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](#_Toc79149606)

[2. State of the Art 3](#_Toc79149607)

[2.1. Music Education 3](#_Toc79149608)

[2.2 MIR approaches to onset/offset detection 3](#_Toc79149609)

[2.3 Measuring duration 5](#_Toc79149610)

[2.4 Abessers Research 6](#_Toc79149611)

[2.4 Literature Research Summary 7](#_Toc79149612)

[2.5 Revision of Algorithm Evaluation for Music Pedagogical Purposes 10](#_Toc79149613)

[2.6 Perception of Onsets and Offsets 11](#_Toc79149614)

[2.7 Observations 12](#_Toc79149615)

[3. Datasets 13](#_Toc79149616)

[3.1. IDMT BASS SINGLE TRACK 13](#_Toc79149617)

[3.2. TCL Dataset 15](#_Toc79149618)

[4. Methodology 20](#_Toc79149619)

[4.1 Algorithm Evaluation 21](#_Toc79149620)

[4.2 Indexed Energy Checker ( IEC) 22](#_Toc79149621)

[4.3 Accuracy Metrics 24](#_Toc79149622)

[4.4 Deviation Metrics 28](#_Toc79149623)

[4.5 Correlation with Student Grades Metrics 31](#_Toc79149624)

[5. Experiment 33](#_Toc79149625)

[5.1. Pysimmusic End-to -End test 33](#_Toc79149626)

[5.2. Student Recording Portal 34](#_Toc79149627)

[5.2.1. Latency Test 34](#_Toc79149628)

[5.3. Grading 38](#_Toc79149629)

[6. Results 43](#_Toc79149630)

[6.1. Yellow 45](#_Toc79149631)

[6.2. Billie Jean 46](#_Toc79149632)

[6.3. Just Looking 49](#_Toc79149633)

[6.4. Brown Eyed Girl 50](#_Toc79149634)

[6.5. Roadrunner 53](#_Toc79149635)

[6.6. Walking on the Moon 54](#_Toc79149636)

[6.7. Technical Control Grades 56](#_Toc79149637)

[6.8. Teacher Comments 58](#_Toc79149638)

[7. Conclusion 59](#_Toc79149639)

[7.1. Student Recordings and Teacher Gradings 59](#_Toc79149640)

[7.2. Technical and Musical issues with Experiment 60](#_Toc79149641)

[7.3. Improving prediction strategies 61](#_Toc79149642)

[7.4. Improving Duration measurements 61](#_Toc79149643)

[7.5. Going forward 62](#_Toc79149644)

[8. List of figures 63](#_Toc79149645)

[9. List of tables 65](#_Toc79149646)

[10. Bibliography 66](#_Toc79149647)

[11. Appendices 66](#_Toc79149648)

[11.1. Appendix 1 66](#_Toc79149649)

[11.2. Appendix 2 69](#_Toc79149650)

Dedication

I would like to thank those near to me personal that made this whole Masters program possible, my family in Barcelona. One of the music biographies, I read around 2016, “U2 on U2” [1]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, how the sonic identity developed and how the producers played a big role. Fast forward to November 2020 and my thesis topic was chosen, I am very grateful for having this opportunity to do research on the electric bass guitar for Music Education with Vsvelod and Xavier as supervisors. Thanks to Marti for diligently correcting the Student recordings and to Ramon for sharing his ideas with me.

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

Acknowledgments

Student Recording Assessments: Marti Benach Obradors, Bass Teacher and Post Graduate Student at the Escuela Superior de Música Moderna i Jazz ,Taller de Músics.

Ideas for the Indexed Energy Checker ( IEC) algorithm.

Ramon Romeu, Researcher in Sound and music computing, Aalborg University Copenhagen, Denmark.

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 having prepared 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 the 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, brown, durations, electric bass guitar, expression style, expressive performance analysis, fretboard, ground -truth, Indexed Energy Checker (IEC), Machine Learning, Mean Absolute Error (MAE), Music Assessment, Music Education, Music Information Retrieval, Music Performance Analysis, offset, onset, playing technique, plucking, position, Rhythm, PRF, Sound Archipelagos, Sound Islands, Sound Onset Processor (SOP), source separation, string detection, style, the deviation statistics , TCL Dataset, Timing, wotm (Walking on the Moon), yellow,

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 et all gives us a brief history of the role of MIR in Music Education [2] . 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 on tuning note accuracy, pitch accuracy metronome accuracy. Despite the engaging front ends (e.g., real time feedback, scoreboards for highest accuracies)there are important musical qualities that are not assessed: duration, articulation, good use of dynamics etc.

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 learnt their instruments. Neill McCormack [1] describes how U2 got together in the early days, how they relied on feedback given by friends and Dik Evans (the Edges brother) and how later on, U2 Bassist, Adam Clayton sought bass lessons from Patrick Pfeiffer, author of the book “Bass for Dummies” [3] after having made a series of successful albums.

So, the performance assessment of a particular instrument, in this case the bass guitar can happen at any stage in one’s journey. It has to be placed in the context of the goals that a musician wants to achieve. If the student wants to deviate a bit from the score as permitted in playing jazz style, there is a separate topic that deals modelling expressive performance [4]. This research focuses on the score adherence that Rock and Pop Music demands and bases the performance assessment on the Trinity Rock and Pop Bass Syllabus [5]. 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 thesis opens with the State-of-the-Art Chapter and refer to examples from the two Datasets under study: the Fraunhofer IDMT Single Track Dataset[[2]](#footnote-2) and thesis termed “TCL Dataset”. These two datasets are described in more detail in Chapter 3, focusing on how they help achieved the goals of the Thesis. Chapter 4 then describes the methods used to do the first stage evaluation of the algorithms using these datasets and the seconds stage evaluation which involves testing them on the music student recordings. Chapter 5 describes the experiments carried out to test the current state of the Art algorithms for onsets and to prepare the TCL Dataset for doing the Student Assessment. Chapter 6 summarizes the results of these experiments and points out the main observations. Chapter 7 expands on the observations and implications of the experiment results. It also discusses in detail the lessons learnt from carrying out the experiment. This is important as it is hoped that the methodology used in the experiment can be reused (perhaps configured differently) for further research into music performance assessment for Rock and Pop repertoire.

1. State of the Art

The core of this thesis centres around the research and development of a model that can automatically assess a student’s performance of the four-string fretted electric bass guitar and provide them with the useful feedback that can help them improve. The research will focus on the specific qualities that playing the instrument entails for rhythm and timing aspects: onsets, notes duration and spacing.

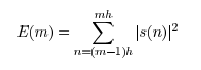
* 1. Music Education

The mastery of any musical instrument without the live presence of teacher is a challenging task. However, there has been interesting developments MIR research that can support Music Education. The algorithms that are used in automated music transcription (AMT) are capable of extracting score information from polyphonic recordings as performed by Salamon/Gomez(SG) [6]. In our music education context, we have the advantage of isolated bass stems as input, and we can take advantage of recent success in transcribing bass lines using data driven methods based on the U-Net Architecture [7] and signal processing methods of bass solos [8] and also tests the effectiveness of extracting plucking features (right hand techniques) and expressive features(left hand techniques) in playing bass. AMT methodologies for bass guitar has been seen useful applications in genre classification [9]and sound synthesis and this research aims to apply these techniques to improve the formal context in which bass music students can improve on their rhythmic precision

The effectiveness of a PAT is can only be accurately measured on the readiness level it provides to a student in Assessment of Learning scenario which can be a final exam, competitions, or audition. This thesis uses the final exam approach with the Trinity College London (TCL) Rock and Pop from Initial level to Grade 3 [5] as the quality standard. The Teacher Grading for the experiment is based on the score ranges used by the Trinity Examiners: Distinction 87-100; Merit 75-86; Pass 60-74 and below pass.

* 1. MIR approaches to onset/offset detection

This thesis aims to build on what current technologies can offer. Onset Detection shall be discussed focusing on the properties of the bass which is a Pitch Percussive instrument. The peak picking algorithm introduced by Bello [10] is one of the techniques used for measuring onsets. It requires experimentation to optimise 3 parameters. His subsequent paper on considers the energy in addition to phase for the onsets [11] and it is the equation that considers the local energy of a signal that gives us a technique to measure offsets.



*Figure 1: Local Energy Equation.*

A big question in this research is the following: when does the offset end?

This introduces the topic of signal perception in the human auditory system and this has been researched outside of the context of Music Education by Kopp-Scheinpflug [12] In the scope of this work, an attempt is made to consider the output of this research in gap-detection and the limits of human auditory temporal acuity which varies from 2/3 ms to 30 ms depending on the level of spectral disparity in the signal

As part of a trade study of current apps that use MIR approaches for onset detection , the performance assessment for singing in Yousician[[3]](#footnote-3) is very different to how bass (and other stringed instruments) are assessed. The piano roll format as used for vocal training gives better visual feedback to note durations and although no comprehensive test have been carried out on the this, on some songs it has been found to correctly assess note duration but for other songs in the Yousician curriculum (e.g. “Fire” by REM) it has been clearly shown not too work. The scope of this research is limited to doing non-realtime assessments, i.e. the assessment of a stem recording as opposed to a live performance, so front end displays are not the priority, but this example does raise awareness of format issue for displaying timing feedback.

In Bass guitar, we consider two components of onset: the attack part of the rapidly increasing amplitude envelope and the exponentially decaying part. Bello [13] has provided guideline for choosing the right method depending on the requirements. The Wavelet method places focus on precise time localisation, an important aspect of bass rhythm. There are other options that have a high computational cost ranging from complex-domain spectral difference to using training sets and statistical methods. Non-real-time assessments are not constrained by computational cost requirements.

* 1. Measuring duration

As mentioned earlier, experiments with Yousician have shown that duration feedback is missing for some tracks in Vocals Performance assessment, but it is missing for all the tracks for Bass guitar. e g. If you have a half note duration on the score and you play a quarter note duration, you will not be penalized. If you play 4 crotchets, 4 quavers, or 4 semiquavers, the scoring would be the same by all apps, even though there is a clear musical difference in each of the displayed three bars below



*Figure 2: 4 crotchets, 4 quavers and 4 semi-quavers.*

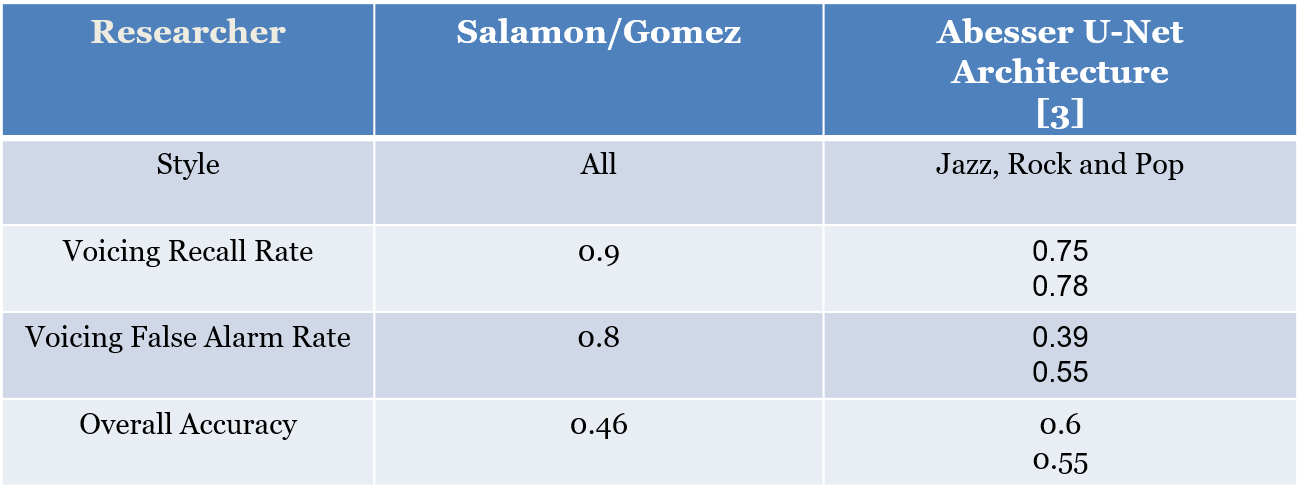
This shortcoming is really surprising when you consider that that Yousician is capable of correctly assessing left hand techniques covered in some of by Reboursiere [14] such as slides, Hammer-ons and Pull-offs. The challenge of durations measurement is that it is not an instantaneous measurement. It’s a bit like the SPECS Speeding camera system where you have start point and an end point. An offset (end point) cannot exist without a start point onset. For offsets, the “exit point” is really an open question for non-muted playing technique.

For offset measurement the algorithms under study focus on capturing the energy of the bass stem and determining the point of “drop off”, when the bass note is no longer audible. In any given song you may find a variation in the finger style techniques used for playing bars. A staccato style results in shortening the duration of the notes but also lengthening the inter-note interval. A legato style will result in the offset of a given note to run into the onset of the next note. Clearly a different strategy needs to be applied to each scenario. The case where the offset position of a given note exceeds the onset of the subsequent note is not considered. This would be like holding down the sustain pedal on a piano while playing another note or letting an open string play while you pluck another string.

* 1. Abessers Research

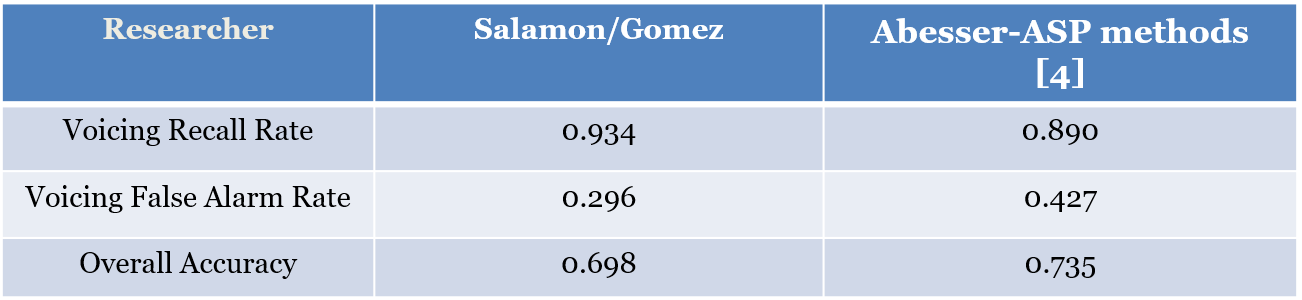
Jacob Abessers applied research in music information retrieval & machine learning / deep learning and audio signal processing is focussed primarily on bass. For timing measurements, the voicing classification (independent of which pitch it is ) is an important metric. The Salamon and E. Gómez (SG) [6] techniques mentioned earlier have good voicing recall metrics but Abessers-Müller Data-Driven [7] yields lower false alarms.

