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

Universitat
Pompeu Fabra
Barcelona

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

Universitat
Pompeu Fabra
Barcelona

Table of Contents

1. Introduction ... 1
2. State of the Art ... 3
2.1. Music Education ... 3
2.2. Onset Detection ... 4
2.2.1 Perceptual considerations for onsets ... 4
2.2.2 Technical Discussion of Onsets ... 4
2.3. Offset Detection ... 8
2.3.1 Perceptual considerations for offsets ... 8
2.3.2 Musical considerations for offsets ... 9
2.3.3 Voicing Classification ... 9
2.2.3 Technical Discussion of Offsets ... 10
2.4. Neural Network Strategies ... 11
2.5. Score and Instrument Level Parameters ... 13
2.6 Summary ... 15
2.7 Observations ... 15
2.8 Machine Learning Techniques in related areas ... 16
3. Datasets ... 18
3.1. IDMT BASS SINGLE TRACK ... 18
3.2. Rock and Pop repertoire ... 20
3.3. Musical Properties of R&P repertoire ... 23
4. Methodology ... 24
5. Algorithm Development and evaluation ... 26
5.1 Music Critic End-to End test ... 26
5.2 Algorithm Evaluation ... 27
5.3 Indexed Energy Checker (IEC) ... 28
5.4 Accuracy Metrics: IEC vs MELODIA for R&P ... 31
5.4.1 Long Notes ... 31
5.4.2 WOTM Bridge and Verse Comparison ... 31
5.4.3 Classifying Muteness ... 34
5.5 Combined IEC-SOP for R&P ... 35
5.6 Accuracy Metrics for Neural Network Methods ... 36
5.7 Deviation Metrics on Reference ... 36
5.8 Summary ... 38

6. Prediction of Student Performance 39
6.1. Student Recording Portal 39
6.2. Processing of student recordings 40
6.3. Teacher Grading Policy 41
6.4. Teacher Grading Results Analysis 44
6.5. Yellow 45
6.6. Billie Jean 46
6.7. Just Looking 49
6.8. Brown Eyed Girl 50
6.9. Roadrunner 53
6.10. Walking on the Moon 53
6.11. Technical Control Grades 54
7. Conclusion 56
7.1. Student Recordings and Teacher Gradings 56
7.2. Musical and Technical issues with Recording Experiment 56
7.3. Improving analysis and prediction strategies 58
7.4. Improving Duration measurements 59
7.5. Going forward 59
8. List of figures 61
9. List of tables 62
10. List of symbols 63
11. Bibliography 64
12. Appendices 67
12.1. Code for generating data. 67
12.2. Reproduction of Predicted Grades Plots 69
12.3. Extra Plots for Predicted Grades 71
12.4. Student Portal 77
12.5. Steps for Processing Student Recordings 78
12.6. Teacher Portals 79
12.7. Overall Student Performances 81
12.8. Billie Jean Student Performances 82
12.9. Performance of Neural Net methods 84

Dedication

I would like to thank those closest to me that made this whole Master’s program possible, my family in Barcelona. In 2016, I read the biography, “U2 on U2” [1]. This work by Neill McCormack, inspired me to take up the bass again, take lessons, do exams, look at how music technology could help and eventually do the SMC Masters full time. I was fascinated about the musical journey of U2, how the sonic identity developed and how the producers and played a big role in giving feedback to the musicians. I am very grateful for having the opportunity to do research on the electric bass guitar for Music Education purposes with Vsevolod and Xavier as supervisors. It has been a long academic year and I hope the contribution to performance assessment of electric bass guitar can continue in 2021/2022. Thanks to Marti for diligently correcting the student recordings and to Ramon for sharing his ideas with me.
Thanks to Dmitry for helping me build up my knowledge of Essentia during the Internship and to Frederic, Vsevolod and Fabio for volunteering to record basslines.

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 were from 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, they don’t provide feedback on note length. My 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 (R&P) repertoire.

Keywords:

Audio Signal Processing, Automatic Music Transcription, Bass transcription, durations, electric bass guitar, expression style, expressive performance analysis, fretboard, Indexed Energy Checker (IEC), Machine Learning, Mean Absolute Error, Music Assessment, Music Education, Modern Music, Music Information Retrieval, Music Performance Analysis, offset, onset, playing technique, plucking, PRF accuracies, position, Rhythm, Rock and Pop (R&P), Sound Archipelagos, Sound Islands, Sound Onset Processor (SOP), source separation, string detection, style, deviation statistics, Timing.

1. Introduction

Using technology to assist in Music Performance Assessment in the context of Music Educations is the core subject matter of this thesis. Audio Signal Processing (ASP) and Music Information Retrieval are the technologies used and Dittmar [2] gives us a brief history of the role of MIR in Music Education. A key step forward was the transition to digital formats for both recorded and symbolic notation and hence the transition from CDs and score books to today's smart phone apps. Applications such as "Yousician"[1] and Fretello offer performance assessment for learning help guide the student on tuning note accuracy, pitch accuracy and 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, and good use of dynamics.

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 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 formed a band before even knowing how to play individually. They relied on feedback given by friends and some tutoring by Richard Evans (the Edge's brother) and it took them nearly eight months before they produced a sound that later got them signed to Island Records. Later in his career Adam Clayton (U2 Bassist) sought bass lessons from Patrick Pfeiffer, author of the book "Bass for Dummies" [3]. Bass lessons are not only for novice students, even the most accomplished professionals seek to improve their musical knowledge.

If the student wants to improve playing the jazz style, there are tools that can assist you play alternative interpretations of the score in a topic known as expressive performance modelling [4]. This research focuses on the modern music session skills that a music producer expects when recording bass tracks. 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 and technical control of the instrument to produce a good

[1] https://yousician.com/, https://fretello.com/

sound. The objective is the research and development of a model that can automatically assess a student's performance of the bass guitar and provide them with the useful feedback that can help them improve.

The thesis opens with the State-of-the-Art Chapter and refers to examples from the two datasets under study: the Fraunhofer IDMT Single Track Dataset[2] and a Rock and Pop Repertoire Dataset with six selected popular songs. These two datasets are described in more detail in Chapter 3, focusing on how they helped to achieve the goals of the thesis. Chapter 4 describes the overall methodology of the thesis. Chapter 5 then describes the evaluation of the algorithms using the reference bass recording as audio input. Chapter 6 describes the experiments leading to predicting student performances and summarizes the results of the analysis of recordings. Chapter 7 discusses observations, implications and lessons learnt from experiments for continuing research into performance assessment methods. The appendices contain links to the experiment data: student portal information, Data exports (raw data and histograms) of Teacher Feedback, GitHub links to code running algorithms, GitHub links to Notebooks for reproducing the grade predictions and with some extra example tables and plots.

[2] https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/bass.html

2. State of the Art

2.1. Music Education

The mastery of any musical instrument without the live presence of teacher is a challenging task. Taking one application introduced in Chapter 1, “Yousician”, any casual navigation of the tool for bass clearly shows that no points are lost when you don’t hold a note for the correct length or play “imposter notes” during silences. It doesn’t matter you if play the first four crotchets like quavers in the example below. If the onsets are correct, you still get the “Perfect” feedback score.

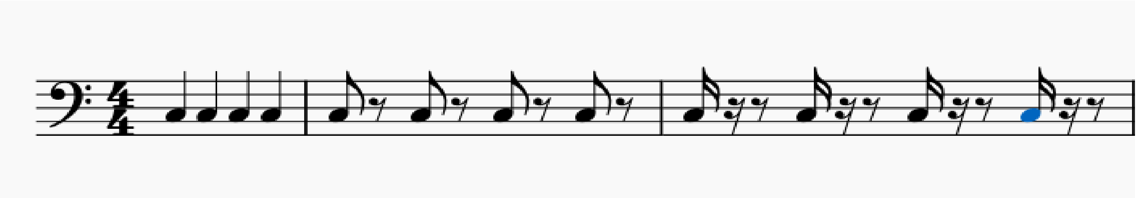


Figure 1: 4 crotchets, 4 quavers and 4 semi-quavers.

Duration is not controlled on any of the note above. Also, if you add a note (imposter note)during silent periods of large enough duration, you will not lose marks.

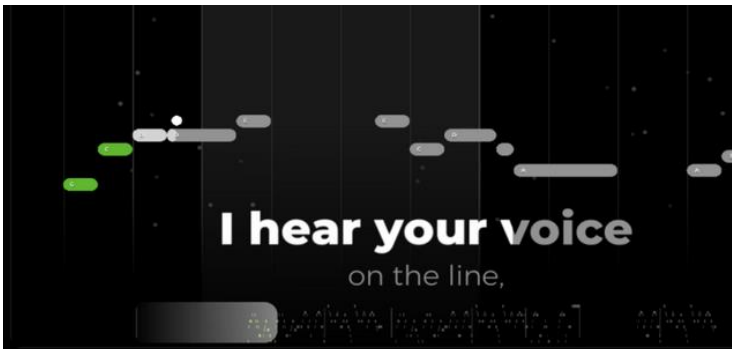


Figure 2: Yousician example with Vocals

“Yousician” uses piano roll format display for giving feedback on vocals and this gives the impression that note length is detected. However, a test on the “Yousician-Premium Plus” version of the song with long notes, e.g. “The one I love” by REM, demonstrates that it doesn’t work. The effectiveness of a Performance Assessment Technology (PAT) is the readiness level it provides to a student for auditions or music exams (e.g., TCL R&P

[5]). Music Critic[3] is currently developed for Guitar at the MTG[4] (Music Technology Group) and has addressed rhythm assessment using the Spectral Onset Processor(SOP)[5] from the Madmom library. Music Critic is non-realtime: it does not flash green or red on each individual note while playing like in Yousician. The feedback is delivered with an annotated score diagram along with a numeric grade data based on global pitch and timing statistics. This information is delivered after the performance is submitted on the Music Critic portal. This setup is more suitable for student preparing for real exams than real-time assessments.

2.2. Onset Detection

2.2.1 Perceptual considerations for onsets

Research on the auditory system proposes that we have a temporal energy integrator [6]. The integration time is estimated to be between 50ms and 200ms with some researchers proposing longer values for low frequencies. This relationship between the temporal resolution of the auditory system and onsets is beyond the scope of this thesis. However, this observation of longer integration for lower frequencies is interesting and it has implications for the perceived attack time for onsets for bass guitar.

