4 | 0.102 | 0.098 | 0.1 | 0.013 | 0.001 | 0.018 | 0.33 | 76.5 | 3.6 |
| 5 | 0.206 | 0.21 | 0.208 | 0.04 | -0.033 | 0.133 | 0.39 | 76.5 | 2.7 |
| 6 | 0.201 | 0.206 | 0.203 | 0.023 | 0.005 | 0.026 | 0.68 | 81 | 3.645 |
| 7 | 0.107 | 0.108 | 0.108 | 0.326 | -0.326 | 0.411 | 0.7 | 76.5 | 1.8 |
| 8 | 0.239 | 0.259 | 0.248 | 0.149 | -0.115 | 0.326 | 0.65 | 63 | 2.7 |
| 9 | 0.088 | 0.091 | 0.09 | 0.014 | -0.001 | 0.018 | 0.94 | 36 | 0.9 |
| 10 | 0.18 | 0.154 | 0.166 | 0.042 | -0.02 | 0.103 | 0.52 | 36 | 0 |
| 11 | 0.568 | 0.455 | 0.505 | 0.045 | 0.03 | 0.181 | 0.4 | 76.5 | 3.6 |
| 12 | 0.525 | 0.469 | 0.495 | 0.026 | -0.005 | 0.078 | 0.57 | 63 | 2.7 |
| 13 | 0.437 | 0.374 | 0.403 | 0.019 | 0.011 | 0.022 | 0.67 | 76.5 | 3.15 |
| 14 | 0.512 | 0.448 | 0.478 | 0.017 | -0.002 | 0.021 | 0.22 | 90 | 4.5 |
| 15 | 0.468 | 0.43 | 0.448 | 0.029 | -0.017 | 0.087 | 0.39 | 49.5 | 1.98 |

1. https://yousician.com/ [↑](#footnote-ref-1)