*Table 1: Abessers Data Driven algorithms vs Salamon/Gomez*



The experiment on bass tracks that Abesser [8] performed when comparing the Signal Processing methods with SG [6] also yielded better overall accuracy (considering pitch) but the on just the Voicing Recall and False alarm rate, the SG [6]techniques have higher accuracy.

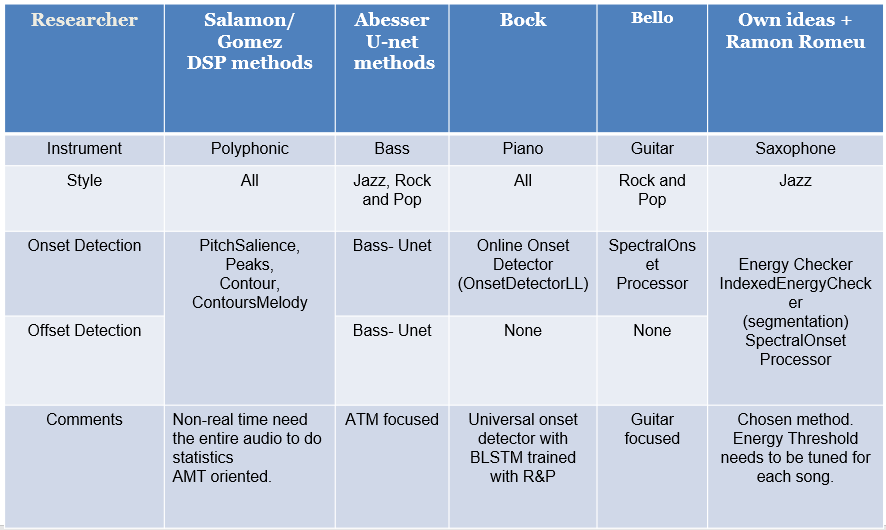
*Table 2: Abesser Signal Processing vs Salamon/Gomez*



* 1. Literature Research Summary

Onset detection is well researched topics with many off-the-shelf library functions available from Madmom[[4]](#footnote-4) and Essentia[[5]](#footnote-5) that can be readily applied to a bass stem. The following table summarizes the state of the art of algorithms used for onsets and offsets, many of which use these library functions.

*Table 3: State of the Art algorithms for Onset/Offset detection*



The first columns are based on pitch extraction and have been tested on polyphonic music and work quite well. The main function is the PitchMelodia[[6]](#footnote-6) function, which is designed to extract the predominant melody from polyphonic music and has 18 parameters which can be configured. PitchMelodia consists of 4 algorithms that are called in a chain: PitchSalienceFunction, PitchSalienceFunctionPeaks, PitchContours and PitchContoursMelody. Using these library functions for testsing offers more configuration flexibility than the combined PitchMelody function and that was the approach used in this Thesis.

The second column is based on a CNN (Convolutional Neural Network) Streamlined Encoder/Decoder Architecture for Melody Extraction. It has been tested on the following Dataset:   
- Real World Computing (RWC) [15]  
- MDB-bass-synth [16]

- Weimar Jazz Database (WJD) [17]   
It the implementation code for bassunet[[7]](#footnote-7) is open source. In the 2017 [8] paper is based on Onset and Offset detection using Fo Contour Tracking, Abesser used the following annotated Database, which is also used in this Thesis.  
- IDMT-SMT-BASS-SINGLE-TRACKS (Fraunhofer) [18]

The matlab/python code for calculating onsets and offsets is closed source but the Fo Tracking library code is available[[8]](#footnote-8) in the pymus libraries.

Columns 3 to 5 are all based on the Madmom[[9]](#footnote-9) libraries. The Online Onset Detector[[10]](#footnote-10) based on recurrent neural networks by Bock is a universal onset detector with BLSTM and was trained with music mixtures including R&P. The Spectral Onset Processor (SOP) implements several (up to 11) onset detection functions considering phase and energy information and was considered in the set of algorithms that were developed by Bello [11] based on Peak Picking. This algorithm was considered for the first experiment using the guitar focused pysimmusic [[11]](#footnote-11)tools (back end for Music Critic). This approach forms the core of one of the onset detection methods used in this thesis. The following peak picking formula forms the basis of the [11]



*Figure 3: Equation: Peak Picking formula*

The peak picking strategy used here is the equivalent of the one contained in the S/G Essentia algorithm PitchSalienceFunctionPeaks. Ramon Romeu[[12]](#footnote-12) developed a wrapper based on the above formula with the following values determined for C\_t, H and delta in sweep experiments

#Dynamic threshold

C\_t = 0.99

H = 100

delta = 0.1

din\_th = np.zeros(len(det\_function\_norm))

for m in range(H, len(det\_function\_norm)):

din\_th[m] = C\_t\*np.median(det\_function\_norm[m-H:m+H])+delta

*Figure 4: dynamic threshold setting*

The values given for C\_t, H, delta can be customised for different Dataset tracks. A sweeping method can then be sued to optimise the detection for a given musical piece.

The final column summaries the method developed that can return an offset for every detected onset, using a “sound island” approach.

* 1. Revision of Algorithm Evaluation for Music Pedagogical Purposes

In Trinity’s “exam guidance: marking “ section [5]33% of the Music Assessment is directed at Fluency, synchronisation with the backing track, security in notes and rhythm. This is probably the easiest to objectively define and benchmark with MIR solutions currently available. Music Critic[[13]](#footnote-13) is currently developed for Guitar at the MTG (Music Technology Group) and requires adaptation for bass to deal with the importance of timing and locking in with the drum pattern.

The Guitar version has an online version and works well for simple demos. Music Critic gives the student feedback at the end of a performance. It does not flash green or red on each individual note “on the fly” like in the previously mentioned edutainment apps. This leaves some room for discretion in terms of discriminating between the different bars and different notes played for a specific musical piece. Post-Analysis is also more appropriate in preparing the student for an exam and it leaves margin for customizing deviation in timing/tone particular to each song and/or a part. With a defined set of duration formula, an annotated dataset of onset and offset times, the next stage is to gather new annotated data by obtaining real student recording using a Student Portal in chapter 5.

The second part of the TCL R&P exam is on Technical Control. This is ability to control the instrument effectively, achieving the various technical demands of the song and sound quality. Of the three songs a student chooses, one song has to be “Technical Focus” (TF) which means more weight is given to this section (12 point instead of 9). This prompts, the question, how do we measure and gather data to provide useful feedback on instrument aspects? Is there a Dataset that we can use to train a model to identify good technical focus performances? Abesser addressed the issue of extracting instrument features [8] and to lay the foundation for building this research, the same dataset [18] is used in this thesis. Even though, the scope of the thesis is limited to timing and rhythm aspects, the TCL dataset shall be built in a way so that it can be expanded to consider all technical aspects of playing bass.

The Dataset prepared in this thesis consists of 8 Student recordings of 6 different TCL R&P syllabus and is discussed in Chapter 3. These recordings were graded to a Bass Teacher with a customised questionnaire based on the TCL examiners report of Pass, Merit or Excellent in each of the examination areas of Fluency & Security and Technical Control. With this dataset, a set of histograms of onset deviations in 4 different levels were developed using the analysis techniques applied by Eremenko et all [19].

* 1. Perception of Onsets and Offsets

Onset detection is huge research topic with many alternative formulas for calculation [13] however offset, particularly the perceptual aspects of them, have historically not been given the same level of attention. Perception of offset detection is dealt with very well by Kopp-Scheinpflug [12]in which they researched literature on neurons with sound-offset responses in the auditory system. In sound perception, the concept of “just noticeable difference” is important and what this means for offsets is that a sound will have to be of sufficient duration to allow an offset response. For music education and performance assessment, this means time window needs to be defined that specifies the minimum duration for detecting an onset and an offset. Typically, the onset window [20]has a time value of 50ms being cited as being the minimum time window of detecting onsets, but for Music Education purposes, this is too long and pending further measurements and observations, a time window of about 12ms would probably be more appropriate. A more rigorous approach to measuring onsets would distinguish between perceived and actual onsets, but this distinction is not considered in this Thesis.

* 1. Observations

There are a lot of musical aspects to consider in assessment, that are difficult to measure using transcription methods based on score. They are swing ratio, attack displacement (playing exactly, ahead or behind the quantised beat), to mention a few. To capture these characteristics, it is necessary to obtain ideal student performances at the excellent level and also some performances and intermediate and novice level. Ideally these could be obtained with from recording with a microphone capturing the bass performance separate from the backing track. However, if this dataset is limited, Source Separation techniques based on the Spleeter model [21] that have been successfully applied to the songs from TCL gradebooks can be applied to previously recorded performances.

The challenge in this project is to find the most suitable machine learning techniques to extract the most relevant parameters from the limited number of annotated student performances. Support vector machines to perform not only score transcription for bass but also the extraction of the plucking and expressive styles [8]AMT-domain Datasets such as the IDMT [18] consist of small single note extracts for the style, short 15 sec tracks for the score but Music Education requires larger audio segments and more widely ranging quality performances.

The area of Technical Control (TC) remains a challenge for machines to evaluate. Section 5.1 tests MusicCritic’s effectiveness of onset detection in the Bass, thus building on the work performed for the guitar in [19].

1. Datasets

This chapter explains the two Datasets that are used to evaluate algorithm accuracy

* 1. IDMT BASS SINGLE TRACK

The IDMT dataset [18]from Fraunhofer consists of 17 audio tracks with accompanying score and annotated onsets/offsets with various levels of complexity. Each score is accompanied by a WAV audio file and an XML file with various annotations including MIDI pitch, onset, offset and other instrument characteristics as shown below:

<event>

<pitch>36</pitch>

<onsetSec>2.4</onsetSec>

<offsetSec>2.5552</offsetSec>

<fretNumber>3</fretNumber>

<stringNumber>2</stringNumber>

<excitationStyle>FS</excitationStyle>

<expressionStyle>NO</expressionStyle>

<modulationFrequencyRange>0</modulationFrequencyRange>

<modulationFrequency>0</modulationFrequency>

</event>

<event>

<pitch>36</pitch>

<onsetSec>2.7</onsetSec>

<offsetSec>3</offsetSec>

<fretNumber>3</fretNumber>

<stringNumber>2</stringNumber>

<excitationStyle>FS</excitationStyle>

<expressionStyle>NO</expressionStyle>

<modulationFrequencyRange>0</modulationFrequencyRange>

<modulationFrequency>0</modulationFrequency>

</event>

*Figure 5: Extract from IDMT Dataset. File 002.xml*

The above annotations show that there is a clear difference in duration when you subtract offset from onset: 159ms for staccato and 300ms seconds for normal.

This Dataset is not used for Student performances. IDMT contains many tracks that have complexity exceeding that required for grading required for this thesis. It also has some very short notes where the plectrum is used for the style. The objective of this Dataset is to measure the effectiveness of the onset/offset algorithms. It is available under a creative commons licence.

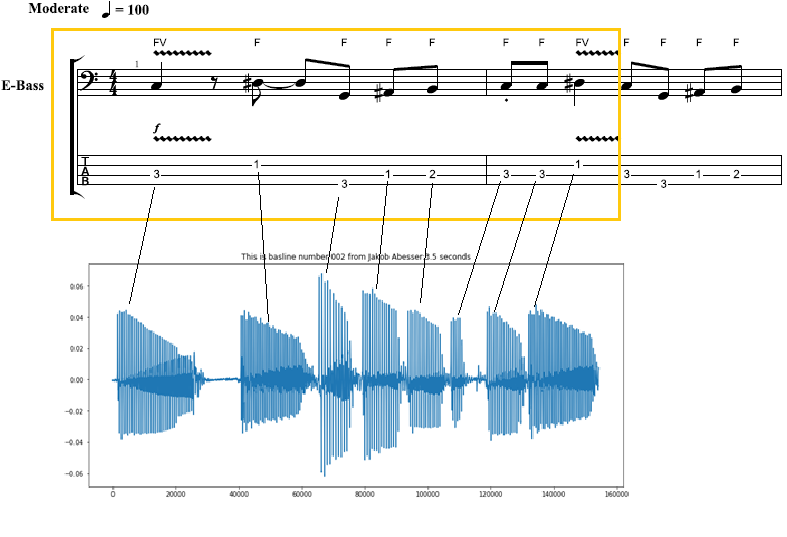
IDMT-Summary

IDMT (Fraunhofer) : 17 tracks

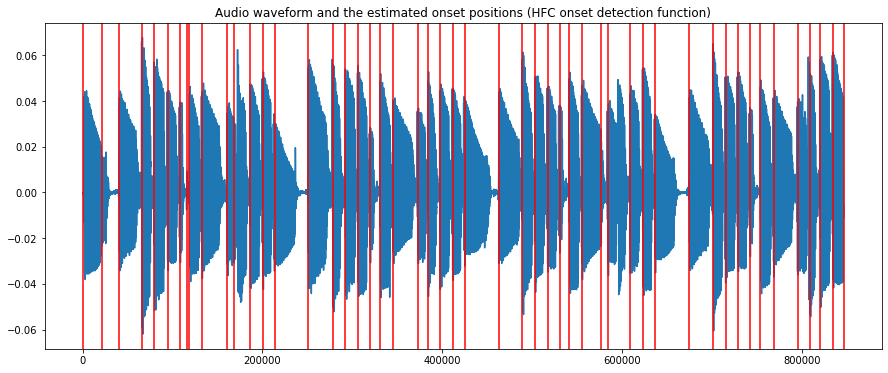
* Varying plucking techniques (Pick, Finger, Muted, Slap)
* PDFs of score and XML of parameters (onsets, offsets, pitch, fret number)
* Expression Style annotated
* 4 tracks within musical and timbral complexity of Trinity Grade 3

IDMT-Example on note length

The annotation of a musical note as “staccato” or “legato” can have a big impact on the intended duration. In the figure below you can see that the first C note in the 2nd bar is almost 1/3 the duration of the second C note.