2.2.2 Technical Discussion of Onsets

Before a technical discussion of onset detection can begin, the relevant background concepts (energy and events) need to be described. For a given hop size “h”, hop number “m”, the local energy of the frame of an audio signal “x” can be calculated as follows

$$E(m) = \sum_{n=(m-1)h}^{mh} \left| x[n]^2 \right|$$

Equation 1 Local Energy Equation

[3] https://musiccritic.upf.edu/
[4] https://www.upf.edu/web/mtg
[5] https://github.com/CPJKU/madmom/blob/7e560700987d2800f1c78fdb2a5faba77ecece0d/madmom/features/onsets.py#L585

The energy “E” is typically measured for a specific energy band (e.g., for bass guitar typically from 25 Hz to 500Hz), so the Energy is calculated from bin index k1 to k2:

$$E = \sum_{k=k1}^{k2} \left| X[k]^2 \right|$$

Equation 2 Energy in a Frequency Band

where X[k] is the STFT of x(n). An onset detection function (ODF) is used to detect a musically meaning event in an audio signal and returns one value per audio frame. In the case of bass guitar, a pitched percussive instrument, we are concerned with the plucking event. A simple ODF calculation can be made by taking the energy difference of two consecutive audio frames:

$$ODF(frame\ index) = E(frame\_index)\text{-}E(frame\_index\ \text{-}1)$$

Equation 3 Onset Detection Function Equation

A more general strategy for detecting onsets involves various stages:

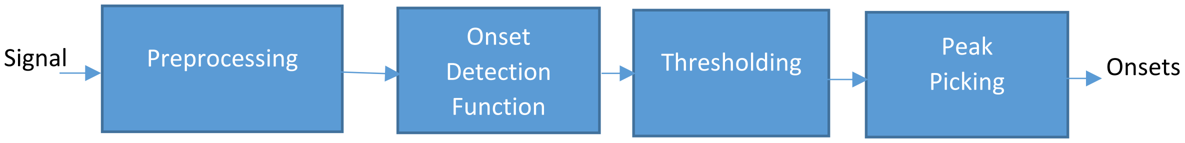


Figure 3: Onset Detection Stages

- pre-processing is an optional step that emphasises or attenuates aspects of the signal, for example separating the signal into different frequency bands.
- ODF reduces the signal to subsampled occurrences of transients using “ODF Methods”.
- The “Thresholding” eliminates peaks which are not related to the onset event. Bello discussed how a median filter is used to obtain an adaptive threshold curve [7].
- Peak Picking involves choosing the local maxima after threshold filtering.

ODF Methods

The ODF method can be chosen based on instrument class: pitched or non-pitched; percussive or non-percussive. The Spectral Onset Processor[6] (SOP) implements several onset detection functions based on magnitude or phase information from the spectrogram of an audio signal. Out of Bello's fourteen ODF Methods, Spectral Difference and Phase Deviation [7] scored the highest accuracy metrics for pitched percussive instruments. In the Essentia[7] library, the Onset Detection Function has six methods with examples[8]. In the following table, four ODF methods from both libraries that have similar names and implementations are summarised.

Table 1: Some ODF methods from Essentia and Madmom

Essentia ODF Method name	**Madmom ODF Method name**	**Comment**
hfc	high_frequency_content	Detects percussive events based on the High Frequency content spectrum
complex	complex_domain	Considers significant energy changes in Magnitude and deviations in Phase
complex_phase	phase_deviation	Considers phase changes weighted by magnitude. Good for bowed string instruments
flux	spectral_flux, spectral_diff	Spectral Flux9 Detection function. The difference in two consecutive frames of the magnitude spectrum, using L2-norm [8]or L1-norm [9].

[6] https://github.com/CPJKU/madmom/blob/7e560700987d2800f1c78fdb2a5faba77ecece0d/madmom/features/onsets.py#L585

[7] https://essentia.upf.edu/reference/std_OnsetDetection.html

[8] https://github.com/MTG/essentia/blob/master/src/examples/tutorial/example_onsetdetection.py

[9] https://en.wikipedia.org/wiki/Spectral_flux

Another method not included in Madmom/Essentia libraries is the Wavelet method described in Bello's tutorial paper [10]. It places focus on precise time localisation, an important aspect of bass rhythm, however it did not score as high as the Spectral Difference, Phase Deviation or High Frequency Content in accuracy metrics. Bock and Widmer introduced the SuperFlux method in a paper titled "Maximum Filter Vibrato Suppression for Onset Detection" [11] and it is implemented in the Madmom libraries. Music Critic uses the 'superflux' method as a parameter in the SOP algorithm and this has worked well for detecting guitar onsets. The challenge of handling vibrato is also a theme in offsets and is dealt with in a later section.

Peak Picking

Different strategies can be employed varying from the "PitchSalienceFunctionPeaks" used in the MELODIA[12] algorithm to Bello's [13] "Peak Picking" formula which calculates the dynamic threshold as the weighted median of a section of the kurtosis around the audio frame:

$$\delta_t(m) = C_t \operatorname{median} \gamma_2(k_m), k_m \in [m - \frac{H}{2}, m + \frac{H}{2}]$$

Equation 4 Peak Picking Formula

where *C_t* is a predefined weighting value. In a later paper [7], the median filter is used to obtain an adaptive threshold as follows:

$$\delta_t(m) = \delta + \lambda \operatorname{median} \eta(k_m), k_m \in \left[m - \frac{H}{2}, m + \frac{H}{2}\right]$$

Equation 5 Peak Picking Formula (adaptive)

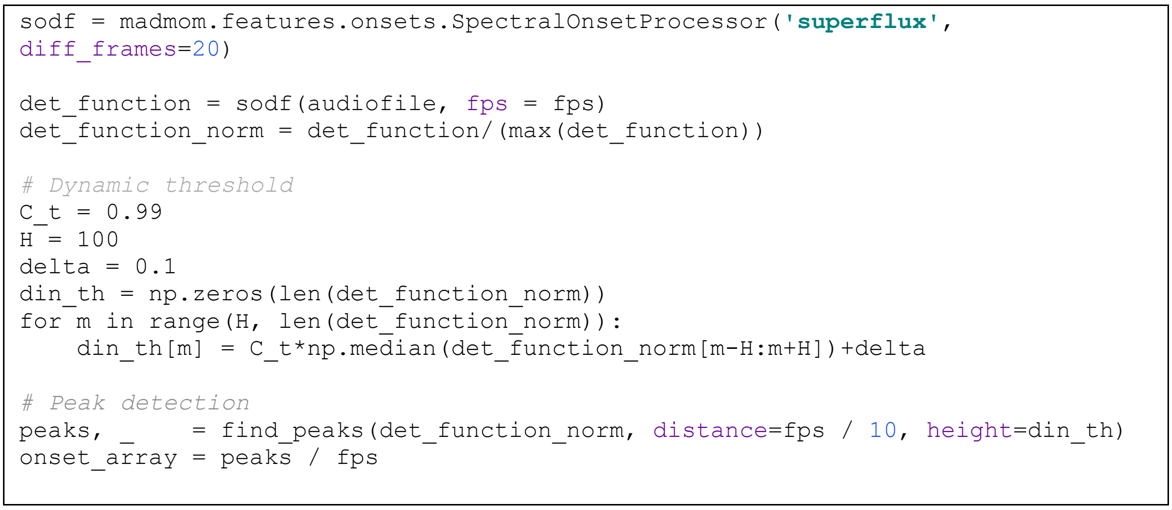
Where delta(δ) and lambda(λ) are constant values. Ramon Romeu[10] implemented the formula (but using *C_t* instead of λ)with the following values for *C_t*, *H* and *delta* determined in sweep experiments for his work isolating saxophone notes.

C_t = 0.99

H = 100

delta = 0.1

[10] https://github.com/RamoonRoomeu/ToneExperiments



```
sodf = madmom.features.onsets.SpectralOnsetProcessor('superflux',
diff_frames=20)

det_function = sodf(audiofile, fps = fps)
det_function_norm = det_function/(max(det_function))

# 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

# Peak detection
peaks, _    = find_peaks(det_function_norm, distance=fps / 10, height=din_th)
onset_array = peaks / fps
```

Figure 4: Peak Picking implementation

The values given for C_t, H, delta can be customised for different Dataset tracks. A sweeping method can then be used to optimise these values for a given musical piece.

2.3. Offset Detection

While onsets have been well researched by Bello [10] and present clearly defined points in time for pitched percussive instruments, offsets present more challenges. These include:

- perception of offsets
- musical considerations
- identification of stable voiced regions

2.3.1 Perceptual considerations for offsets

For pitched percussive instruments such as the bass, defining the “cut-off” point for an offset is more difficult than identifying on onset. For sounds in general, Terhardt [14] showed that the subjective duration of pure 200Hz tone is less than that of a 1kHz tone. Ha also showed that the perceived duration of a low tone can be increased when more harmonics are added.

For perception we are concerned with “just noticeable difference” or JND when it applies to note length. The question is whether there is sufficient duration to allow an offset response. Kopp-Scheinpflug [15] considers gap-detection and the limits of human auditory temporal acuity which varies from **2/3ms to 30ms** depending on the level of spectral disparity in the signal. This would suggest a more appropriate value of about 12.5ms compared to an onset time window of 50ms [16] used typically in transcription applications. For songs such as “Brown Eyed Girl” by Van Morrison, 12.5ms is almost equal to the shortest inter-onset interval in the song.

2.3.2 Musical considerations for offsets

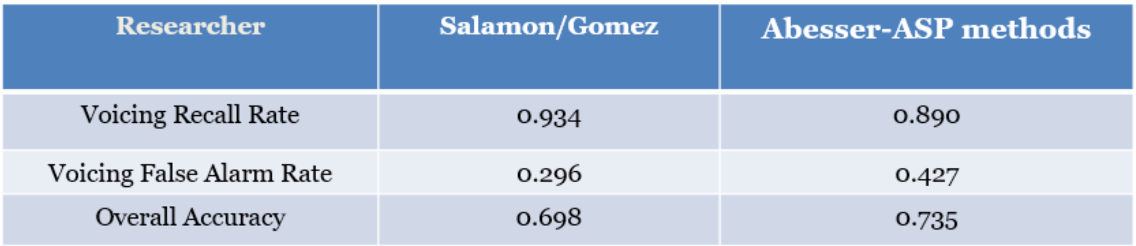
Note duration is the difference between onset and offset and this is the key metric that needs to be identified for student feedback. An offset (exit point) cannot exist without an onset (entry point). For offsets, the “exit point” depends on the playing technique.

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 means the offset aligns with onset of the next note. A “dead note” style could result in such a short duration that the pitch is barely perceptible. A “sustain” style where the offset position of a given note exceeds the onset of the subsequent note is not considered (i.e., letting an open string ring, then plucking another). In short, different styles suggest different measurement techniques need to be applied for different offset types.

2.3.3 Voicing Classification

Voicing Classification is an algorithm that determines whether a sound is pitched or not. It is a pre-processing stage for fundamental frequency (F0) extraction. The voicing recall statistics for MELODIA [12] techniques compare more favourably against Jacob Abesser’s signal processing methods [17] for bass.