*Figure 6: Note duration of Staccato notes*



*Figure 7: Audio 002.wav (IDMT DATASET) with onsets*

* 1. TCL Dataset

There are two key differences with the previous dataset. First, it is not publicly available and access to the audio and the PDFs is only granted on purchase of the materials from Trinity College London. Secondly, unlike the Fraunhofer stems, these tracks do not come with accompanying Onset/Offset annotations. It was necessary to make these manually using Sonic Visualizer.

For this research, an initial list of twelve songs were identified to cover as many syllabus topics that are necessary for grading. Finally, six songs were chosen from the Grade books. These six songs form the basis of the keyword for this Thesis “TCL dataset”.

*Table 4: Six chosen songs from TCL song list*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Song Nr | **Grade** | Song | **Artist (original)** | **Shortened Name** | **Seconds** |
| **0** | **0** | Yellow | Coldplay | yellow | 93 |
| **1** | **1** | Billie Jean | Michael Jackson | bjean | 55 |
| **2** | **1** | Just Looking | Stereophonics | just | 86 |
| 3 | **2** | Brown Eyed Girl | Van Morrison | brown | 64 |
| 4 | **3** | Roadrunner | Junior Walker and the All Stars | road | 82 |
| 5 | **3** | Walking on the Moon | The Police | wotm | 120 |
|  |  | TOTAL |  |  | 500 secs |

The shortened names in column 4 shall serve as references to the TCL songs in this thesis.

As part of a recent collaboration between the Music Technology Group (MTG) and Trinity College London (TCL), the separate stems of these recordings have been made available for the analysis purposes. The TCL recorded bass stem is to serve as the Ground Truth. The remaining stems are used to build up “minus 1” tracks for mixing the with the student performances. Trinity have also made available the Scores and Grade book data in the form of PDF publications and XML files (for loading into Musescore[[14]](#footnote-14) ). The description of how each song is graded is given before the Tablature is shown, forms an important part of the meta-data for each song.

Throughout the Thesis a Dataset was built up consisting of the following:

* Onset/offset annotations of ground truth (bass stems of TCL recordings)
* Multiple Student recordings of each of the six songs
* Mixes of the Student Recordings with the Minus 1 tracks
* Music Teacher Numeric and Descriptive grades allocated to each Mix.
* Algorithm generated data on the Student and Ground Truth stems.

Duration plays an important role in the TCL R&P criteria for assessing the Grade 1 song “Float On” [22] it says, “*the chorus features some sustained dotted minims, which should be held for their full length*.” Table 4, list the songs chosen for analysis in this Thesis and the shortened name/acronym to be used hereafter in the thesis.

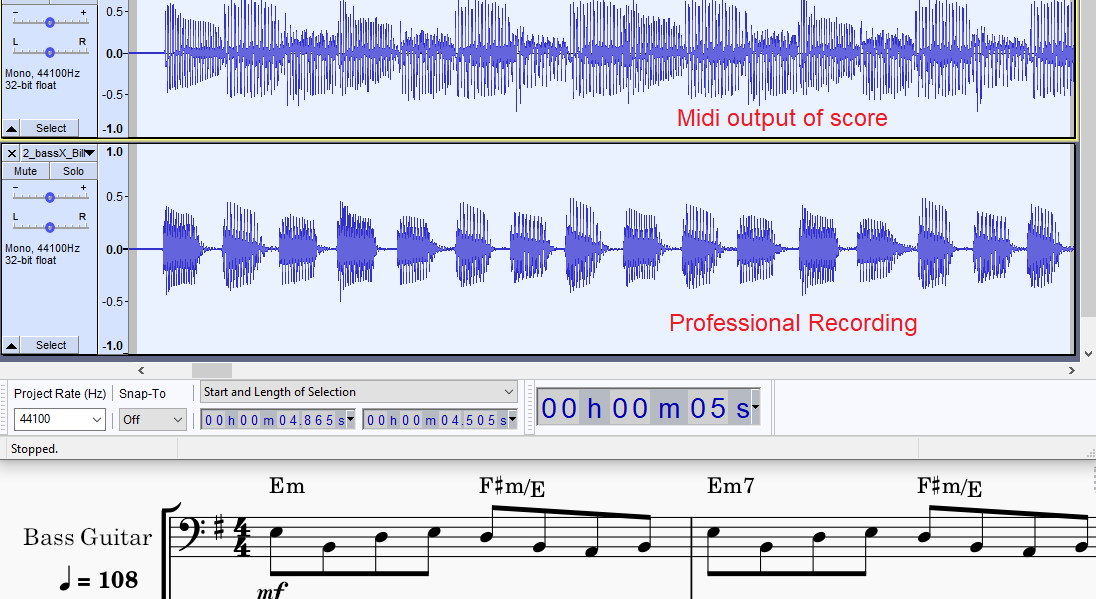
The Dataset includes XML files which provides rich information on the score note duration . There is also technical information related to the fret position and articulation information, e.g., the accent in Billie Jean

|  |
| --- |
| <notations>  <technical>  <fret>**2**</fret>  <string>**2**</string>  </technical>  </notations>  ..  <articulations>  <accent/>  </articulations> |

*Figure 8: Technical Information in XML files*

The Onsets and Offset were initially done by visual inspection in Sonic Visualizer[[15]](#footnote-15). This was improved by running the best performing onset and offset algorithms on the TCL GT Stems, removing the false positives and add the visually adding the missing notes. This resulted in a much more consistent set of annotations, so the mean and standard deviations of the onset/offset deviations of ground truths were close to zero.

The XML files can be viewed in Musescore and by removing the chord information and exporting of a WAV file, a “mechanical rendering” can provide a useful reference in testing the algorithms. It demonstrates the importance of the incorporating stylistic elements in a real performance especially where duration is concerned. It also indicates that there is information provided in the Song Description which is not reflected in the actual score. For “bjean” it is required to pluck with short/jerky movement in the verse/chorus, but no dotted notation is used to reflect this.



*Figure 9: Billie Jean: Midi Rendering vs Human recording (shorter notes)*

The first function of the TCL dataset is to validate the algorithms short listed state of the art algorithms after testing with the IDMT dataset. The second function is to provide a basis to measure student performances.

The TCL professional or “ground truth” stem is taken to represent a grade of 100% in timing and rhythm.

The following table summarizes the individual musical features of each of the tracks a described in the Grade Books [22].As stated in the introduction, the objective is to direct the numerical grading and verbal assessment to reflect as much as possible the rhythmic and timing qualities a Trinity Examiner is looking for. Ultimately the goal is to match what are human assessments to objective measurements of extracted audio features. In keeping with the importance of considering Technical Control of the instrument (Chapter ) , Songs 1-5 were chosen since they were marked “TF” song in the TCL syllabus. The sixth song (WOTM) was chosen because it had widely contrasted note duration features.

Table 5 summarizes the Technical Control parameters of the 6 chosen audio tracks. This table lays the basis for deciding on how to phrase the questions on the Grading Sheet given to the Bass teacher and also the description of the quality levels in each of the Multiple Choice answers.

*Table 5: Tech. Control Parameters for 6 TCL songs*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Song** | **Coord.** | **Syncop.** | **Repetition** | **Dynamics** | **Articulation** | **Note Len.** |
| yellow |  | chorus | No rushed feel |  |  | Play evenly (v) |
| bjean | intro | Accent just before chorus |  | -- | separate jerky quavers+ smooth, melodic material |  |
| just |  | Chorus: accented, syncopated motif.,  hard accent |  | Unexpected  subito p at bar 25. |  |  |
| brown |  |  |  |  |  | different note lengths and rests |
| road |  |  |  |  | Tenutu (underscore)  loud on beat 1 |  |
| wotm |  | syncopated  repeated notes |  |  |  | correct  separation |

In the experiment, truncation is applied to the songs, to facilitate grading a section of music once, for example “yellow” was truncated to remove the repeated verse, since no new musical features were introduced. The song “wotm” has been cut to only include the first 50%. The “wotm” song is symmetrical, the second half is a repeat of the first half. This opens the possibility to split the data and create more performances out of the student recordings, however for this Thesis this option was not performed. Technically speaking the second half of the song “wotm” has an ad-lib section on the bridge, but this was not observed by any student.

The WOTM verse has a long note duration with zero inter-note gaps and the bridge has the opposite shorter “reggae” notes with some noticeable internote gaps. The BPM of this song is 146 and for Billie Jean it is 108. Both require very different thresholds. The Grade1 “Billie Jean” song also manifests similar contrasting sections. Evaluations were set up to evaluate the quality of onset and offset measurement algorithms

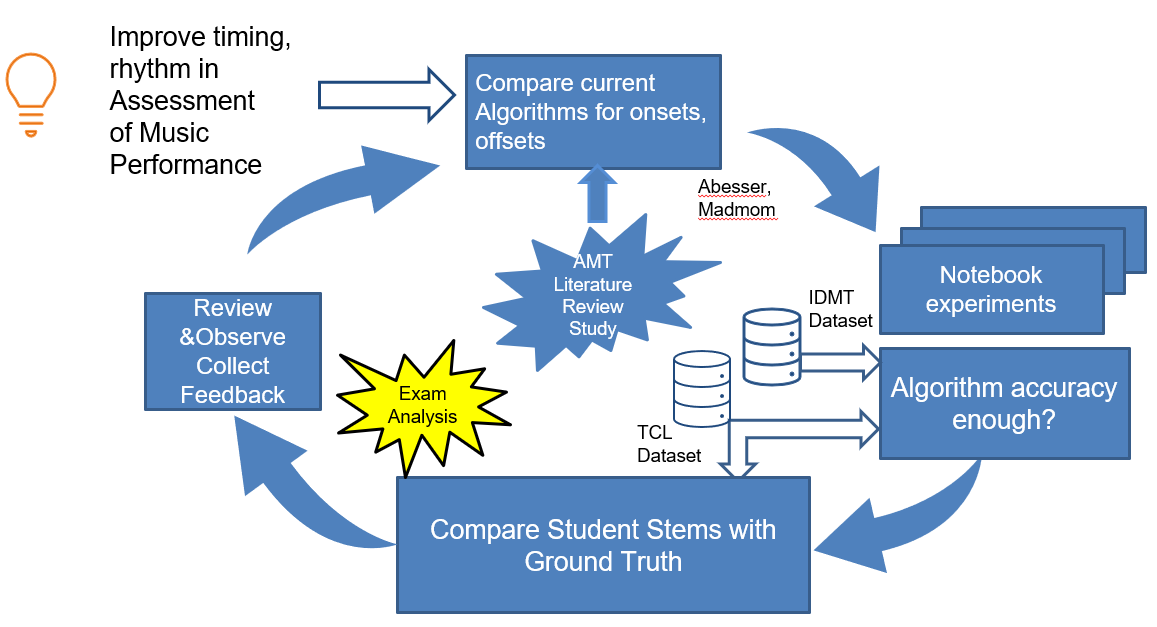
1. Methodology

Music Education demands more precise timing measurements than state of the art automatic music transcription. Onset detection for electric bass has different requirements for guitar in two key areas. First, as a rhythm section instrument the bass plays a key role in synchronizing with the drum pattern. Secondly, the bass does not require the handing of playing chords that the six-string guitar does. The accuracy levels for bass onset detection should exceed the accuracy achieved by playing guitar melodies. The main proposal of this thesis is to add duration measurement and to achieve this offset detection is required and this places a constraint on the choice of onset detection algorithm. Abesser used the standard MIR evaluation of 50ms for measuring onset accuracy in Guitar onsets and increased this to 200 ms for offsets due to difficulty in handling smoothly decreasing note envelopes [23].

* Benchmark algorithms for accuracy
* Testing SOTA algorithms to measure student performances
* Using Multi-variable Linear Regression to train new a model with teacher graded student performances
* Analysis of teacher comments in student performances and correlation with student grades.

There was no “straight run through” to get the best algorithm to grade all the Trinity songs equally well. A few iterations were required that involved modifying the algorithms to consider plucking style, developing a hybrid algorithm that would apply different threshold parameters to different songs and different offset strategies to different sections of a given song.

Even getting good PRF metrics for the professionally recorded ground truth it was a difficult task to produce student recordings that would return a precision value higher than 60%. Collecting Sound Quality grades taken by the teacher were useful in annotating these recording imperfections.



*Figure 10: Strategy to experiment, test to find best accuracy*

Fig. 8 illustrates the complexity of the research carried out. The accuracy of the algorithms needed continuous improvement in order to be meaningful for music education standards. Even after finding a reasonable PRF necessary, it was necessary to obtain bespoke arrangements for different TCL Dataset songs.

* 1. Algorithm Evaluation

In the algorithms introduced in the literature research in chapter 2 there are two strategies involved: “paired” and “non- paired”. Non paired are the classical approaches to capturing the onset without regard to monitoring the end of the onset. The paired approach is to measure start and stop time with the same audio frame. The non-paired approach refers to a method where the onset data on it own.

The 4 methods introduced in table 3 ,chapter 2 were tested alongside a fifth method to implement combined approach.

* Madmom Online Onset Measurement (Non-paired, Data driven)
* SOP: Spectral Onset Processor (Non- paired, ASP driven)
* AbesserUNet Algorithm (Paired, Data driven)
* SG: Salamon and Gomez (Paired, ASP driven)
* IEC: IndexedEnergyChecker (Paired ASP driven)

Precision and Recall, and F-measures shall be made on Onsets and Durations.  
Duration = Onset-Offset

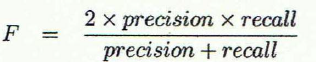
• **Precision**: exactness – How often did the algorithm incorrectly detect an onset/duration? (imposter notes)



• **Recall:** completeness – How often did the algorithm fail to detect an onset/duration? (missing notes)



• ***F* measure (*F1*** or ***F*-score)**: harmonic mean of precision and recall



Deviation measurements shall be made on the difference between Ground Truth and Measured Value. From this Histograms are derived following methodology used in [19]  
the Mean Absolute Error and Standard Deviation are calculated.We are concerned with two key measurements criteria:

Attack (for onset detection) and Release /Decay (for offset detection).

The attack is the most noticeable for performance assessment. Since for bass we are discounting chords, we do not have to consider multiple attack times, and we will consider the small difference between the actual attack and the perceived attack to be negligible. The resulting PRF for onsets for each of IDMT with a 20ms evaluation window averaged over 4 songs that are most like the TCL Dataset.