Table 2: Abesser Signal Processing vs MELODIA[11]



Researcher	Salamon/Gomez	Abesser-ASP methods
Voicing Recall Rate	0.934	0.890
Voicing False Alarm Rate	0.296	0.427
Overall Accuracy	0.698	0.735

2.2.3 Technical Discussion of Offsets

In this section the relevant background concepts to offsets (fundamental frequency detection, segmentation, and transcription) are described. For voiced segments of an audio file, the next key component is the estimation of the fundamental frequency (F0), in Hz or in cents. For timing considerations, we are not interested in the actual frequency, but the beginning and end of the segment with a given F0. Detection of F0 facilitates the segmentation process for music transcription and it can also be applied to our performance assessment purposes. One example of F0 detection is the Two-Way Mismatch (TWM) algorithm used sms-tools[12]. It starts with F0 candidate choices based on peaks of the magnitude spectrum within a specific range of frequencies. Its current implementation has challenges in finding vibrato-resistant stable regions and does not consider sub-harmonics of peak frequencies. Abesser developed a F0 Tracking function with code available[13] in the pymus libraries. His research on Offset detection [17] considers the problem of Vibrato and is tested with example of this the IDMT dataset [18] in Chapter 3. The full matlab/python code Abesser used for calculating onsets and offsets is closed source and has not been tested in this thesis.

The “PitchMelodia”[14] function from the Essentia libraries is based on the MELODIA [12] algorithm. A customised version of “PitchMelodia” is the “PredominantPitchMelodia” function which can work with polyphonic music signals. Both are based on four algorithms that are called in a chain.

[11] Abessers overall accuracy is higher at 0.735, but that metric considers pitch which is not of concern here.
[12] https://github.com/MTG/sms-tools/blob/master/software/models/harmonicModel.py
[13] https://github.com/jakobabesser/pymus/tree/master/pymus/sisa/f0_tracking
[14] https://essentia.upf.edu/reference/std_PitchMelodia.html

Table 3: MELODA Algorithm steps

Function Name	**Input**	**Output**
PitchSalienceFunction	Spectral Peaks of Audio	Pitch salience function.
PitchSalienceFunctionPeaks (like SOP peak picking)	Pitch salience function.	Bins Saliences
PitchContours	Bins Saliences	Contour Bins Contour Saliences Contour Start Times Duration
PitchContoursMelody	Contour Bins Contour Saliences Contour Start Times Duration	Pitch Pitch Confidence

The tutorial example[15] shows how the individual algorithms and parameters choices (min. and max frequency, voicing tolerance, maximum number of peaks) are employed to estimate the frequency contour of a given audio. The algorithm "PitchContour" has two important parameters for capturing timing values:

- timeContinuity: the maximum allowed gap duration for a pitch contour and is defaulted to 100ms
- minDuration: the minimum allowed contour duration

2.4. Neural Network Strategies

Current State of the Art for MIR research focuses more on machine learning / deep learning techniques and onsets/offset detection are no exceptions.

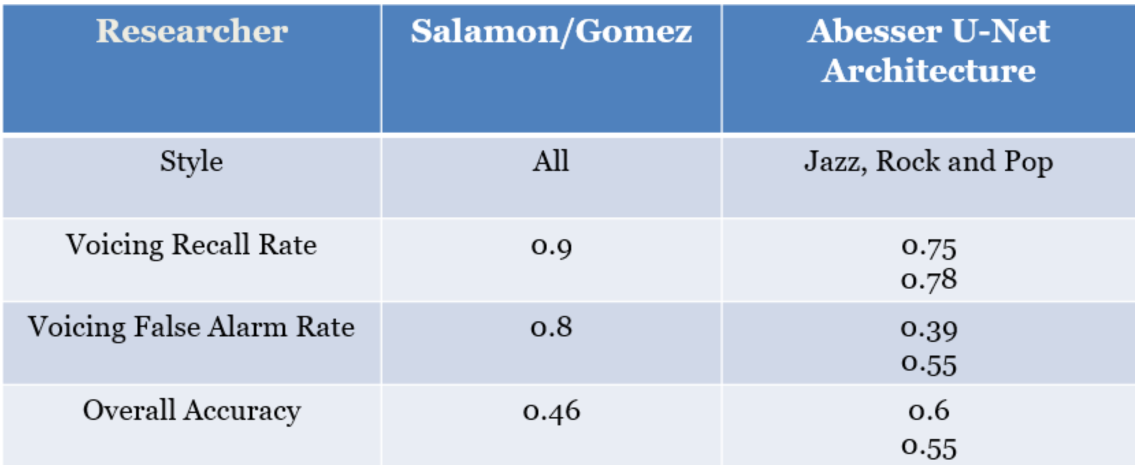
[15] https://github.com/MTG/essentia/blob/master/src/examples/tutorial/example_predominantmelody_by_steps.py

Bass U-Net[16] is based on Abbesser-Müller Data driven algorithms [19]. It uses a CNN (Convolutional Neural Network) Streamlined Encoder/Decoder Architecture for Melody Extraction. It has been tested on the following Datasets:

- Real World Computing (RWC) [20]
- MDB-bass-synth [21]
- Weimar Jazz Database (WJD) [22]

The MELODIA algorithm [12] by Salamon/Gomez as well as producing better voicing recall rate than Abessers [17] F0-detection-based techniques has also produced better recall rates against Abessers-Müller`s Bass U-Net as shown in following table However, Bass U-Net[17] produced lower false alarms.

Table 4: Abessers-Müller Data Driven algorithms vs Salamon/Gomez



Researcher	Salamon/Gomez	Abesser U-Net Architecture
Style	All	Jazz, Rock and Pop
Voicing Recall Rate	0.9	0.75 0.78
Voicing False Alarm Rate	0.8	0.39 0.55
Overall Accuracy	0.46	0.6 0.55

The Voicing accuracy rates still fall short of the requirements for effective F0 extraction for Music Education applications, but there is a trend towards data driven techniques.

Bocks Online Onset Detector[18] is a universal onset detector with BLSTM based on current neural networks and was trained with music mixtures including R&P genre. The main obstacle in all data driven approached in going forward is training using the actual audio used in the Performance Assessment music exams and to drive up the "F measure" to above 90%.

[17] https://github.com/jakobabesser/bassunet

[18] https://github.com/CPJKU/madmom/blob/master/bin/OnsetDetectorLL

2.5. Score and Instrument Level Parameters

In the section on Offset Detection, Abesser's ASP methods [23] were introduced in terms of their accuracy for identification of voicing section. Besides aiming the research at score level parameters (onset, time pitch and duration), he also used Feature Selection to extract the plucking and expression style from single note recordings. In the context of R&P performance assessment, this technique could be applied to validate if a student is using the correct technical parameters in a recording. In the single note recordings of the IDMT dataset, you can hear the sonic difference in between plucking styles: e.g. (FS) finger style plucked note and a plectrum picked note (PK). He also introduced five different expression styles: normal (NO), vibrato (VI), bending (BE), harmonics (HA), and dead note (DN). He also made possible the estimation of the additional instrument-level note parameters string number, fret number. The proposed system achieves accuracy above 0.88 for both left(expressive)- and right-hand (plucking) techniques

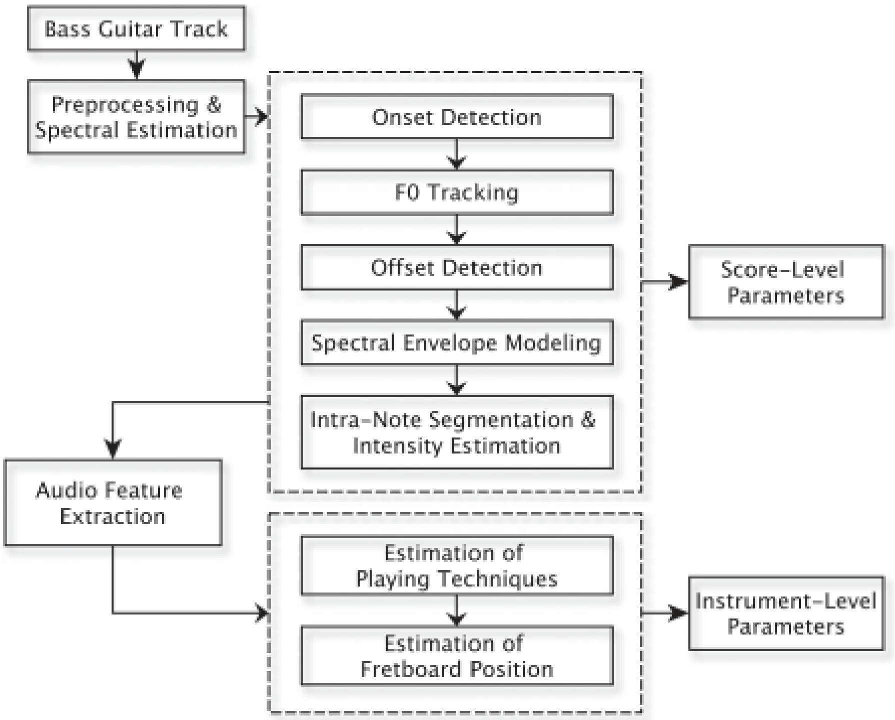

Bass Guitar Track
Preprocessing & Spectral Estimation
Onset Detection
F0 Tracking
Offset Detection
Spectral Envelope Modeling
Intra-Note Segmentation & Intensity Estimation
Score-Level Parameters
Audio Feature Extraction
Estimation of Playing Techniques
Estimation of Fretboard Position
Instrument-Level Parameters

Figure 5: Processing Flow chart of the instrument level extraction process

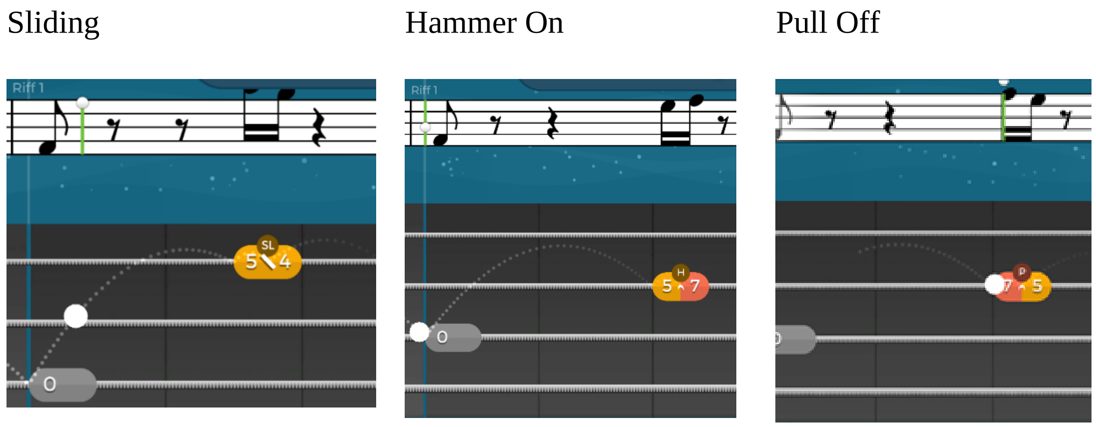
(Borrowed from [17] Instrument-Centered Music Transcription of Solo Bass Guitar Recordings, Jakob Abeßer and Gerald Schuller, Senior Member, IEEE

Reboursiere [24] also conducted research on detecting left and right-hand techniques for guitar using a hexaphonic piezo guitar-pickup to assist music students.

Let us examine the score and instrument parameter extraction flow chart in fig 5. The first three blocks of the Score-Level Parameters deal with fundamental frequency tracking to calculate onsets and offset. The constant-Q transform[19] and Instantaneous Frequency Spectrograms were proposed to get better resolution for the lower frequency bands. The Intensity Estimation part segments a note into Attack and Decay parts. By examining the framewise features over the duration of the Attack and Decay sections of a bass note, the timbral qualities are captured, helping estimate fretboard position and playing styles. Identification of wide band characteristics in the attack part of the magnitude spectrum can indicate Slap-Pluck (SP) or dead notes (DN). The frequency resolution of 10 bins per semitone allows the capture of small F0 modulations that indicate :

- Vibrato (VI) depending on its periodicity
- Bending(BE) if there is a big increase and decrease in F0
- Slide (SL) if there is an initial stable pitch followed by continuous increase or decrease in pitch.

In the context of state-of-the-art tools on the market, Yousician detects one of the audio features just described: Slide (SL). Additionally, it also detects Hammer On (HO), Pull Off (PO).



Sliding

Hammer On

Pull Off

Figure 6: Yousician Detecting slides, hammer-ons, pull-offs on 16th notes.

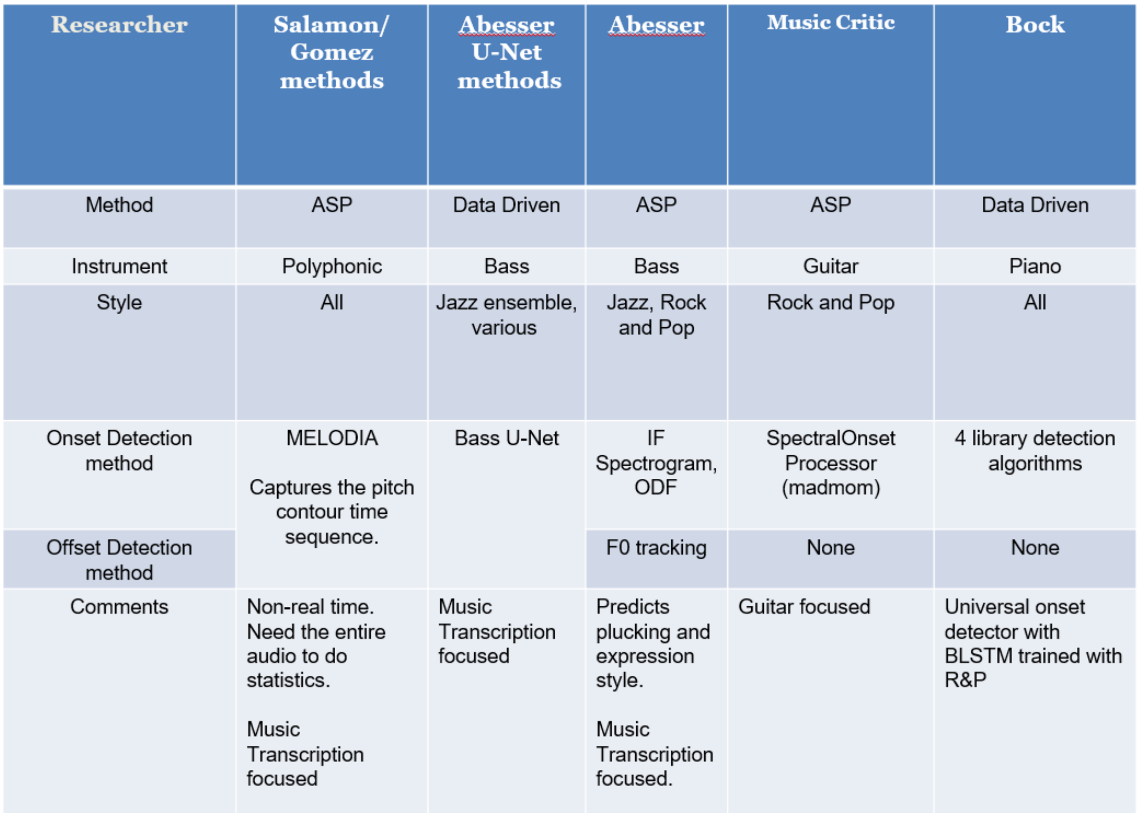
[19] https://essentia.upf.edu/reference/std_ConstantQ.html

Yousician provides real-time recommendations of what finger to use when holding a particular fret position but doesn't give feedback on this. Overall, the state of the art research and tools can classify the use or non-use of certain technical parameters and predict the use of correct string and fret position. All of this can assist in classifying the technical criteria of music performance.

2.6 Summary

The following table summarizes the state of the art of algorithms used for onsets and offsets, many of which use these library functions.

Table 5: Algorithms for Onset/Offset detection



Researcher	Salamon/ Gomez methods	Abesser U-Net methods	Abesser	Music Critic	Bock
Method	ASP	Data Driven	ASP	ASP	Data Driven
Instrument	Polyphonic	Bass	Bass	Guitar	Piano
Style	All	Jazz ensemble, various	Jazz, Rock and Pop	Rock and Pop	All
Onset Detection method	MELODIA Captures the pitch contour time sequence.	Bass U-Net	IF Spectrogram, ODF	SpectralOnset Processor (madmom)	4 library detection algorithms
Offset Detection method			F0 tracking	None	None
Comments	Non-real time. Need the entire audio to do statistics. Music Transcription focused	Music Transcription focused	Predicts plucking and expression style. Music Transcription focused.	Guitar focused	Universal onset detector with BLSTM trained with R&P

2.7 Observations

One can observe that applications on the market, despite having impressive front ends, they fail to measure duration and permit "imposter notes". The choice of suitable machine learning techniques to extract the most relevant parameter is constrained by the limited number of annotated student performances. Source Separation techniques based on the Spleeter model [25] have been successfully applied to the songs from Rock and Pop repertoire and could be used to create a large dataset if the recorded examinations were made available. Scaling up would mean wider ML options, like the Support Vector Machines used in the style analysis [17] discussed, but that was focussed on single note samples, which meant it was easier to scale up to large numbers of recordings.

For music education, the best way forward would be to build an additive Dataset while experimenting. This means a large initial experiment is performed to build a trial performance assessment system based on Linear Regression. At the end of a trial course, advanced students submit their recordings along with their own "self-predicted grades", this could build new data set that could be used for training a new database for the Bocks Online Onset Detector[20] or Abesser's Bass U-Net[21].

2.8 Machine Learning Techniques in related areas

Giraldo [4] used a variety of ML algorithms Support Vector Machines (SVM), k-NN, and artificial Neural Networks) to induce duration, energy and embellishment transformation models. Expressive performance modelling (EPM) also requires data acquisition that Music Education applications require. EPM uses descriptors to model notes (attack level, sustain duration, amount of legato with previous note). The aim of EPM is different: it wants to model and induce different deviations and add "imposter notes" whereas Performance Assessment wants to catch and capture these deviations from real performances and grade them accordingly.

For music education, segmenting an audio stem into individual notes as is done in EPM, would be computationally expensive for any song over a minute. On the other hand, this would not accurately model the essence of the overall performance. Data augmentation is one way to produce more data, but this would involve taking different student performances and adding slight modifications of onsets and offset in different areas with DAW (Digital Audio Workstation) tools or with scripts. Although it could result in some artificial sounds, the increased data would allow the testing of different ML strategies.

[20] https://github.com/CPJKU/madmom/blob/master/bin/OnsetDetectorLL

[21] https://github.com/jakobabesser/bassunet

The next chapter will show examples from the IDMT [18] dataset and the R&P Repertoire that will be used for evaluation.