*Table 6: Benchmark results of Algorithms IDMT DS*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Researcher** | **Simple Energy Checker** | **Indexed Energy Checker** | **Bock (SOP)** | **Salamon/Gomez** | **Abesser U-net** |
| Precision | 0.425 | 1.0 | 0.894 | 0.933 | 0.67 |
| Recall | 0.518 | 0.649 | 1.0 | 0.927 | 0.797 |
| F-measure | 0.465 | 0.783 | 0.943 | 0.93 | 0.728 |

Although PRF measures were taken for the first two columns, they are not included here since the main driver of accuracy is the onset detection, from which dependent offsets are derived.

* 1. Indexed Energy Checker ( IEC)

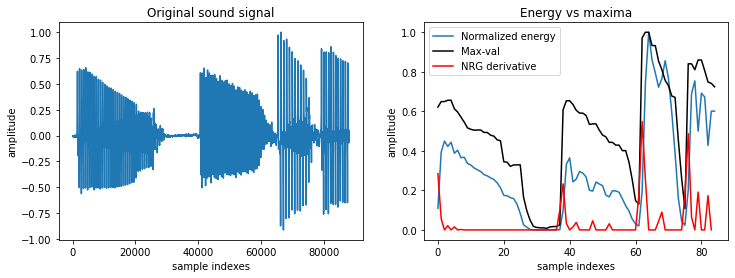
Originally a simple Energy Checker function called “calculateOffsetOnset()” was developed to capture RMS band values of an input signal. Subsequently those RMS band values would then be checked in sequence to see if they dropped below a certain threshold, in which case they would be added to a new offset array and a flag would be set to indicate an offset was detected. The flag would have to be cleared before threshold drop off detection could start again.

index= 0  
array\_of\_time\_offsets= []  
flag = False  
last\_index=0  
while index < len(rms\_bands):  
 if (abs(rms\_bands[index])<threshold) and flag == False and (index!=0):  
 *# Skip very first* array\_of\_time\_offsets.append(index)  
 flag = True  
 last\_index=index  
 index+=1  
 *#We set flag back to false after determined time period* increment\_factor= int(hopSize/hopSizeScaleFactor)  
 if index > last\_index+increment\_factor:  
 flag = False

*Figure 11: Original Energy Checker for offset detection*

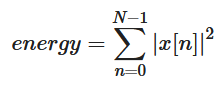
The main problem with this algorithm was that it did not calculate corresponding onsets within the same frame. The Onsets were calculated separately using standard HFC method, but these values were independent of the offset values. Too much effort would be needed to align them and fix this “un-paired” problem.

The IndexedEnergyChecker algorithm pairs the onsets and offsets, based on an Energy Threshold. It is based on the concept of a Sound Island. Depending on the energy threshold level set, it decides on how to split the wave boundaries. The splitting is represented by set of start and stop indices representing onset and offset. This threshold parameter was hard coded to 0.05 in his experiments with the saxophone. For the TCL Dataset, it was configured uniquely for each song.



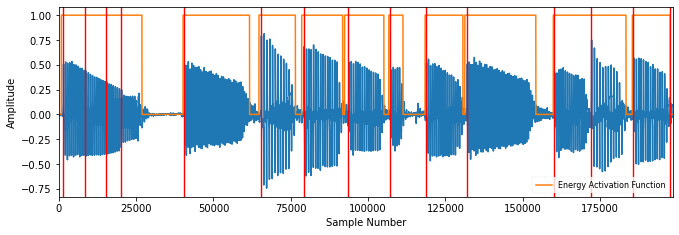
*Figure 12: Normalized Energy of Audio 002.wav sample (IDMT DATASET)*

Energy is calculated using the normalise energy function



*Figure 13: Normalized Energy Function*

The returned parameter “split\_decision\_func” function is an array of 1s and 0s that can be plotted as an overlay to the sound wave to give a graphical view of start and stop times of each voiced section



*Figure 14: Sound Islands of Audio 002.wav (IDMT). FS = 1024, HS = 512.*

This algorithm finds the best matching pairs with the following criteria  
 - distance between elements is no greater than matching\_window\_size

- sum of all distances is minimized

* 1. Accuracy Metrics

To calculate Accuracy measures for SOP and IEC algorithms, PRFs were captured for onsets for each of the TCL bass recordings (stems) with a 20ms evaluation window compare both IEC with SOP

*Table* *7: Accuracy Metrics for 6 TCL ground truths (combined vs non combined)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Song | P | | R | | F | |
| IEC | SOP | IEC | SOP | IEC | SOP |
| yellow | 0.939 | 0.98 | 0.942 | 1.0 | 0.94 | 0.99 |
| bjean | 0.949 | 0.983 | 0.969 | 0.997 | 0.958 | 0.99 |
| just | 0.682 | 0.9 | 0.828 | 0.869 | 0.748 | 0.884 |
| brown | 0.904 | 0.806 | 0.822 | 0.852 | 0.861 | 0.828 |
| road | 0.837 | 0.847 | 0.839 | 0.804 | 0.838 | 0.825 |
| wotm | 0.732 | 0.952 | 0.975 | 1.0 | 0.836 | 0.975 |

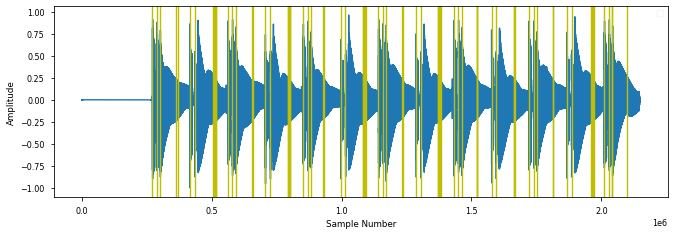
*Table 8: Accuracy Metrics for Billie Jean Hybrid Algorithm*

|  |  |  |  |
| --- | --- | --- | --- |
| Song | P | R | F |
| Hybrid | Hybrid | Hybrid |
| bjean | 0.9893 | 0.97555 | 0.9824 |

*Table 9: Selected comparison: IEC vs SG*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Song | P | | R | | F | |
| IEC | SG | IEC | SG | IEC | SG |
| just | 0.682 | 0.78 | 0.828 | 0.745 | 0.748 | 0.761 |
| wotm | 0.732 | 0.794 | 0.975 | 0.653 | 0.836 | 0.716 |

The worst performing IndexedEnergyChecker (IEC) results were then benchmarked against the SG [6] results. For the first half of the track, “wotm” , using the IEC with Threshold set to 0.06, yielded a lot of false onsets, hence the low precision in the PRF reading: P= 0.514, R = 0.949 F = 0.667

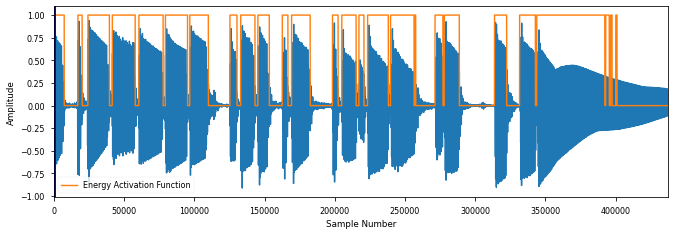


*Figure 15: IEC onsets for WOTM verse, P = 0.514.*

For the bridge section of the same track “wotm”, the IndexedEnergyChecker performed better, but again the PRF accuracies improved when considering the last long note.

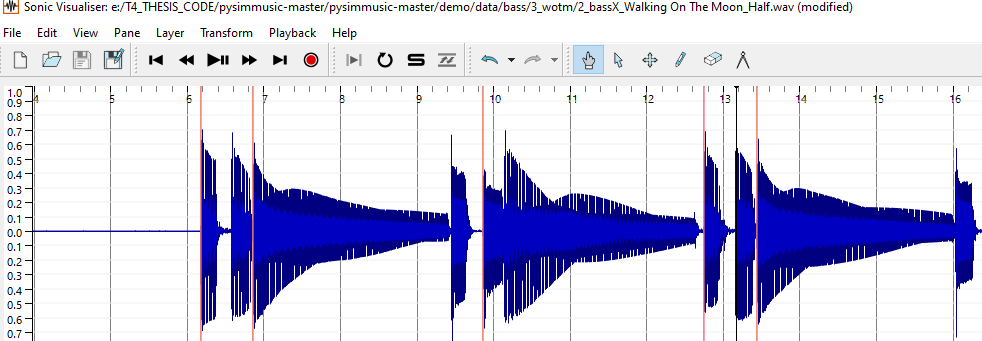
0.937 0.949 0.943 (With last note)

0.961 0.948 0.954 (Without last note)

****

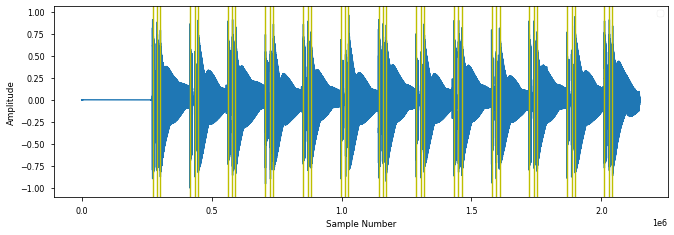
*Figure 16: Calculated IEC Sound Islands for WOTM bridge, P = 0.943*

This song has the longest note length of all the tracks. With (thresh = 0.06, we still see a lot of hysteresis on the last sustained note in Figure 10. The IEC algorithm does not work well in the verse, a lot of onsets are missed.



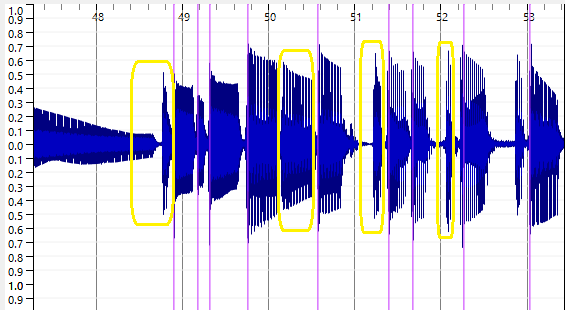
*Figure 17: Calculated IEC Onset positions for WOTM verse*

In contrast the SG [6] algorithm performed a lot better in the “wotm” verse with PRF measures of P = 0.974, R = 1.0 and F = 0.987.



*Figure 18: PitchMelodia derived Onsets for WOTM verse , P = 0.97*

But SG [6] missed a lot of onsets in the bridge section of “wotm”



*Figure 19: PitchMelodia derived Onsets for WOTM bridge*

The same behaviour was observed for “Just Looking” which also had long sustained notes. There are a lot of “misses” for the short notes in SG[5] method in the bridge as seen in fig 12 marked in yellow.

It is important to clarify some local definition of “mute”, that it does not relate to the concept of the left-hand muting technique described by Abesser [8]. Normally muting strings on the bass means damping them with left or right hand to short sounds with rapid decays. The term “soft mute” is introduced here to signify a gap of at least 10-20 milliseconds. When there is no gap., there is no offset. The offset in this case is equal to the next onset.

The annotation file has 3 columns: onset, muted and offset.

Muted is either “Y” or “N”. When it is Y this means an offset exists. When it is ‘N’, the current offset ground truth value is equal to the next onset ground truth value.

* 1. Deviation Metrics

The following plots illustrate the deviation statistics for particular songs for onsets, considering the SOP method and the IEC for onsets, considering only the IEC for offsets.

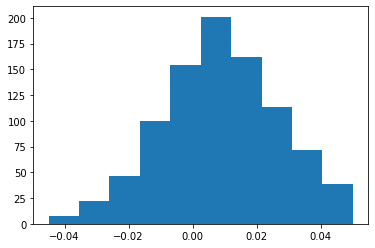
myOnsetEnergyCheckerThreshold = 0.05

matching\_window\_size  = 0.05 # 50 ms MIREX reference

Average precision: 0.947217, recall: 0.632121, f-measure: 0.750529

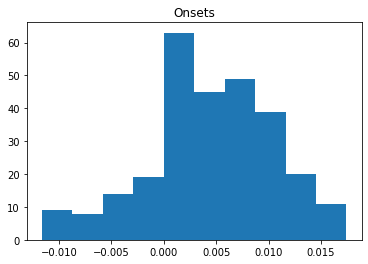
|  |  |
| --- | --- |
|  |  |
| Onsets  Mean: 0.003599  Deviation from zero: 0.019611 | Offsets  Mean: 0.109678  Deviation from zero: 0.282235 |

*Figure 20: IDMT Dataset GT Onset + Duration deviations ( IEC ) algorithm*

**

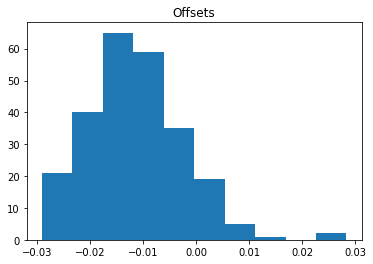
*Figure 21: IDMT Dataset GT Onset deviations ( SOP) algorithm*

SOP:  
Average precision: 0.921253, recall: 0.966997, f-measure: 0.942816



*Figure 22: Billie Jean GT onsets deviation (*Window 20ms)

Onsets  
ABS Mean: 0.005915, Mean: 0.004029,   
Dev. from 0: 0.007258



*Figure 23: Billie Jean GT offsets deviation (? Algorithm)*

misses 39 Percentage Miss 13.636363636363637

Window 30ms

Offsets ABS Mean: 0.012817, Mean: -0.011807, Dev. from 0: 0.014876

The same deviation metrics were used alongside PRF metrics to assess the student performances.

* 1. Correlation with Student Grades Metrics

The next chapter describes the main experiment of the thesis in which real teacher grades are obtained from real student recordings with the aim of allow us to predict grades on a set of test recordings.

The method used is linear regression and will consider as X inputs the PRF results and the mean absolute error and standard deviation of the onset/offset deviations. The Y values shall be the specific grades the teacher assigns for metronome accuracy, for note length and the specific demands that each son requires for Technical Control. There has been a slightly higher deviation noted in the offsets. This is to be expected since there is no clear end point for long sustained notes.