3. Datasets

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

3.1. IDMT BASS SINGLE TRACK

The IDMT Single Track Track dataset [18] 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 relevant for expression styles [17] as shown below. The annotated onset and offset values shown below (in seconds) are extracted using functions inside the scripts (see Appendix 12.1) and are used as ground truths for evaluating the algorithms presented in the State of the Art.

```
<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 7: Extract from IDMT Dataset. File 002.xml

Notice the expression style is Normal (NO). The Vibrato (VI) example discussed in previous chapter would require non-zero values for modulation frequency and modulation frequency range.

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 not used for Student performances in this thesis. It is available under a creative commons licence.

IDMT (from 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

IDMT-Example on note length: 002.wav

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

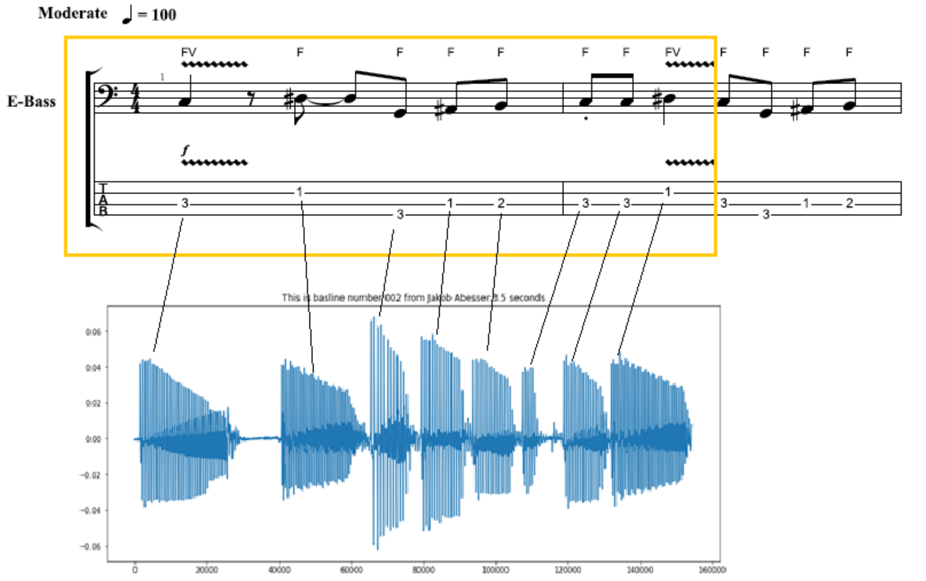


Figure 8: Note duration of Staccato notes

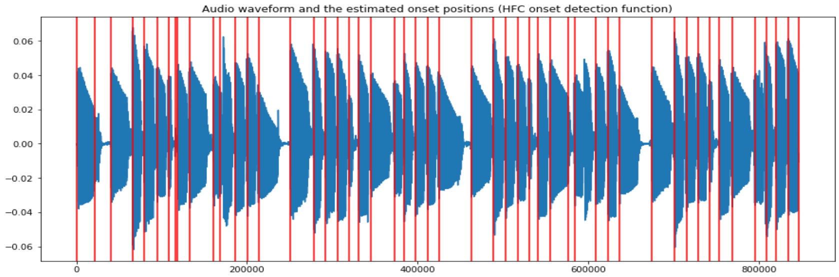


Figure 9: Audio 002.wav (IDMT DATASET) with onsets using HFC method

By examining the musical properties of the IDMT dataset of 17 songs, 4 songs were shortlisted for performing the formal evaluations of the signal processing algorithms for onset and offset detection:

- 002.wav
- 004.wav
- 010.wav
- 012.wav

The criteria of track selection were based on note the minimal use of semi-quavers, finger style plucking and overall complexity level suitable to lower to medium level grades for music academy exams.

3.2. Rock and Pop repertoire

Six songs with increasing levels of difficulty were chosen from the TCL London Rock and Pop Gradebooks.

Table 6: TCL dataset

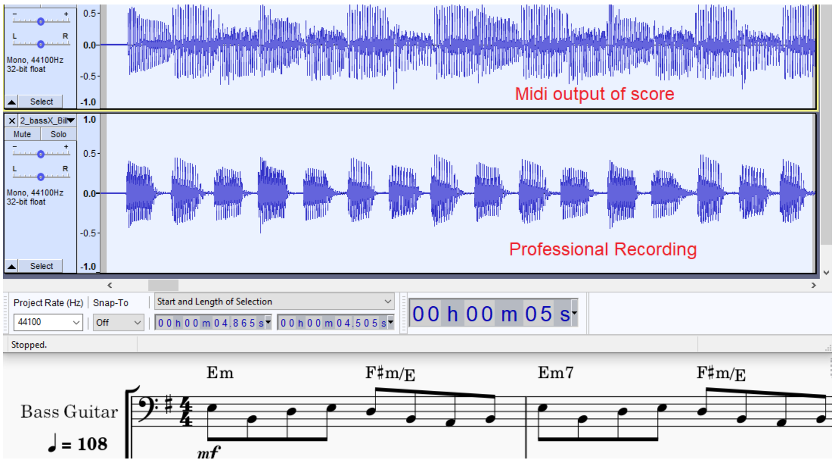
Song Nr	Level	Song	Artist (original)	Shortened Name	Length
0	Grade 0	Yellow	Coldplay	yellow	93
1	Grade 1	Billie Jean	Michael Jackson	bjean	55
2	Grade 1	Just Looking	Stereophonics	just	86
3	Grade 2	Brown Eyed Girl	Van Morrison	brown	64
4	Grade 3	Roadrunner	Junior Walker and the All Stars	road	82
5	Grade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

The shortened names in column 4 shall serve as references to the songs. The “Length” columns indicate the duration of the song from the beginning to be considered for grading.

The Dataset was constructed for all six songs with the following attributes

- Song Descriptions based on information in the printed score
- Professional reference performance
- Onset/offset annotations of the reference bass recording
- Multiple Student recordings of each song
- Mixes of the Student Recordings with the given Minus 1Audio
- Music Teacher Grades (Numeric and Descriptive)
- Algorithm generated data on the student and reference stems.

The Onsets and Offset were initially done by visual inspection in Sonic Visualizer22. This was improved by running the best performing onset and offset algorithms on the reference, removing the false positives, and adding the missing onsets. This resulted in a much more consistent set of annotations on the reference. A midi rendering of the score can check for song description elements that are not shown in the score. For “bjean” it is required to pluck with “short/jerky movement” in the verse/chorus, but no dotted notation is shown, so generating a MIDI output will not reflect desired playing style.



Project Rate (Hz): 44100
Snap-To: Off
Start and Length of Selection: 00 h 00 m 04.865 s 00 h 00 m 04.505 s
00 h 00 m 05 s
Stopped.

Figure 10: Billie Jean: Midi Rendering vs Human recording (shorter notes)

The first function of the R&P dataset is to validate the state-of-the-art algorithms after testing with the IDMT dataset. The second function is to provide a basis to measure student performances.

22https://www.sonicvisualiser.org/

The reference bass recording is assumed to represent a grade of 100% in timing and rhythm. Table 7 summarizes the individual musical features of each of the tracks, with their Grade Level (difficulty level) and technical description included [26]. These are the topics of the Grading Sheet questions given to the Bass teacher.

Table 7: Tech. Control Parameters for 6 TCL songs

Song	Syncopation	Repetition	Dynamics	Articulation	Note Length
Yellow **Grade 0**	chorus	No rushed feel			Play evenly (verse)
Bjean **Grade 1**	Accent just before chorus		Leading into chorus	separate jerky quavers+ smooth, melodic material	
Just **Grade 1**	Chorus: accented, syncopated motif., hard accent		Unexpected subito p at bar 25.		
Brown **Grade 2**					different note lengths and rests
Road **Grade 3**				Tenuto (underscore) loud on beat 1	
Wotm **Grade 3**	syncopated repeated notes				Correct separation

In the experiment, truncation is applied to the songs to make grading faster and easier, for example "yellow" was truncated to remove the repeated verse, since no new musical features were introduced.

3.3. Musical Properties of R&P repertoire

Two songs, “wotm” and “bjean” from the dataset were given more attention because they had interesting note-gap and note length differences between verse and bridge. The “wotm” verse has a long note duration with zero inter-note gaps while the bridge has the opposite: shorter “reggae” notes with some noticeable inter-note gaps. The BPM of this song is 146 and for Billie Jean it is 108. The Grade1 “Billie Jean” song also manifests similar contrasting sections.

4. Methodology

Music Education demands more precise timing measurements than state of the art automatic music transcription. Onset detection for bass requires a different approach from current assessment technologies in three key areas. First, as a rhythm section instrument the bass plays a key role in synchronizing with the drum pattern, so a tighter precision is required than the one used for guitar. 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. Finally, the choice of onset detection algorithm needs to include an offset detection algorithm to measure duration.