1. Experiment

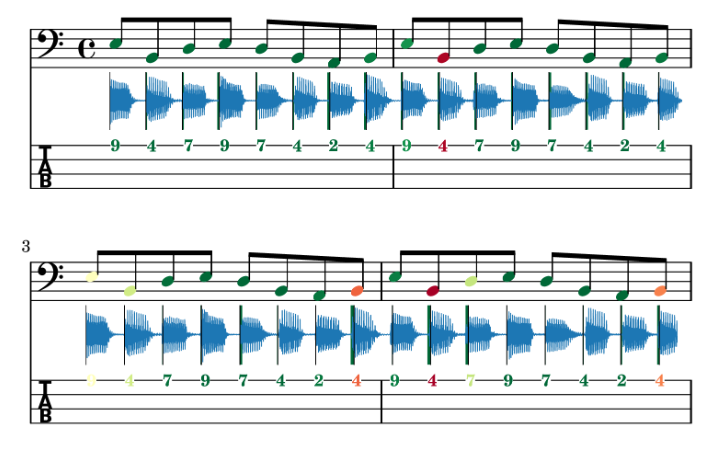
In this chapter, two experiments are described. The first one is an end-to-end test on the chosen dataset against the current State of the Art methods used in the pysimmusic tools [24]. The second experiment consisted of gathering and collecting data for the missing part of the TCL Dataset: onset/offset annotations, the student stem recordings, the mixes, the grades & comments on the mixes

* 1. Pysimmusic End-to -End test

In the pysimmusic program [24]a JSON file is used to mark the overall duration of the songs the beat locations in time. The LY file marks the note pitch, its length and its location in the overall on pattern. To perform the End to End test with pysimmusic software for Billie Jean, the following steps were executed:

* Calculation of Beat positions using Madmom.
* JSON, Lillypond file preparation for Billie Jean.
* Creation of New “Minus-1” with Trinity Stems.

The diagram below shows the graphical output m of the first experiment



*Figure 24: End to End test of Music Critic with Billie Jean GT*

The alignment of the ground truth stem was not aligned 100%. One possible explanation might be the customisation of the wrapper of the Onset detection function to consider multiple onsets that can occur with guitar chords. The outcome of this experiment is that the Onset detection methods as used currently for guitar need to be improved.

* 1. Student Recording Portal

The second part of building the TCL Dataset involved a large campaign to find eight students to do at least one recording of each of the six chosen bass tracks. The name “Bass Critic” was given to the Student Portal” which was implemented as a Google Form containing instructions and links to each of the six songs to perform the live recording with a backing track. The Student presses the play button, listens and plays along with the option to listen back on the stem before pressing “Submit”. Details of this Portal can be found in the appendix. Good headphones, that are capable of isolating the backing track from the microphone are required. A direct audio interface for the Bass is preferred instead of relying on positioning the microphone close to the bass amp.

* + 1. Latency Test

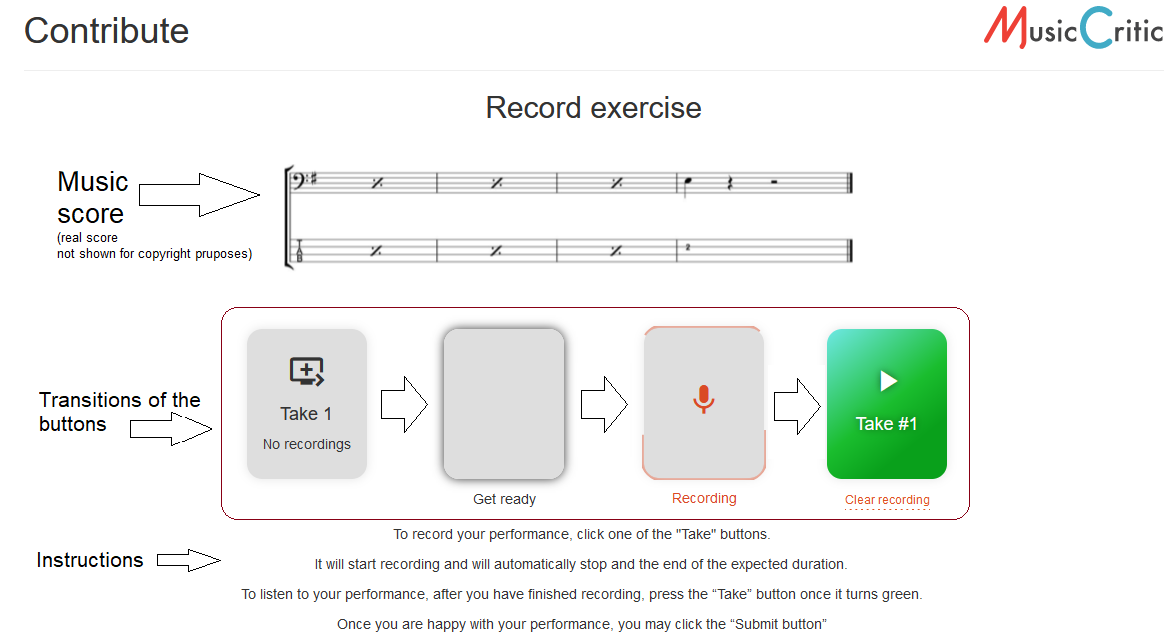
A pre-requisite for using Music Critic is to do the Latency Test. This involves placing the microphone close to the headphone speaker while recording the Click track. This allows to calculate a latency value which is then stored in the submissions.json file.

Ideally the microphone should go through the audio interface for the latency test. However, the hardware adapter components were not available at the time so the regular headset connection with the laptop was used for microphone input for latency test.

As with the first experiment with Billie Jean in section 5.1, the “Minus 1” tracks for the remaining five were created from the individual stems.

In the portal the student is requested to provide a description of hardware (external or internal microphone, Audio Interface), soundcard and driver (e.g. Realtek Audio), Browser (e.g. Chrome) and Operating System (Windows/ Mac / Linux)

Refer to Appendix full details on Student Instructions. The middle section of the diagram below illustrates the transitions on the web portal for capturing a student recording.



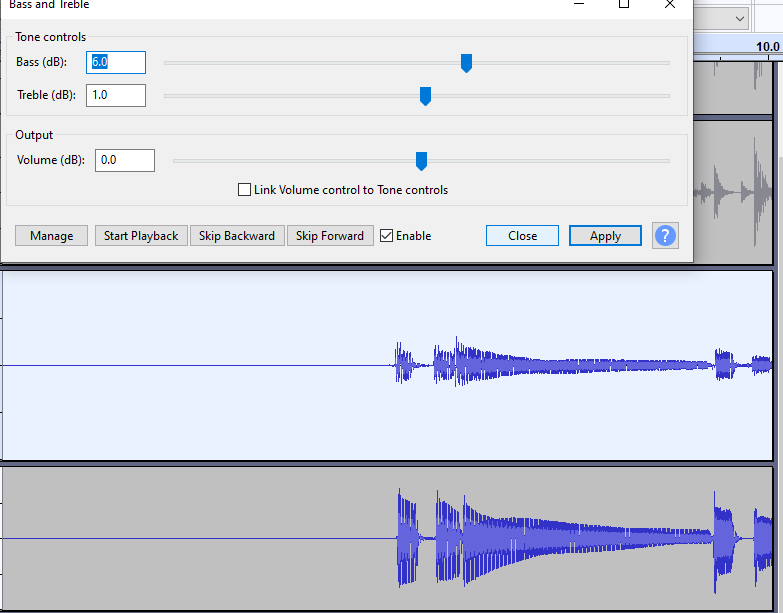
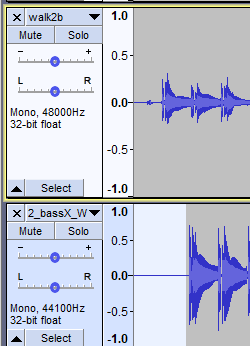
*Figure 25: Workflow on Student Portal*

After removing the latency from the recordings, it was noticed that the json calculated value still did not align with the first onset. As a result, the student recording stems were aligned manually with the first onset. This assumption was reasonable since it reflected the reality of the “best effort approach” [[16]](#footnote-16)of the student recording. Apart from aligning the student stem like this, some amplitude boost was given to equalize the loudness of the bass stem with the backing track. This was achieved using as greater Signal Boost ( Audacity 2.4) to increase the volume, to be in line amplitude of the Minus-1 track. A copy of the post processed student stem was then stored for analysis and also copy mixed with the minus 1 track to be provided to the teacher for grading.

In the initial recordings, the microphone on the headset was used, there was some noticeable background noise which adversely affected the teachers sound quality grading. However, the algorithms were still capable of onset detection for noisy stems. The steps involved in going from having a bass stem recording on a server, to a clean bass stem ready for analysis and a clean mix ready for grading are quite elaborate.

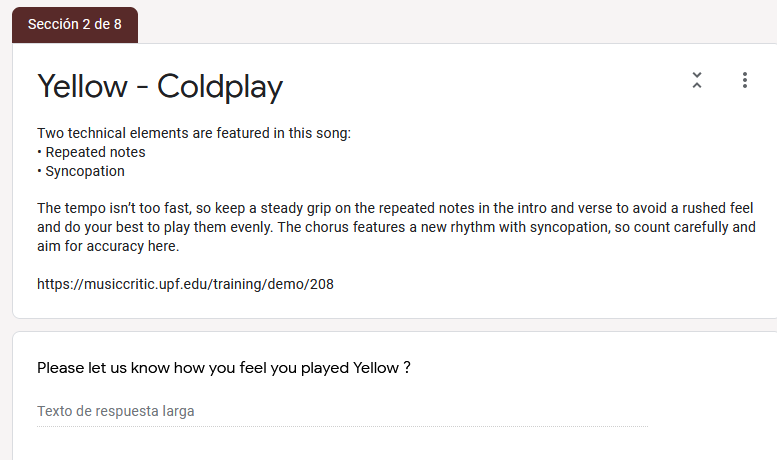
|  |
| --- |
| Steps:   1. Download the student recording from the server. IT has a name like this: e-e-g- "submissions/1805\_52a4886b326c4301b2760c8df6404c96.wav" 2. Rename it to the Student name. 3. Check that the playback rate is 44100Hs If its 48000Hz, left click on the audio file in Audacity and choose the rate 44100Hz, make sure it is also this rate in the Project settings, then go to Tracks menu and choose “Resample”. 4. Import Isolated Student Stem to Audacity 5. Import Ground Truth and make a split track to Mono 6. Zoom in on initial onsets and align them manually. Align the first onsets. 7. Boost the Bass 6db and the Treble 1db and after words add another 1edb to align amplitude with stem. 8. Add other tracks to audacity playback and check synchronisation and volume mix. 9. Boost the bass volume so you can hear it clearly. You may need to attenuate the other tracks, particularly vocals. 10. When you are happy with the mix so that you can grade the bass as a teacher export as WAV file. 11. Remove the other tracks and export the student stem also as a WAV file. It is recommended to have a good naming system to distinguish Bass stems from mixes.   Use the “\_m” suffix to signify a mix (wotm1\_m.wav)and leave bass stem with student name (e.g. wotm1.wav)   1. Don’t wait too long before you having the mixes graded |

Any future development that would require the collection of recordings on a large scale would require automating some or preferably all these steps.



*Figure 26: Processing a student stem in Audacity*

Here is an example of how the link to student recording looks.



*Figure 27: Technical instruction and link to recording portal*

5.2.2 Criteria for grading

For each song there were grade given each of the following categories:  
Onset, Duration, Technical Focus and Sound Quality

Technical Focus (TF) is song dependent, e.g for “bjean” the main TF is “Articulation and Coordination”. Exam Grades are allocated on a decreasing scale as shown in the box below starting from Excellent (100%) and decreasing “Unreliable” 37%

|  |
| --- |
| Fluency and Security  Trinity classifies according to the following scale for Fluency, Synchronisation & Security  -----------------------------------------------------  Excellent sense of fluency and synchronisation (100%) Very good of fluency, synchronisation with only momentary lapses. (88%) Good sense of fluency and synchronisation though with occasional lapses. (80%) Generally reliable level of fluency and synchronisation though with some lapses.(63%) Unreliable fluency, synchronisation (37%)  We want to focus on two aspects of fluency and that is on the 2 key timing aspects:  (i) Note Onset (hitting at the right time) (ii) Note Duration (holding it for correct length)  Please Note, the Song is organised as follows. (Ignore the NO BASS INTRO)  Part 1: Verse. Bars 5-16 Part 2: Chorus. Bars 17-24 Part 3: Bridge. Bars 25-32  Please refer to particular sections or bars of the song when making comments. |

*Figure 28: Extract of guide given to teacher for grading*

These percentage points (88,80,63,37) were chosen to fit within the ranges of the 4 categories of the TCL syllabus (excellent, merit, pass, below pass)

A Song length limit was set to reduce load on Bass Teacher and the Students. Teacher Grading was done on 89% of recordings

* 1. Grading

The Onset and Offset Grades are common to all songs. There are variations in the content and number of Technical Control related questions. The table below summarizes the topics that the different questions cover

*Table 10: Technical Control Topics (TODO refer to other table related)*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Q1 | Q2 | Q3 |
| Yellow | Repeated Notes | Syncopation | Sound Quality |
| Bjean | Articulation | Dynamics | Sound Quality |
| Just | Accented syncopation | Dynamics | Sound Quality |
| Brown | Groove | Sound Quality | ----- |
| Road | Syncopation | Articulation | Sound Quality |
| WOTM | Syncopation | Sound Quality | ----- |

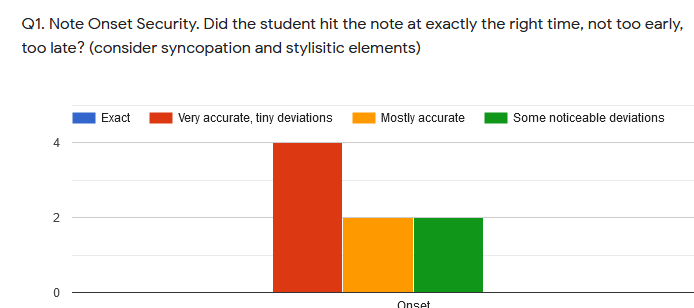
The original “bjean” Google Form has Q2 covering “Sound Quality” and Q3 covering Dynamics. In order for the parsing programs to work these columns had to be switched to read Q2 -Dynamics and Q3 Sound Quality as shown in table. Brown and WOTM

The worst performing EnergyChecker results were then benchmarked against the SG [6]results

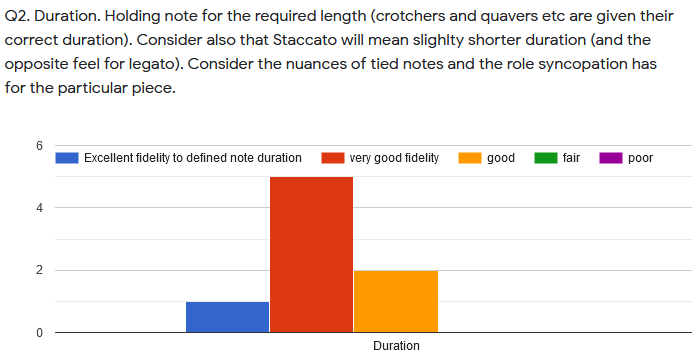
*Table 11: Comparison IEC vs SG*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Song | P | | R | | F | |
| IEC | SG | IEC | SG | IEC | SG |
| Just Looking | 0.682 | 0.78 | 0.828 | 0.745 | 0.748 | 0.761 |
| Walking on the Moon | 0.732 | 0.794 | 0.975 | 0.653 | 0.836 | 0.716 |

As mentioned in the State of the art the TCL R&P exam serves as the syllabus that we will use as a guide for teacher grading. Only the timing criteria in the fluency section is considered, splitting the evaluation to match onset and duration accuracy.



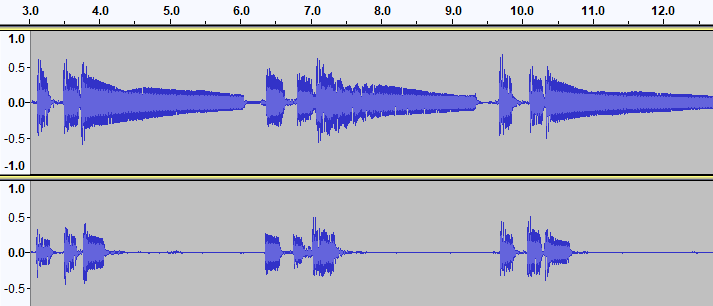
*Figure 29: Grading Histogram Onsets: Billie Jean*



*Figure 30: Grading Histogram Offsets: Billie Jean*

Five of the six songs are “technical focus” songs and the grading here is considered to add more insight into the rhythmic and timing aspects of the students’ performance and perhaps insight into volume handing and dynamics.

Three comment sections and have been added to emulate the comments that are given:



*Muy buena "note duration". Buena diferenciaciòn entre las notas largas de la "intro“ y el "verse" y las cortas del "bridge".*

Duration: **5 (Excellent)**

Overall: **4**

*En esta cancion las notas deben ser mas largas al prinicipio y màs cortas en el "bridge".*

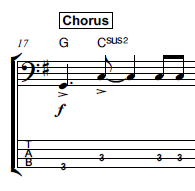
**Duration: 2 (Fair)  
Overall: 2**

*Figure 31: Contrasting comments of WOTM performances*

In TCL exams [5]comments alongside grades are obligatory, and applying this principle to the experiment, meant we have more rationale behind why a particular recording was graded a certain way. The tagging is also another means of securing the integrity of the recordings by appending comments to the tags field.

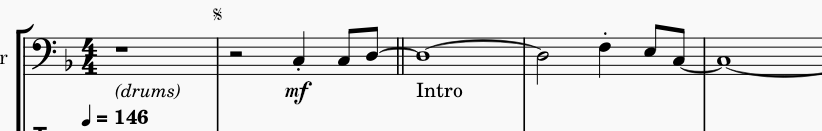
An additional global grade was added to help do consistency checks in the combination of the other grades allocated.

As mentioned in the State of the Art, it is difficult to map technical focus skills such as syncopation, dynamics to numeric measurements in audio features. The TCL R&P template is the chosen benchmark reference for performance assessment, so this requires the collection of Technical Control Grades. This also requires that each song has customized assessment policy depending on the Technical Control parameters as summarized in table X.Y in chapter 3. Even though these T.F teacher grades were collected, without knowledge to direct mapping to an extracted audio feature that we can offer some suggestions. But first it is important to understand its meaning in the musical context of a particular song. In Just Looking the score shows that the emphasis is off the beat in the chorus, i.e. on the “and” beat after 2, when counting 1+ 2**+** 3+4.



*Figure 32: Syncopation example Just Looking*

To measure its effectiveness, the amplitude of the G and first C note would have to be greater than the last two C notes. One way would be to add a field int the Rhythm CSV file that indicates Syncopation and this could flag the onset detection algorithm to check for greater energy. The same musical property occurs in the chorus of “Yellow”. For WOTM the syncopation manifests itself in the audio, with the first bass note occurring half a beat after where it is indicated in the score.



*Figure 33: Syncopation example: WOTM*

The dot is used to indicate syncopation and it also rendered to have half the duration of a regular crotchet not only in the ground truth audio but also in the WAV file generated from the XML file in Musescore. So, the syncopation is characterised by both the note position and note length both of which are annotated in the ground truth onsets and offsets. Therefore, the Syncopation Technical Focus grade should correlate with the combination of onset and offset measurements.

So, in the end how was it possible to make good use of the Technical Control (e.g. Syncopation or Dynamics) and Sound Quality Information? This question is dealt with on a song-by-song basis in the Results chapter.

1. Results

The student recordings were analyzed for their onsets and offset measurements and compared against the ground truth stems. Linear Regression Machine Learning algorithms were used to train the models with multiple variables. The Cross Validation strategy was that 30% of the total cases are used as Test Cases. The X inputs were the Precision, Recall and F-Measure (PRF for short) for both predicting both Onset and Duration Grades.

*Table 12: 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 | 76.50  68.85  49.50  49.50  76.50 | 68.763970  55.128501  42.868460  63.165241  78.933811 | 76.5  81.0  63.0  36.0  76.5 | 64.39  72.57  67.02  64.31  74.54 |
| just | 79.19  72.00  90.00  56.70 | 72.854928  73.866432  90.70  79.140241 | 90.00  79.2  90.00  79.2 | 78.19  78.90  92.32  83.46 |
| Brown | 79.2  79.  72? | 77.53  81.1  77.53? | 72  90  90 | 79.56  86.55  79.56 |
| Road | 79.2  79.2  72 | 76.87 79.54  92.96 | 79.2  90  79.2 | 94.5  86  74.6 |
| Wotm | 72  79.2  72  72 | 77  69.7  69.07  65.68 | 72.0  79.2  79.2  72 | 82.99  74.50  81.06  75.15 |

*Table 13: 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 | 8.62 | 12.25 | - | 6.54 | 8.26 |  |
| just | 11 | SOP | 7.84 | 11.7 |  | 4.67 | 6.38 | Mean |
| brown | 8 | IEC | 3 | 3.5 |  | 7.14 | 7.7 | - |
| road | 8 | IEC | 7.87 | 12.17 | Mean | 7.95 | 9.5 | Mean |
| wotm | 11 | SOP/IEC | 5.92 | 6.38 |  | 5.173 | 6.249 | Mean |

Table 17 gives an overview of how the different songs responded to predicting the main grades: Onset and Duration. The additional inputs are the Mean or Absolute Mean and Standard Deviation of the Onset/Offset Deviations.

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. The remainder of this chapter discusses the relationship between the predicted grades and the nuances and musical properties of the audio tracks. 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.

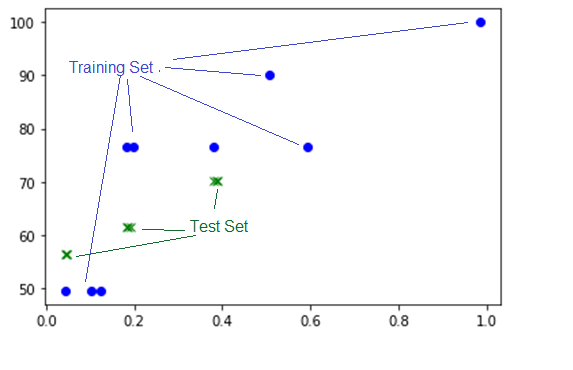
The analysis core results are contained in the “StudentStatistics\_<song>.csv file and each student performance has an associated Deviations file for Onsets and Offsets

* 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 the SOP algorithm. While the PRF score of the Ground Truth was close to 100% the highest graded Student 5 only scored a 50% precision while 2nd highest graded Student 4 for a P.R.F score with a 59% precision.

*Table 14: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 34: Plot of Onset Grades vs Precision*

Green plot: Test Set  
Blue Dots: Training Set

One explanation for the higher grade given by the teacher for Student 5 was because it was a different student with different instrument, with an overall superior sound quality.

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 was an outlier. Apart from a slight lack of consistency in hitting the A 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 Student 9,10,11 have been made to allow for future validation checks.

* 1. Billie Jean

The song was divided into three sections: 1st verse (Muted), Bridge (Non-Muted), Chorus (Muted).

The M.A.E increased to 11% when including 'Onset Mean', 21%, when including ‘Absolute Mean’ and 11% and when including Onset Standard Deviation. Overfitting can be a problem with five or more input variables. Low PRF results in a lower deviation count because “bad onsets” are filtered out. If you have two student recordings with very similar Precision, Recall , F Measure Values, then you can do a fair comparison of 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 is included Although relatively speaking, the Mean Absolute Errors, compare well against the other songs, but 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 have very low PRF scores. However, if these two rows are removed and the Linear Prediction of the Duration Grade is applied on the reduced set , the MAE jumps to 63%, but removing the Duration Standard Deviation would then bring it back down to 14%. Billie Jean is the song with most Student recordings; however they are more divergent in quality than other songs.

*Table 15: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 (which is give out of 5) does not fit well with the prediction for test 3. Figure 33 shows 3 over estimated outliers and two underestimated outliers.

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

*Figure 35: Plot of Precision (X) vs Onset Grades (Y) Billie Jean*

Lets examine the 50% grades with precision value above 0.5. Listening back on the stem and looking at the waveform there are no major differences with the ground truth, but the problem was the synchronization with the bass drum. The teachers comment was

*Although the bass follows the song quite well, there is a tendency to play behind the "beat". Try to match the bass and kick drum.*

So this outlier could be explained by making a bad mix. 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. Nevertheless, the teacher overlooked this and gave a high grade. Subsequently this grade was further scaled down by 90% to compensate this score deviation.

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.

|  |  |
| --- | --- |
| High Grade: Student 14 157 deviations | Worst Grade: Student 10  62 deviations |
|  |  |

*Figure 36:Best vs Worst bjean 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 a very poor PRF 0.1). 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.  
TDO : Point out Student 6 and 4., look at comments

|  |  |
| --- | --- |
| Onsets | Duration |
|  |  |

*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.

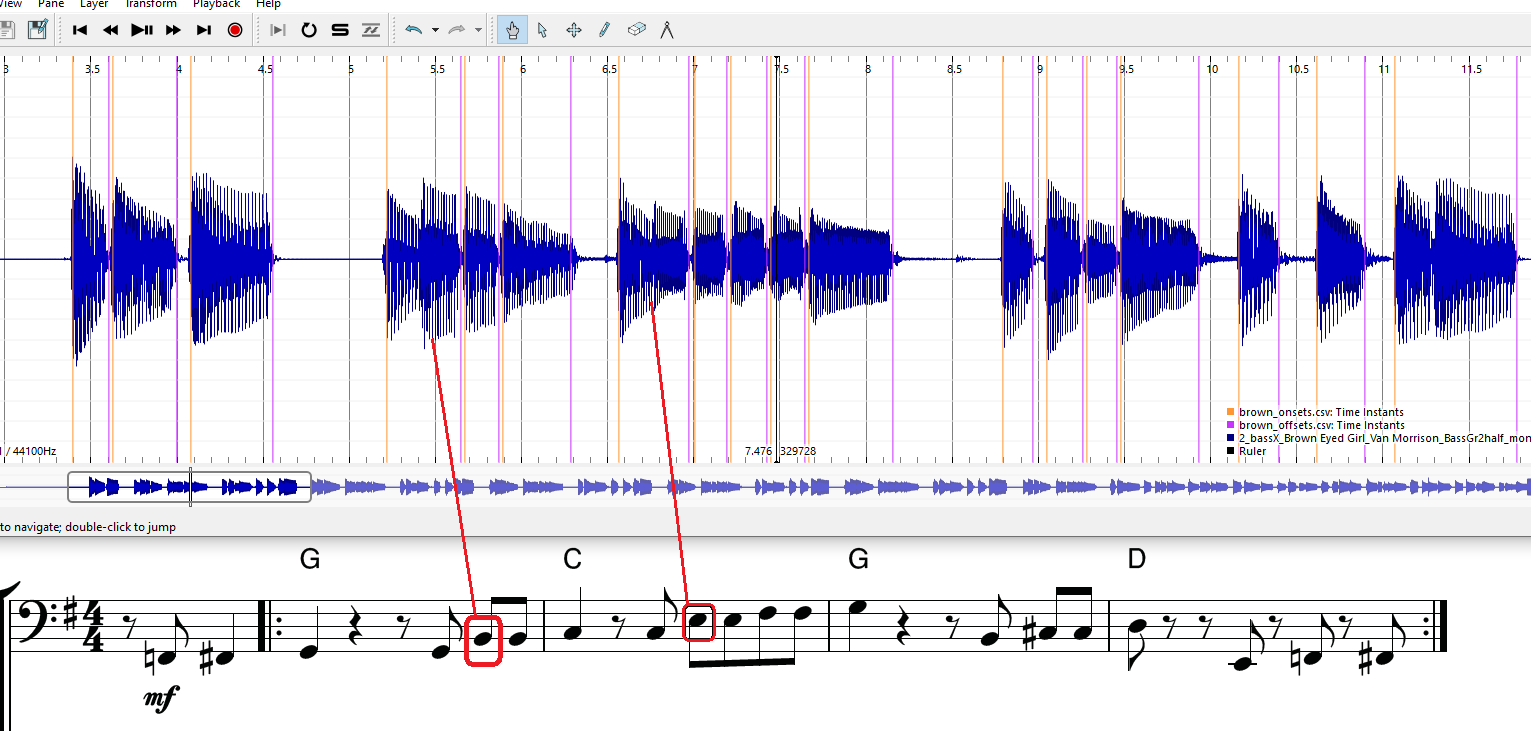
* 1. Brown Eyed Girl

“brown” from grade 2 is a good example in highlighting the limits of all the state of the two main algorithms shortlisted. The Recall metric for the IEC methods was only 82 % and the SOP not much better at 85%. Although IEC had higher precision the closeness of the notes generated “Sound Archipelagos” instead of sound islands.

This song tests the limitations of the MIR evaluation window of 20 ms, when you consider the Onsets for the first 8 notes of the song, 5 of the gaps are less than 25ms.