The basic workflow was to check the algorithms from the summary table of State of the Art in Chapter 2 with starting with IDMT Dataset then moving onto R&P Repertoire Dataset. The circular workflow would allow continuous improvement of accuracies, by combining algorithms, checking algorithms on different dataset and different song sections in the R&P Dataset. The goal is to find the best numerical measurement of onset and offset.

Initial tests used the standard MIR evaluation of 50ms. This value was for measuring onset accuracy in guitar onsets and was increased to 200ms for offsets due to difficulty in handling smoothly decreasing note envelopes [27]. However, when reducing this to 20ms, the accuracy measurements dropped as expected and this required further iterations of algorithm improvement. It was expected that a hybrid algorithm that would apply different threshold parameters to different songs and different offset strategies to different sections of a given song would be required.

Getting good Precision, Recall and F-Measure accuracy values for the professionally recorded reference was only half the battle. When gathering and testing the student recordings it proved very difficult to obtain a precision value higher than 60%.

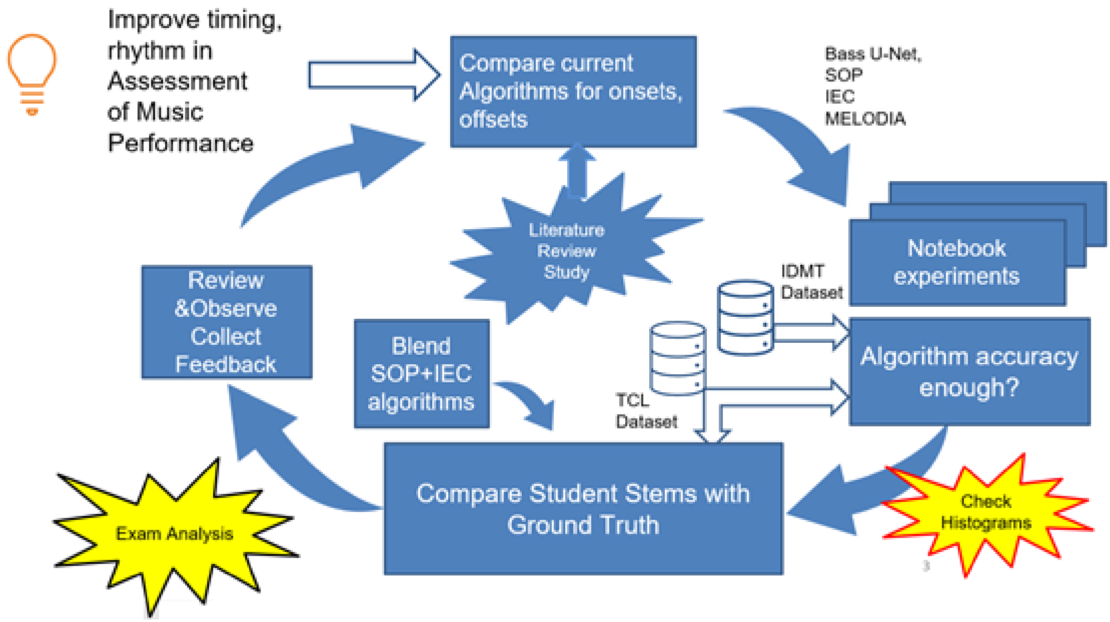

Improve timing, rhythm in Assessment of Music Performance
Compare current Algorithms for onsets, offsets
Bass U-Net,
SOP
IEC
MELODIA
Literature Review Study
Notebook experiments
IDMT Dataset
Review &Observe Collect Feedback
Blend SOP+IEC algorithms
TCL Dataset
Algorithm accuracy enough?
Compare Student Stems with Ground Truth
Check Histograms
Exam Analysis

Figure 11: Strategy to evaluate, test and find the best accuracy

The next chapter tests different onset and offset detection strategies with the Datasets discussed in Chapter 3.

5. Algorithm Development and evaluation

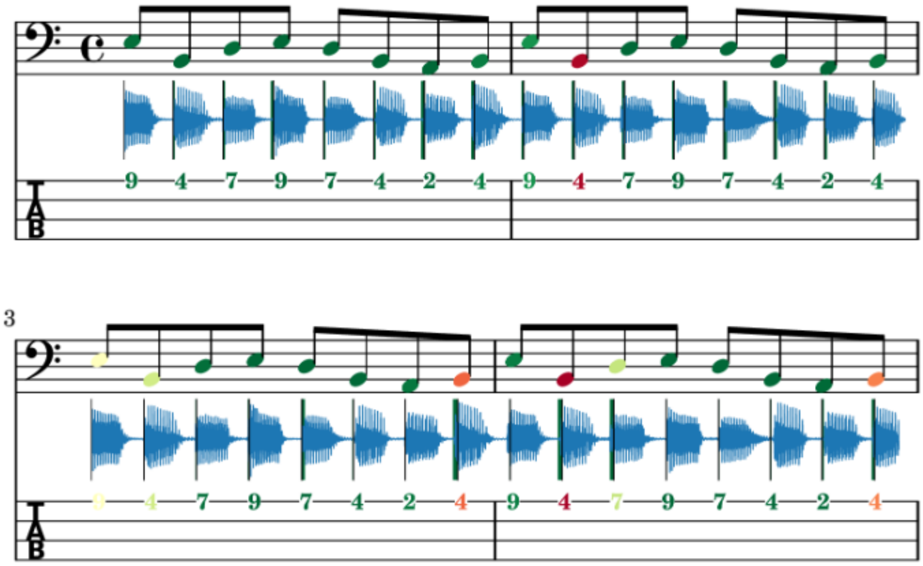
The objective in this chapter is to find out how far the algorithms are from the ground truth. It opens with the end-to-end test on a R&P repertoire song using the current State of the Art method used in the Music Critic.

5.1 Music Critic End-to End test

In Music Critic, a JSON file is used to mark the overall duration of the songs the beat locations in time. The Lillypoint (LY)[23] file marks the note pitch, its length and its location in the overall on pattern. To perform the End-to-End test for Billie Jean, the following steps were executed:

- Calculation of Beat positions using Madmom.
- JSON, Lilypondfile preparation for Billie Jean.

To reproduce there results you must execute the python script “demo_bass.py” in a Music Critic repository with private access. In the same folder there is a subfolder called “data/bass”. Inside there you have the JSON file (l2ex1.json) and Lillypond (l2ex1.ly). Running the “demo_bass.py” generates PNG file with appearing as follows:



[23] http://lilypond.org/

Figure 12: End to End test of Music Critic with Billie Jean GT

The alignment of the ground truth stem was not aligned 100%. The outcome of this experiment is that the Spectral Onset Processor (SOP) as its currently parameterized in Music Critic needs to be improved. It uses the *SuperFlux* method described in the state of the art. One possible explanation might be that the peak picking formula is configured to consider multiple onsets that can occur with guitar chords. Another speculation might be that detection methods behave differently for the lower frequency range. In any case, the tools provide by Ramon in the state of the art, do give us a mechanism to customize the adaptive threshold formula by running sweeping experiments to choose the optimal values for *C_t*, *H* and *delta.*

5.2 Algorithm Evaluation

The algorithms proposed are either "paired" or "non- paired". Non paired means capturing the onset without regard to monitoring the end of the note, for example the SOP function. The paired approach measures start and stop time of a musical note within an audio frame. The algorithms are also classified as ASP (audio signal processing) or Data Driven

So given the current state-of-the-art strategies:

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

a fifth strategy is added for evaluation and comparison

- IEC: IndexedEnergyChecker (Paired ASP driven)

For accuracy measurements, Precision and Recall, and F-measures were made on the Onset measurements. The definitions are as follows:

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

$$precision = \frac{TP}{TP + FP}$$

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

$$recall = \frac{TP}{TP + FN}$$

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

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Deviation measurements were made on the difference between the Ground Truth and Measured Value. From this, Histograms are derived following methodology used in [28]. The Mean Absolute Error , Absolute Mean Error Standard Deviation were calculated on the following

- Onset deviations
- Duration deviations

5.3 Indexed Energy Checker (IEC)

Originally a simple Energy Checker function called “calculateOffsetOnset()” was developed to capture RMS band values of an input signal. 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 Superflux method, but these values were independent of the offset values and could not be paired.

The IndexedEnergyChecker[24] (IEC) algorithm, derived from combining functions in the song assessment python notebooks, pairs the onsets and offsets, based on an Energy Threshold. Depending on the energy threshold level set, it decides on how to split the wave boundaries into “Sound Islands”. The splitting is represented by set of start and stop indices representing onset and offset. This threshold parameter was uniquely configured for each song.

[24] https://github.com/RamoonRoomeu/AutomaticAssessmentSax/tree/main/Training%20and%20Experiments

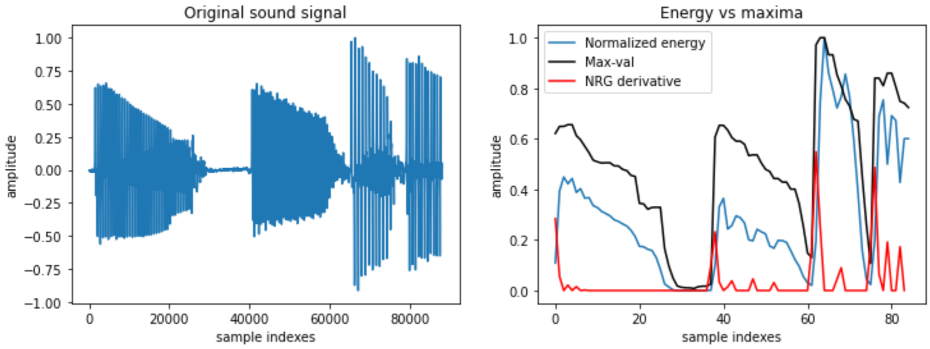


Figure 13: Normalized Energy of Audio 002.wav sample (IDMT DATASET)

Energy is calculated using the normalise energy function

$$energy = \sum_{n=0}^{N-1} |x[n]|^2$$

Equation 6 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.

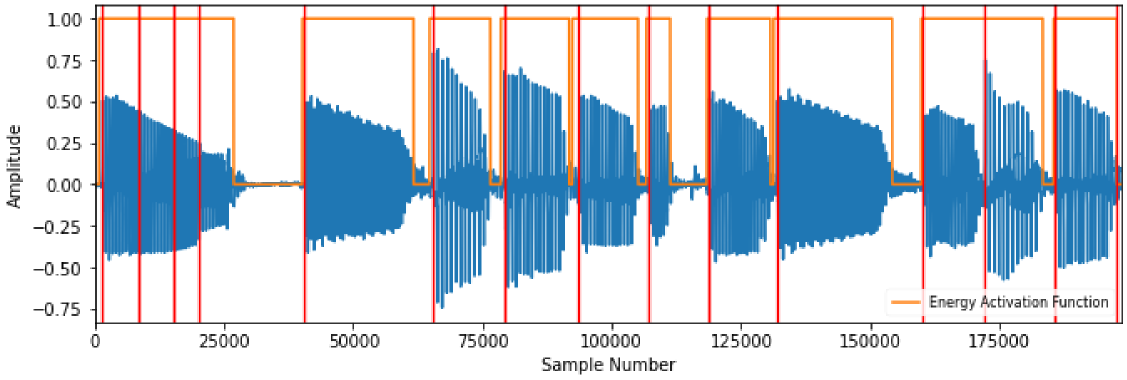


Figure 14: Sound Islands of Audio 002.wav (IDMT). FS = 1024, HS = 512.

The red lines in figure 14 show that the onsets detected by the SOP algorithm throw more false positives for long notes than the IEC.

Table 8 presents results Precision, Recall and F-measure for initial onset detection tests.

Table 8: Benchmarking Onset detection algorithms using IDMT Dataset

Researcher	**Indexed Energy Checker (IEC)**		**SOP Superflux**	
evaluation window	20ms	50ms	20ms	50ms
Precision	0.27	0.44	0.623	0.894
Recall	0.26	0.415	0.594	0.849
F-measure	0.264	0.425	0.607	0.869

Going from, 50 to 20ms window reduced the PRF for SOP dramatically. For IEC, it was problematic, because each audio track has different acoustic characteristics so one threshold value of 0.05 does not fit well with all WAV files. Although it was possible to locally optimise the threshold for IEC and get better results, more effort was focused on the R&P dataset in the next section.:

To reproduce the results for IEC and SOP run the script[25]

python3 abesser_test.py

Set *matching_window_size* = 0.05 (ms) or 0.02 (20ms)

The results for the Bock online detector are not shown because this algorithm was discovered at a later stage in the Thesis and was tested with the R&P repertoire. The experiment also extracted Precision, Recall and F-measure for offset, for the IEC, MELODIA [12] and Abesser U-Net algorithms [19], but they are not shown, since offsets depend on onsets.

[25] https://github.com/cvf-bcn-gituser/bass-critic/tree/main/tclEvaluation/abesser_test.py

5.4 Accuracy Metrics: IEC vs MELODIA for R&P

5.4.1 Long Notes

The songs “just” and “wotm” were interesting candidates for containing long notes in different sections of their songs. The relationship between the “long note-short gap” pattern and the low PRF score for IEC was observed. They were tested with MELODIA [12] to see there were better accuracy alternatives for algorithms that could detect both onset and offset. For the MELODIA algorithm the min. and max. frequencies were chosen as 25 and 340 Hz respectively. For the song “wotm” the timeContinuity was set at 14ms and the minDuration set at 112ms.

Table 9: Selected comparison: IEC vs MELODIA

Song	**Precision**		**Recall**		**F-measure**	
	IEC	**MELODIA**	**IEC**	**MELODIA**	**IEC**	**MELODIA**
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

5.4.2 WOTM Bridge and Verse Comparison

WOTM Verse-IEC Algorithm

Since there was no improvement, I broke the song “wotm” into components to see which song sections could be used for the paired technique (detecting both onsets and offsets). 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

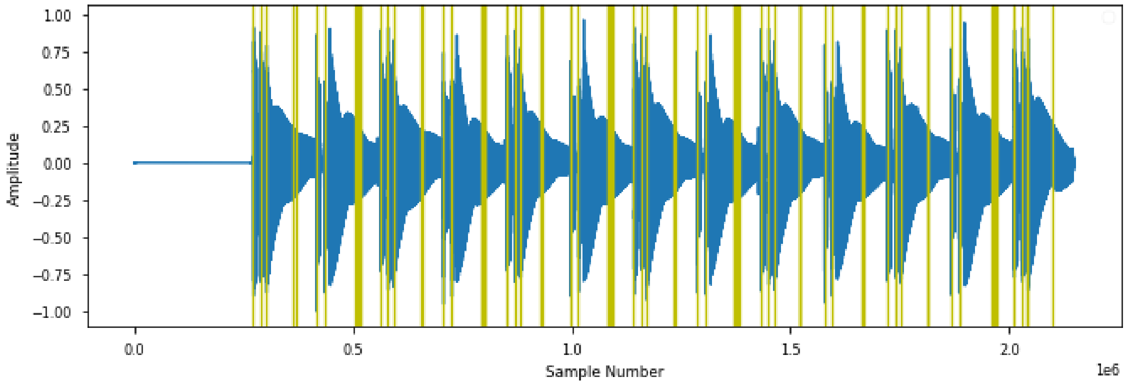


Figure 15: IEC onsets for “wotm” verse

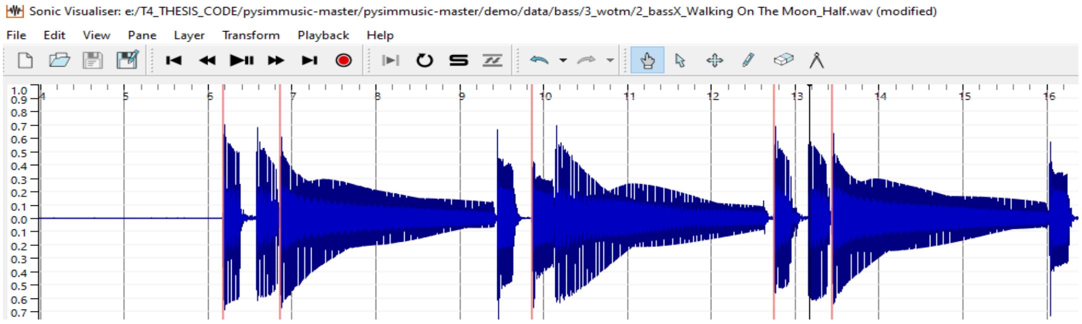

Sonic Visualiser: e:/T4_THESIS_CODE/pysimmusic-master/pysimmusic-master/demo/data/bass/3_wotm/2_bassX_Walking On The Moon_Half.wav (modified)
File
Edit
View
Pane
Layer
Transform
Playback
Help
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure 16: Section of “wotm”verse showing missed IEC onsets.

WOTM Bridge -IEC Algorithm

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

P =0.937 R= 0.949 F=0.943 (With last note)

P =0.961 R= 0.948 F=0.954 (Without last note)

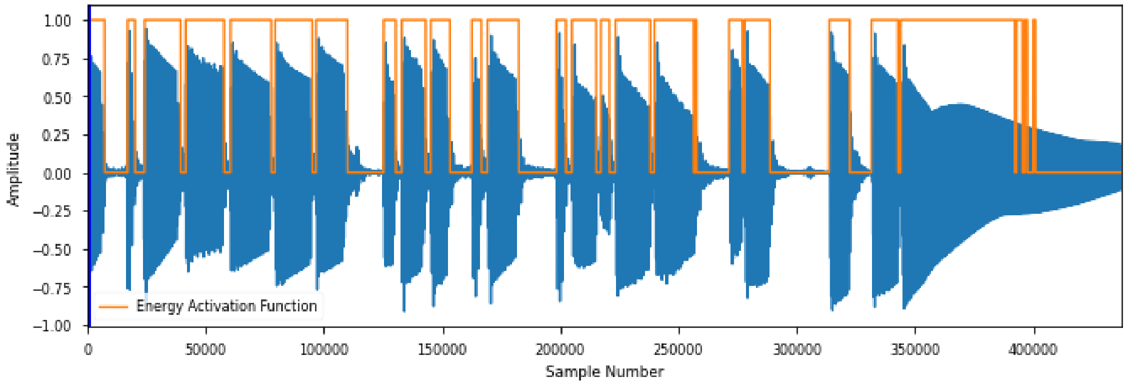


Figure 17: Calculated IEC Sound Islands for WOTM bridge, P = 0.937

WOTM Verse -MELODIA Algorithm

This song has the longest note length of all the tracks. With Threshold = 0.06, we still see a lot of hysteresis on the last sustained note in Figure 17 and this kicks 4 percentage points off the precision.

In contrast the MELODIA [12] algorithm performed a lot better in the "wotm" verse with PRF measures of P = 0.974, R = 1.0 and F = 0.987.

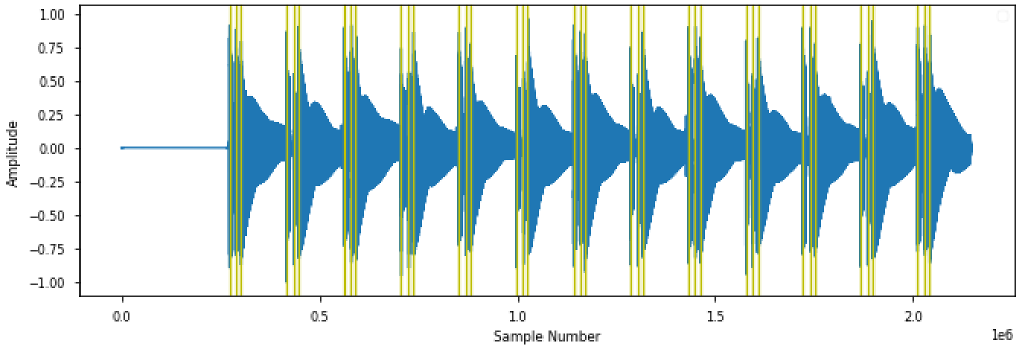


Figure 18: PitchMelodia derived Onsets for WOTM verse , P = 0.97

WOTM Bridge -Melodia Algorithm

But MELODIA [12] missed a lot of onsets in the bridge section of "wotm".

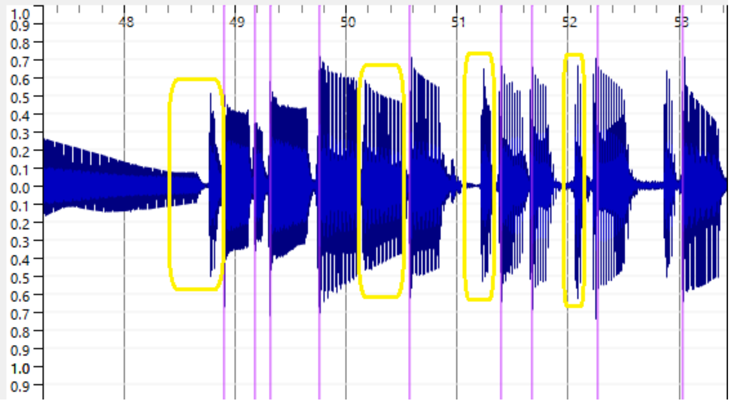
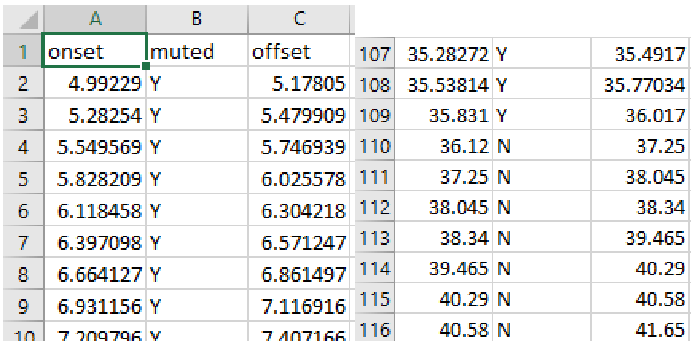


Figure 19: PitchMelodia derived Onsets for WOTM bridge

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 MELODIA [12] method in the bridge as seen in fig 19 marked in yellow.

5.4.3 Classifying Muteness

It is important to clarify our local definition of "mute". It does not relate to the concept of the left-hand muting technique described by Abesser [17]. 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. This introduced the concept of an annotation file called the "”<song>_rhythnm.csv" file with 3 columns: onset, muted and offset. Example extracts of "bjean_rhythm.csv" is shown below for the verse and the bridge sections



	A	B	C
1	onset	muted	offset
2	4.99229	Y	5.17805
3	5.28254	Y	5.479909
4	5.549569	Y	5.746939
5	5.828209	Y	6.025578
6	6.118458	Y	6.304218
7	6.397098	Y	6.571247
8	6.664127	Y	6.861497
9	6.931156	Y	7.116916
10	7.209796	Y	7.407166

107	35.28272	Y	35.4917
108	35.53814	Y	35.77034
109	35.831	Y	36.017
110	36.12	N	37.25
111	37.25	N	38.045
112	38.045	N	38.34
113	38.34	N	39.465
114	39.465	N	40.29
115	40.29	N	40.58
116	40.58	N	41.65

Figure 20: Sample of bjean_rhythm.csv verse (left) and bridge(right)

The middle column is marked Y when a slight muting of the string occurs. If the string is allowed to sound right up until the next onset, then the middle column is marked “N”. You can see in row 110 that the offset is equal to the next onset value in row 111.

5.5 Combined IEC-SOP for R&P