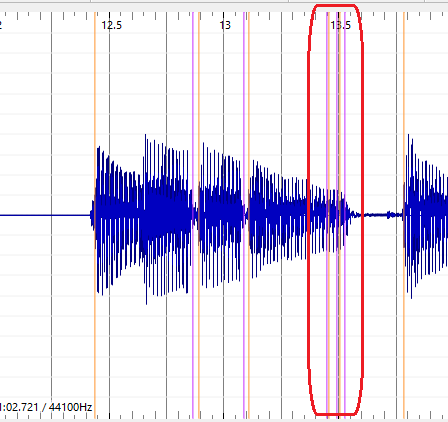
*Table 17:First 8 onsets of “brown”*

|  |  |
| --- | --- |
| **Onset Mark** | **Difference** |
| 3.385 |  |
| 3.62 | 0.235 |
| 3.96 | 0.34 |
| 4.08 | 0.12 |
| 5.215 | 1.135 |
| 5.42 | 0.205 |
| 5.66 | 0.24 |
| 5.88 | 0.22 |



*Figure 38: Brown: Missed Onset pattern*

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 an acceptable amount.



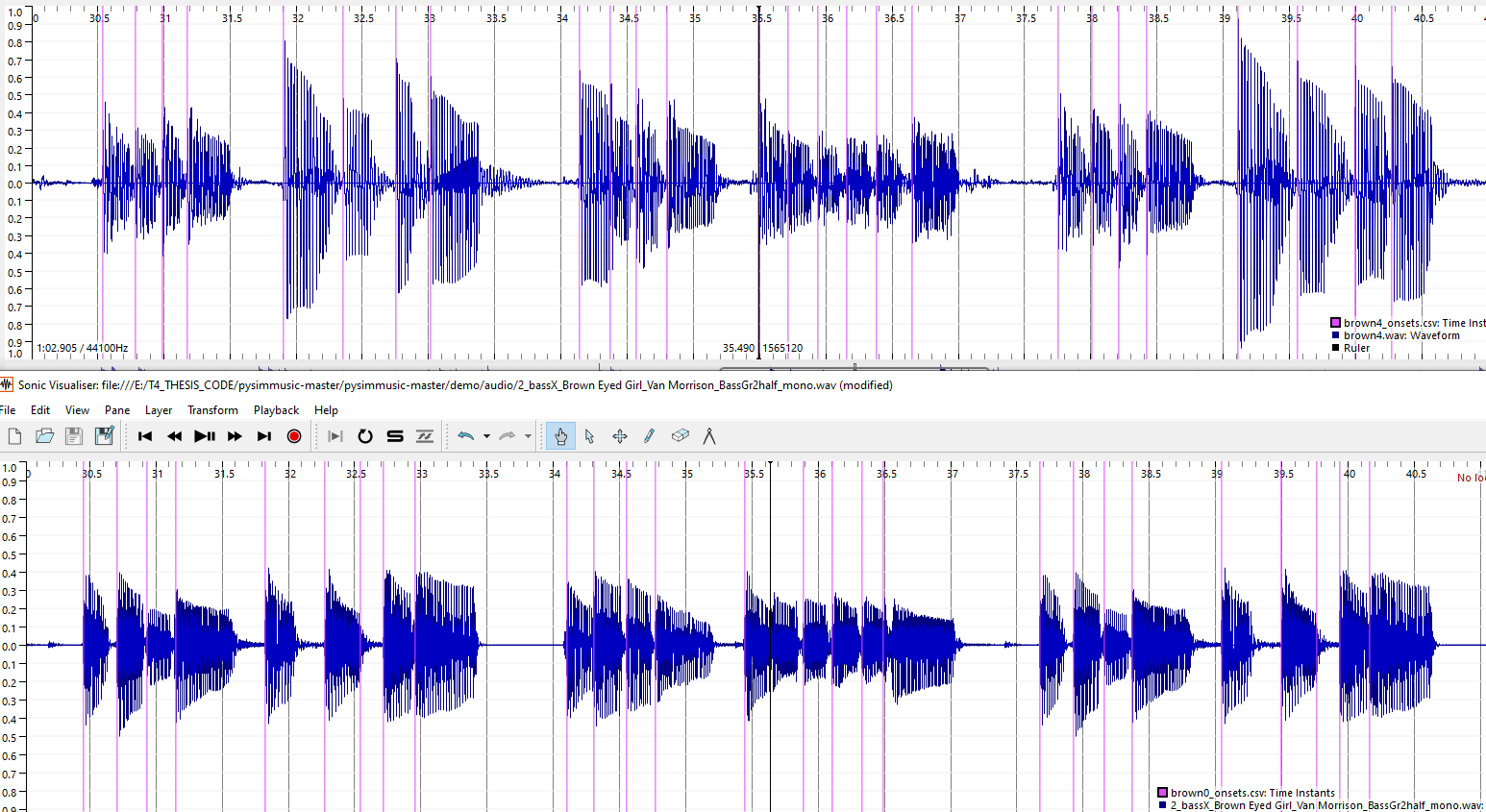
*Figure 39: Brown: False Onsets*

Notwithstanding these limitations, the same procedure of evaluating students was applied like the other songs. However, it would be useful to have “confidence formula” that depended on the best ground truth measurements. However, the student performances were well distributed. The best graded student had PRF results in the 10%, so this raises question about the confidence and robustness of the IEC algorithm to short notes and note intervals.

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

*Figure 40: Brown: Grade Plots*

The big outlier is Student 4 who scored 90% grade , but a precision around 10%. The investigate this the wave from of the Ground Truth and Student are compared in the figure below. At 36.5 secs Student 4 correctly hits 6 offsets but the GT only hits 5 onsets and also as the song progresses. There is almost a 10ms drift at 39.5 second timestamp.



*Figure 41: Brown: GT and Student 4 stems with onsets detected*

The extreme case was discovered in brown that the IEC algorithm returned zero offset deviations for one student (Student 4). This can happen when the PRF is very low These case should be considered “stress test” case for the IEC algorithm in terms of evaluation time limits.

* 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 cutoff point was made at 34.5 seconds to remove them all. This resulted in getting a better PRF score using the IEC algorithm. The Grade Prediction for Onset and Duration worked optimally when considering the Mean Onset and Mean Duration respectively alongside the PRF

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

*Figure 42: Roadrunner: Onset & Duration Grades*

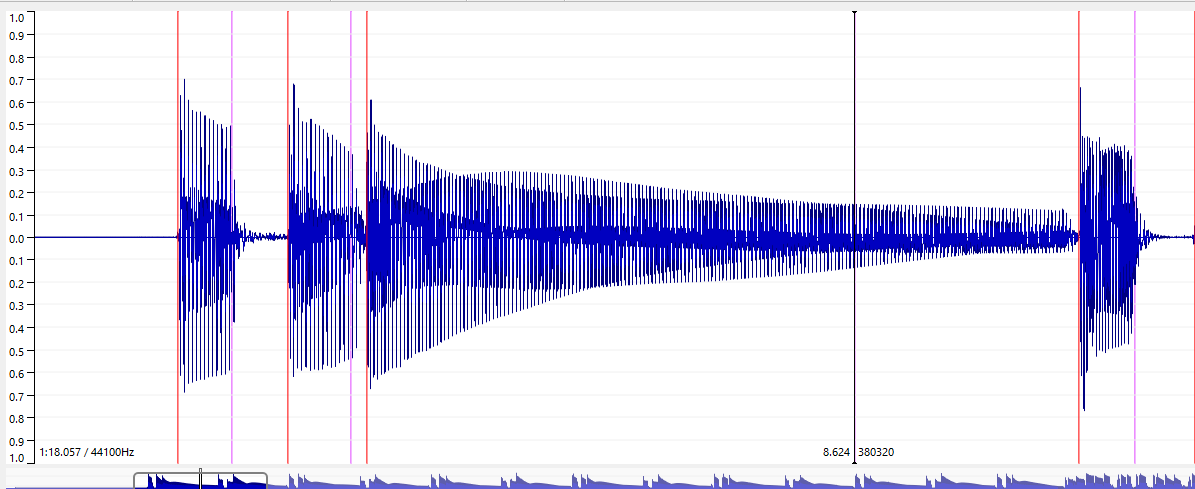
* 1. Walking on the Moon

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. 0.983, 0.991, 0.987.

It is important to point out a tradeoff 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. Given that the Energy Checker didn’t return a high Recall, how would you determine if this subtle gap is respected in a student recording? The Essentia library function “Effective Duration” considers perceptual factors [25]when measuring note length. For the first three notes of “wotm”, the threshold setting of 0.05 returns offset points that approximately align with the human hearing threshold



*Figure 43: WOTM: Offsets using Effective Duration (T= 0.05)*

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.

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

*Figure 44: Wotm: Onset & Duration Grades*

* 1. Technical Control Grades

Bjean having the most recordings 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 18: Tech Control Grades Actual vs Predicted*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Technical Focus 1 | | Technical Focus 2 | |
| **Song**. | Actual | Predicted | Actual | Predicted |
| yellow |  |  |  |  |
| bjean | Actual Predicted  0 72.000000 75.644342  1 90.000000 70.263078  2 79.199997 76.314501  3 33.299999 67.726495  4 79.199997 86.336615 |  | 0 79.199997 82.306743  1 90.000000 77.707844  2 79.199997 84.835356  3 56.700001 75.214065  4 72.000000 86.350437 |  |
| just |  |  |  |  |
| Brown | 79.199997  79.199997  72.0 | 76.876749  77.229587  76.876749 | N/A | N/A |
| Road |  |  |  |  |
| Wotm | 72.0  72.0  79.19  56.70 | 80.018391  82.625873  74.860405  78.773665 | N/A | N/A |

“Brown” yielded a low error prediction. Caution has to be exercised in drawing an conclusions, considering the onset and duration grade error were over 10%. However, it does highlight the case that maybe the teacher should grade the Technical Focus Areas 1 and 2 in one joint grade.

*Table 19: Technical Grade Prediction Errors*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Song. | # S | Algo. | MAE Onset TF1 | RMS Error Onset TF1 | Extra Inputs  TF1 | MAE  Duration  TF2 | RMS Error Duration  TF2 | Extra  Inputs  TF2 |
| yellow | 8 | SOP |  |  |  |  |  | ABS Mean |
| bjean | 15 | IEC/SOP | 13.56 | 18.15 |  | 10.77 | 12.17 | - |
| just | 11 | SOP |  |  |  |  |  |  |
| brown | 8 | IEC | 3 | 3.3 |  | N/A | N/A | N/A |
| road | 8 | IEC |  |  | - |  |  | - |
| wotm | 11 | SOP/IEC | 11.26 | 13.07 | - | N/A | N/A | N/A |

* 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. In the TCL R&P exam the student performs 3 songs and is graded in the previously mentioned 3 areas (Fluency, TF and Communication). A number between 1 and 8 is given and comments between 10 and 40 words are given. In the thesis experiment, separate comments were allocated for each of the five metrics that are graded: onset, duration, technical focus, sound quality and dynamics.

An online Natural Language Processing engine called Sentiment-Analysis[[17]](#footnote-17) as applied to the teacher comments after Translating the comments using the Deepl online translator[[18]](#footnote-18).

1. Conclusion
   1. Student Recordings and Teacher Gradings

It wasn’t not possible to find music students to do the experiment. Five submissions were collected by researchers at the MTG and I submitted the remaining (48) recordings. Some of the bassists requested to participate had issues with music notation literacy and others were not comfortable in having to strictly follow the score and technical control details as laid out.

For future experiments the bass students should be sourced from undergraduate music programs at different third level music schools. This would avoid music literacy problems and any other issues related to following the musical techniques that are evaluated on each song, on the other hand it could exclude bassists who learn in informal contexts.

A comprehensive experiment would require compensating the participating students sufficiently to motivate them to complete the questionnaire, do the latency test correctly and optimize their technical setup to produce the best sound possible on the playback. A compensation scheme could be economic one (e.g., a fixed amount per song for the “informal context” students), academic (e.g., negotiating allocating credit as part of a subject in an undergraduate program or possible prize for best recording.). If an agreement could be established with a Music school, an agreement to allow music-undergrads and post grads to participate as students and teachers in a grading experiment and allocate credits for participation could be mutually beneficial. The benefits of collecting different performances and sounds from different bassists are clear from the experiences in this experiment. The most interesting observations came from the recordings that were not my own. One student managed to obtain the best PRF score with a guitar and another student scored highest on “bjean”, despite having deviated from the score. Having had to do over 40 recordings (and more attempts), I noticed improvements each time I recorded the difficult songs “brown” and “road”, but then the “law of diminishing returns” set in and subsequent performances don’t improve.

Due to time constraints, the Bass Teacher got 43 out of 48 gradings done. One mistake that was pointed out by participating student, was that the Bass Teacher was not given the ground truth under the guise of a student attempt. This would have helped benchmark the real attempts. The subsequent “downgrading compensation” that I applied was to add more validity to the grades rather than push for lower prediction errors.

* 1. Technical and Musical issues with Experiment

Another restriction was that Music Critic is not supported on Tablets and Smart Phones, and one professional bassist reported no having a sound card on his PC, indicating he uses his Tablet for recording. One workaround for this problem would be to set up a recording workshop on a fixed date, with a dedicated laptop running Audacity (or another DAWS), connected with an audio interface and headphone. The students could come in record without any latency test. This would be like an exam simulation and may suit some students who like to time limit their attempts and are driven by “getting it right first time”. On the other hand, their might be students who prefer to prepare the recordings in the comfort of their own environment, taking as many attempts as they wish.

One mistake made in the experiment was not performing the truncation of the backing track and score at a very clear point in time or section of the music. The idea of truncation was to reduce the student and teacher effort. It was also necessary to truncate some songs further, e.g., “road”, to make the algorithms work properly. for subsequent experiments, the score that gets uploaded should clearly show the stop point for student evaluation and the backing track should either fade out or stop accordingly.

The quality of the mix is also a factor in facilitating the Bass teacher allocate a fair score. Perhaps a more thorough grading policy would be to allow the teacher listen to the recorded stem as well as the mix. Another weakness in the grading technique is linearly “going through the students one-by one”. In real exam correcting situations a teacher may prefer to have multiple iterations of grading student. Finally, the onset and offset grade should be mark between 1-8 and the Technical Control grades should be discarded, until such time that significant progress is made on accurately measuring duration.

* 1. Improving prediction strategies

Although I made additional submitted recordings and gradings, the total number was lower than other research feature extraction papers by Abesser and it meant a limited choice of Machine Learning algorithms. I found that in some cases, as shown in Tables 13 and 19*,* showed thatadding the deviation statistics information did not help in improving the predicted grade. For the Onset Grade Prediction, it was the Mean rather than Absolute Mean deviation which added more value. A wider set of tests, with Mean deviation calculated separately for the early and late onset would help to get a better understanding of this relationship. For offsets absolute mean rather than mean deviation, helped reduce the predicted grade error metrics. Apart for Duration Grade prediction in “bjean”, adding the Standard Deviation statistic made the prediction a lot worse. Musically speaking , the song style would dictate the teacher sensitivity is to lateness/vs earliness and note longitude.

No further research was added to Abessers classification of plucking styles, but the three column CSV folder, would be a good continuation point for adding more functionality and parameters to the algorithms. For example, the ground truth input files could not only contain annotated onsets and offsets, but also the plucking style on each note which a detector could compare against. 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. Improving Duration measurements

The main challenge was to do a score accurate check on long notes, typically composed 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., IEC and SOP and the tolerance window for Offset detection was double that of Onset detection. A more rigorous measurement approach would incorporate the Effective Duration as discussed in section 6.6. into the algorithms and narrow the matching window size.

The repeated quavers in the verse for Yellow leaves no gaps for duration so its meaningless here. Theoretically it would be possible to play shorter notes than quavers, but that is actually more difficult to do, so it is not worth checking for a minimum note length

There are some areas that were too complex to analyze, e.g. The Tenuto musical property in “road”. A starting point for looking further 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. Going forward

To end I list suggested future paths for scaling up future experiments to continue the research:

* A pilot project that would involve a selection of up to 20 students who are studying bass guitar in private schools and conservatories.
* Continually annotating more Ground Truth songs
* Source Separation to obtain more annotated data.
* Consideration of other Music Syllabuses e.g. Rockschool[[19]](#footnote-19)

1. List of figures

(Optional)

[*Figure 1: Local Energy Equation.* 4](#_Toc79152027)

[*Figure 2: 4 crotchets, 4 quavers and 4 semi-quavers.* 5](#_Toc79152028)

[*Figure 3: Equation: Peak Picking formula* 9](#_Toc79152029)

[*Figure 4: dynamic threshold setting* 9](#_Toc79152030)

[*Figure 5: Extract from IDMT Dataset. File 002.xml* 13](#_Toc79152031)

[*Figure 6: Note duration of Staccato notes* 14](#_Toc79152032)

[*Figure 7: Audio 002.wav (IDMT DATASET) with onsets* 14](#_Toc79152033)

[*Figure 8: Technical Information in XML files* 16](#_Toc79152034)

[*Figure 9: Billie Jean: Midi Rendering vs Human recording (shorter notes)* 17](#_Toc79152035)

[*Figure 10: Strategy to experiment, test to find best accuracy* 21](#_Toc79152036)

[*Figure 11: Original Energy Checker for offset detection* 23](#_Toc79152037)

[*Figure 12: Normalized Energy of Audio 002.wav sample (IDMT DATASET)* 24](#_Toc79152038)

[*Figure 13: Normalized Energy Function* 24](#_Toc79152039)

[*Figure 14: Sound Islands of Audio 002.wav (IDMT). FS = 1024, HS = 512.* 25](#_Toc79152040)

[*Figure 15: IEC onsets for WOTM verse, P = 0.514.* 26](#_Toc79152041)

[*Figure 16: Calculated IEC Sound Islands for WOTM bridge, P = 0.943* 27](#_Toc79152042)

[*Figure 17: Calculated IEC Onset positions for WOTM verse* 27](#_Toc79152043)

[*Figure 18: PitchMelodia derived Onsets for WOTM verse , P = 0.97* 28](#_Toc79152044)

[*Figure 19: PitchMelodia derived Onsets for WOTM bridge* 28](#_Toc79152045)

[*Figure 20: IDMT Dataset GT Onset + Duration deviations ( IEC ) algorithm* 29](#_Toc79152046)

[*Figure 21: IDMT Dataset GT Onset deviations ( SOP) algorithm* 30](#_Toc79152047)

[*Figure 22: Billie Jean GT onsets deviation (*Window 20ms) 30](#_Toc79152048)

[*Figure 23: Billie Jean GT offsets deviation (? Algorithm)* 31](#_Toc79152049)

[*Figure 24: End to End test of Music Critic with Billie Jean GT* 32](#_Toc79152050)

[*Figure 25: Workflow on Student Portal* 34](#_Toc79152051)

[*Figure 26: Processing a student stem in Audacity* 36](#_Toc79152052)

[*Figure 27: Technical instruction and link to recording portal* 36](#_Toc79152053)

[*Figure 28: Extract of guide given to teacher for grading* 37](#_Toc79152054)

[*Figure 29: Grading Histogram Onsets: Billie Jean* 39](#_Toc79152055)

[*Figure 30: Grading Histogram Offsets: Billie Jean* 39](#_Toc79152056)

[*Figure 31: Contrasting comments of WOTM performances* 40](#_Toc79152057)

[*Figure 32: Syncopation example Just Looking* 41](#_Toc79152058)

[*Figure 33: Syncopation example: WOTM* 41](#_Toc79152059)

[*Figure 34: Plot of Onset Grades vs Precision* 45](#_Toc79152060)

[*Figure 35: Plot of Precision (X) vs Onset Grades (Y) Billie Jean* 47](#_Toc79152061)

[*Figure 36:Best vs Worst bjean students* 48](#_Toc79152062)

[*Figure 36: Just Looking: Grades (Y) Precision (X)* 49](#_Toc79152063)

[*Figure 38: Brown: Missed Onset pattern* 50](#_Toc79152064)

[*Figure 39: Brown: False Onsets* 50](#_Toc79152065)

[*Figure 40: Brown: Grade Plots* 51](#_Toc79152066)

[*Figure 41: Brown: GT and Student 4 stems with onsets detected* 51](#_Toc79152067)

[*Figure 42: Roadrunner: Onset & Duration Grades* 52](#_Toc79152068)

[*Figure 43: WOTM: Offsets using Effective Duration (T= 0.05)* 53](#_Toc79152069)

[*Figure 44: Wotm: Onset & Duration Grades* 54](#_Toc79152070)

1. List of tables

(Optional)

[*Table 1: Abessers Data Driven algorithms vs Salamon/Gomez* 6](#_Toc79152007)

[*Table 2: Abesser Signal Processing vs Salamon/Gomez* 7](#_Toc79152008)

[*Table 3: State of the Art algorithms for Onset/Offset detection* 7](#_Toc79152009)

[*Table 4: Six chosen songs from TCL song list* 15](#_Toc79152010)

[*Table 5: Tech. Control Parameters for 6 TCL songs* 18](#_Toc79152011)

[*Table 6: Benchmark results of Algorithms IDMT DS* 23](#_Toc79152012)

[*Table 7: Accuracy Metrics for 6 TCL ground truths (combined vs non combined)* 25](#_Toc79152013)

[*Table 8: Accuracy Metrics for Billie Jean Hybrid Algorithm* 26](#_Toc79152014)

[*Table 9: Selected comparison: IEC vs SG* 26](#_Toc79152015)

[*Table 10: Technical Control Topics (TODO refer to other table related)* 38](#_Toc79152016)

[*Table 11: Comparison IEC vs SG* 38](#_Toc79152017)

[*Table 12: Grades Actual vs Predicted* 43](#_Toc79152018)

[*Table 13: Grade Prediction Errors* 43](#_Toc79152019)

[*Table 14:Best two “yellow” students* 45](#_Toc79152020)

[*Table 15:Billie Jean: Actual vs Predicted Grades Overall* 47](#_Toc79152021)

[*Table 17:First 8 onsets of “brown”* 49](#_Toc79152022)

[*Table 18: Tech Control Grades Actual vs Predicted* 55](#_Toc79152023)

[*Table 19: Technical Grade Prediction Errors* 55](#_Toc79152024)

[*Table 20:* Inputs (yellow) vs outputs (green) for Billie Jean Onsets 69](#_Toc79152025)

[*Table 21:* Inputs (yellow) vs outputs (green) for Billie Jean Offset 70](#_Toc79152026)

1. Bibliography

|  |  |
| --- | --- |
| [1] | N. McCormick, U2 by U2: Bono, The Edge, Adam Clayton, Larry Mullen Jr, New York: HarperCollins., 2009. |
| [2] | E. C. J. A. a. S. G. Christian Dittmar, "Music Information Retrieval Meets Music Education," in *Multimodal Music Processing*, Dagstuhl, Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik, 2012, pp. 95-120. |
| [3] | P. Pfeiffer, Bass Guitar For Dummies, 3rd edition, Hoboken, NJ.: John Wiley and Sons , 2014. |
| [4] | H. G. S. &. R. Bantula, "Jazz ensemble expressive performance," in *17th International Society for Music Information*, 2016. |
| [5] | T. C. London, “Trinity R&P Bass Syllabus from 2018,” in *Trinity R&P Bass Syllabus from 2018*, London, Trinity College London, 2017, p. 32. |
| [6] | J. S. a. E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," in *IEEE Transactions on Audio, Speech, and Language Processing,*, 2012. |
| [7] | Jakob Abeßer and Meinard Müller, "Jazz Bass Transcription Using a U-Net Architecture," *Electronics,* vol. 10, 12 March 2021. |
| [8] | J. Abeßer and G. Schuller, "Instrument-centered music transcription of solo bass guitar recordings," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (9), 1741*, 2017. |
| [9] | P. D. M. M. D. T. A. K. Prof. Dr.-Ing. Gerald Schuller, *Automatic Transcription of Bass Guitar Tracks applied for Music Genre Classification and Sound Synthesis,* Technischen Universität Ilmenau, 2014. |
| [10] | J. P. Bello, and M. B. Sandler, "Phase-based note onset detection for music signals," in *ICASSP-88*, 2003. |
| [11] | J. Bello, C. Duxbury, M. Davies and M. Sandler, "Phase-based note onset detection for music signals," in *IEEE Signal Processing Letters*, 2004. |
| [12] | C. Kopp-Scheinpflug, J. L. Sinclair and Jennifer F. Linden., "When Sound Stops: Offset Responses in the Auditory System," *Trends in Neurosciences,* no. Special Issue: Time in the Brain, 2018. |
| [13] | J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler, "IEEE A Tutorial on Onset Detection in Music Signals," in *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING,*, 2005. |
| [14] | L. Reboursiere, O. Lahdeoja, T. Drugman, S. Dupont, C. Picard and N. Riche, "Left and right-hand guitar playing techniques detection," in *New Interfaces for Musical Expression*, 2012. |
| [15] | M. Goto, "Predominant-F0 estimation for detecting melody and bass lines in," in *Speech Communication*, 2004. |
| [16] | J. Salamon, R. Bittner, J. Bonada, J. Bosch, E. Gómez and J. Bello, "An Analysis/Synthesis Framework for Automatic F0 Annotation of Multitrack Datasets," in *International Society for Music Information Retrieval Conference*, Suzhou, China, 2017. |
| [17] | J. F. K. P. M. Abeßer and W.-G. Zaddach, "Introducing the Jazzomat project - Jazz solo analysis using Music Inf. Retrieval methods," in *Int. Symposium on Computer Music Multidisciplinary Research (CMMR) Sound, Music and Motion*, Marseille., 2013. |
| [18] | Fraunhofer. [Online]. Available: https://www.idmt.fraunhofer.de/en/business\_units/m2d/smt/bass.html. |
| [19] | V. Eremenko, M. A, N. J and S. X, "Performance assessment technologies for the support of musical instrument learning," in *12th International Conference on Computer Supported Education*, 2020. |
| [20] | Mirex, "music-ir.org," MIREX, [Online]. Available: https://www.music-ir.org/mirex/wiki/2020:Singing\_Transcription\_from\_Polyphonic\_Music. |
| [21] | R. Hennequin, A. Khlif, F. Voituret and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *The Journal of Open Source Software,* 2020. |
| [22] | Trinity College London, "R&P Bass Grade 1," in *R&P Bass Grade 1*, Londoin, 2017, p. 9. |
| [23] | C. Kehling, J. Abeßer, C. Dittmar and G. Schuller, "Automatic Tablature Transcription of Electric Guitar Recordings by," in *DAFx*, 2014. |
| [24] | MTG. [Online]. Available: https://github.com/MTG/pysimmusic. |
| [25] | G. Peeters, "G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," CUIDADO I.S.T., 2004. |
| [26] | "RockSchool," [Online]. Available: https://www.rslawards.com/country/espana/. |

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

Technical Focus Grades are not included in tables 14 and 15.

*Table 20:* 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 |

*Table 21:* 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. Examples include <https://yousician.com/>, https://fretello.com/ [↑](#footnote-ref-1)
2. https://www.idmt.fraunhofer.de/en/business\_units/m2d/smt/bass.html [↑](#footnote-ref-2)
3. http://www.yousician.com/ [↑](#footnote-ref-3)
4. https://github.com/CPJKU/madmom [↑](#footnote-ref-4)
5. https://essentia.upf.edu/ [↑](#footnote-ref-5)
6. https://essentia.upf.edu/reference/std\_PitchMelodia.html [↑](#footnote-ref-6)
7. <https://github.com/jakobabesser/bassunet> [↑](#footnote-ref-7)
8. https://github.com/jakobabesser/pymus/tree/master/pymus/sisa/f0\_tracking [↑](#footnote-ref-8)
9. https://github.com/CPJKU/madmom [↑](#footnote-ref-9)
10. https://github.com/CPJKU/madmom/blob/master/bin/OnsetDetectorLL [↑](#footnote-ref-10)
11. https://github.com/MTG/pysimmusic-experiments [↑](#footnote-ref-11)
12. https://github.com/RamoonRoomeu/ToneExperiments [↑](#footnote-ref-12)
13. https://musiccritic.upf.edu/ [↑](#footnote-ref-13)
14. https://musescore.org/en [↑](#footnote-ref-14)
15. https://www.sonicvisualiser.org/ [↑](#footnote-ref-15)
16. The student tries to give best recording possible, improving on each take with no deliberate mistakes. [↑](#footnote-ref-16)
17. https://deepai.org/machine-learning-model/sentiment-analysis [↑](#footnote-ref-17)
18. https://www.deepl.com/en/home [↑](#footnote-ref-18)
19. https://www.rockschoolespana.com/ [↑](#footnote-ref-19)