Four of the R&P songs song was configured use either SOP or IEC algorithms, or a combination of the two. For “bjean”, the verse was configured to use IEC, while the SOP was used for the bridge. For “wotm”, the verse was configured to use SOP, while the IEC was used for the bridge. The choice was made as follows: for “muted” sections the IEC was used whereas for legato section the SOP was used.

Table 10: Accuracy Metrics for 6 TCL references using IEC-SOP

Song	**Algorithms**	**Precision**	**Recall**	**F-measure**
yellow	SOP	0.985	1.0	0.992
bjean	IEC-SOP	1.0	0.997	0.998
just	SOP	1.0	0.988	0.994
brown	IEC	0.893	0.905	0.899
road	IEC	0.96	0.95	0.955
wotm	SOP-IEC	0.992	1.0	0.996

Data source for above metrics : StudentStatistics_<song>.csv[26] file [Check row 0 of each file]

Reproducibility: Run the command $python osets_song.py and make sure reference recording is in the audio directory.

To achieve these high accuracies, it was necessary to truncate the songs. Details of truncation points are described on a song by song basis in Chapter 6 “Prediction of Student Performance”

[26] https://github.com/cvf-bcn-gituser/bass-critic/tree/main/tclEvaluation/data

5.6 Accuracy Metrics for Neural Network Methods

Using a fork[27] of Bass U-Net was tested using the Mixed Genre and Jazz Training models ('bassunet_mixed.h5', 'bassunet_jazz.h5) the IDMT Dataset was tested for tested for PRF accuracies alongside “bjean”. The tables in the Appendix 12.9 show that the IDMT Dataset precisions range from 0.22 to 0.65. In contrast, “bjean” returned the following: precision, recall, f_measure_value values:
P = 0.815, R = 0.969, F= 0.885.

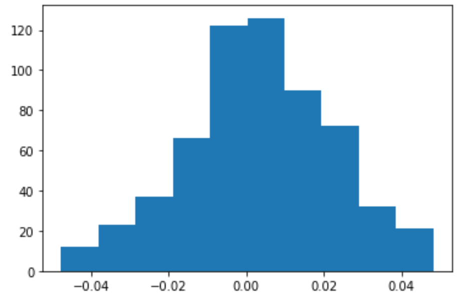
The high recall is interesting, in that it is higher than precision. Overall, these figures make it suitable for Student comparison. Reproducibility instructions are detailed in Appendix 12.9.

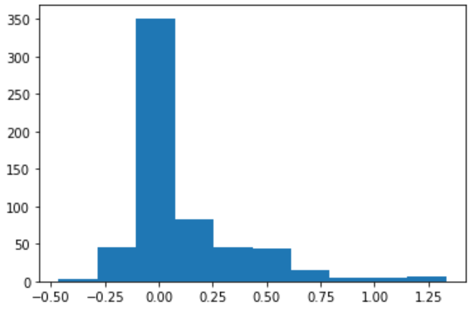
5.7 Deviation Metrics on Reference

The onset and offset deviation statistics show us how far the reference recording is from the ground truth. The following plots illustrate the deviation statistics for onsets for one IDMT Dataset file:

- MIR evaluation window size Onsets = 50ms
- MIR evaluation window size Offsets = 150ms.

IEC Measurements:

<table>
<tr>
<td>


</td>
<td>


</td>
</tr>
<tr>
<td>Onsets

Mean: 0.003599
Deviation from zero: 0.019611</td>
<td>Offsets

Mean: 0.109678
Deviation from zero: 0.282235</td>
</tr>
</table>

Figure 21: IEC Deviations, IDMT Dataset with 002.wav

[27] https://github.com/cvf-bcn-gituser/bassunet

SOP Measurements:

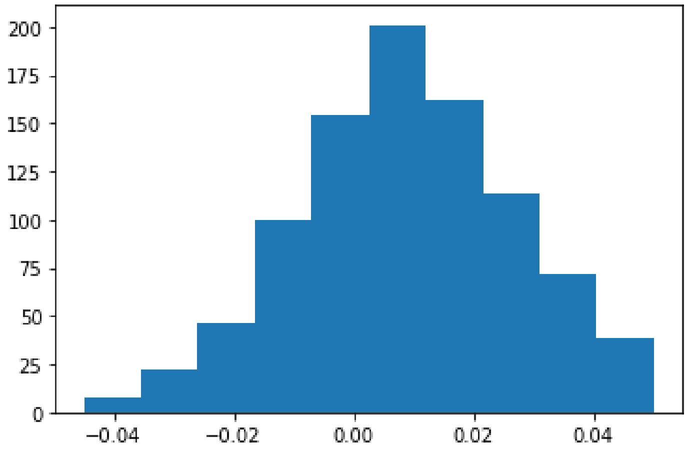


Figure 22: IDMT Dataset Onset deviations (SOP) algorithm with 002.wav

Doing a deviation measurement for Onsets using SOP for “bjean” with the MIR eval window set to 20ms yielded the following:

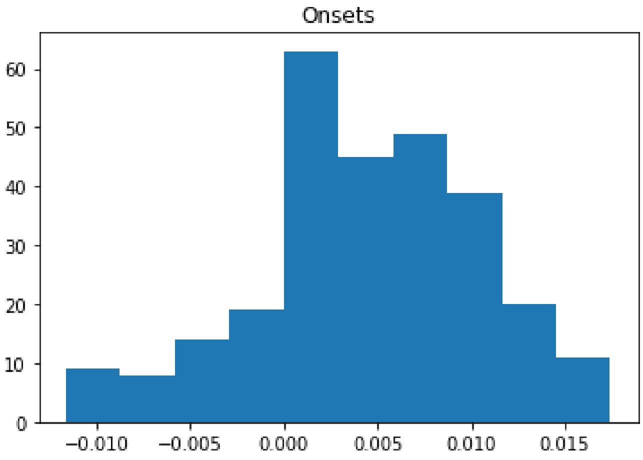


Figure 23: Billie Jean Reference onsets deviation

SOP Onset Deviations:

- ABS Mean: 0.005915
- Mean: 0.004029
- Dev. from 0: 0.007258

The deviation statistics are used to measure how far the student is from the reference performance in the next chapter.

5.8 Summary

Combining IEC and SOP for songs gave the optimal solution. MELODIA [12] was useful for benchmarking particular songs with muted and legato sections, but it is computationally very complex and not suitable for implementation. The PRF results are the important drive for choosing the algorithm. The deviations are explored in more detail for analysing student performances.

Although some tests on the students were made with Neural Network strategies, there was not enough scope in the project to choose both ASP and Data Driven methods. The ASP path was chosen since there was no easy way to prepare H5 models using the R&P Dataset. 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.

6. Prediction of Student Performance

The goal of predicting music grades centers on how far the students' performance is from the reference. This chapter describes the collecting and analysis of data: onset/offset annotations, the student stem recordings, the mixes, the grades and using data this to predict Teacher Grades. It discusses the following:

- Student Portal and Teacher Grading
- The testing algorithms to measure student performances
- Using a Multi-variable Linear Regression to train new a model with teacher graded student performances

The Machine Learning (ML) 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 (outputs) are the grades the teacher assigns for metronome accuracy, note length and the specific demands that each song requires for Technical Control. There has been a slightly higher deviation noted in the offsets as expected since there is no clear end point for long sustained notes.

6.1. Student Recording Portal

Each song has a Google Form containing instructions and links to a song portal (schema of portal shown below) to perform the live recording with a backing track.

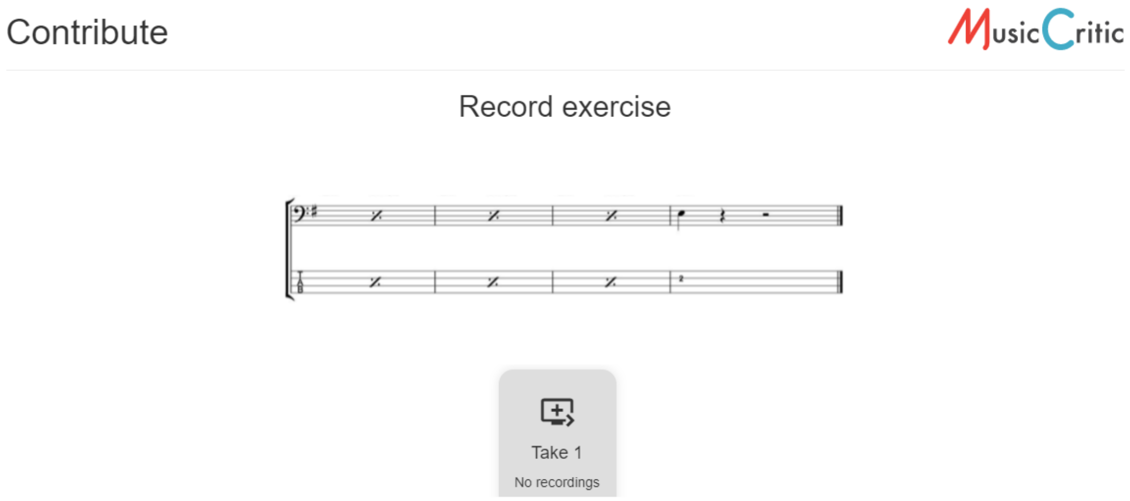
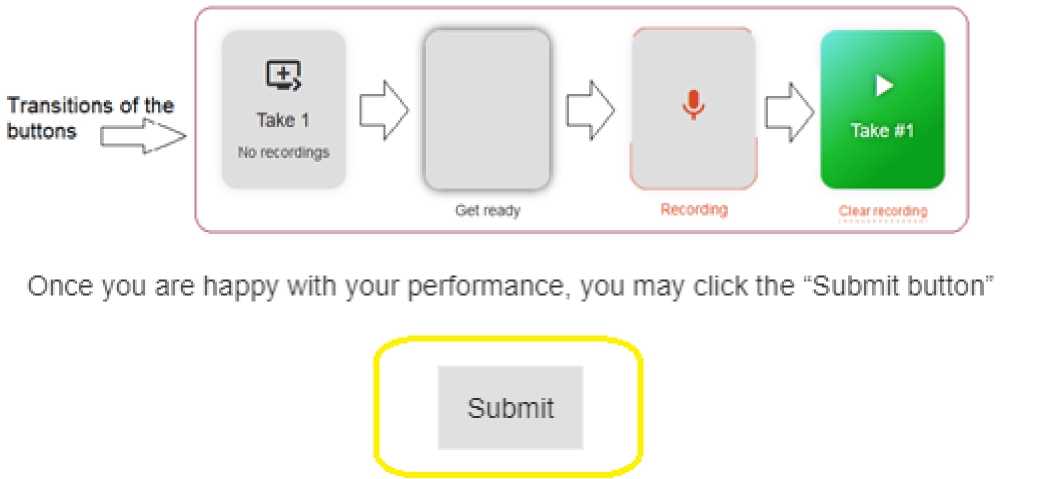


Figure 24: Student Portal (copyrighted score removed)



Once you are happy with your performance, you may click the “Submit button”

Figure 25: Transition states on student portal

The student presses the play button, listens, and plays along with the option to listen back on the stem before pressing “Submit”.More Details of this Portal can be found in the Appendix 12.4.

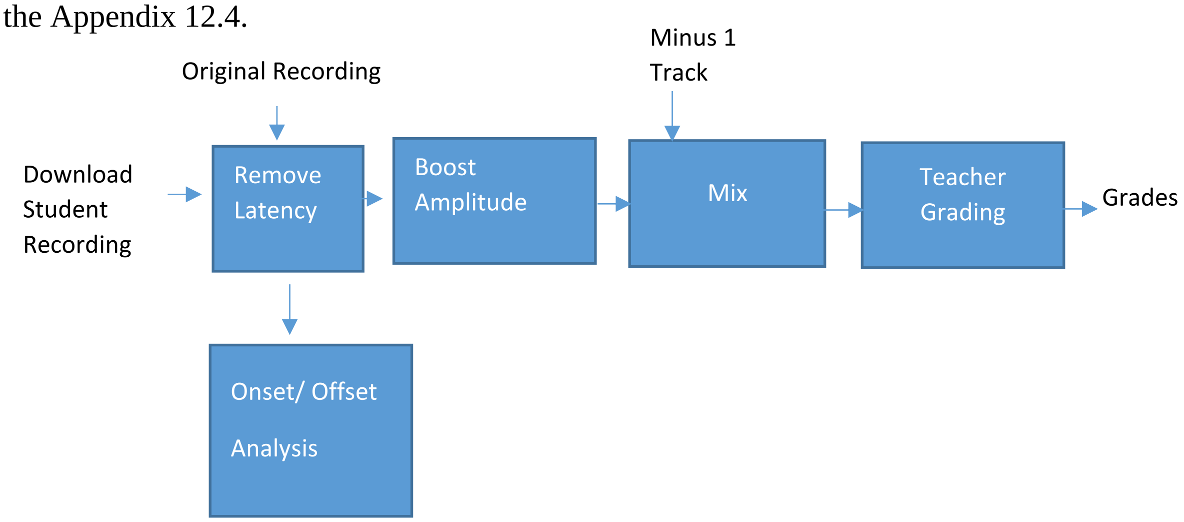


Figure 26: Recording Processing Stages

6.2. Processing of student recordings

After removing the latency from the recordings, it was noticed that the JSON calculated value provided by Music Critic did not align with the first onset, so the student recording stems were aligned manually with the first onset of the bass of the original 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. A copy mixed with the minus 1 track was provided to the teacher for grading. In the initial recordings, the microphone on the headset was used, there was some noticeable background noise which affected the teachers sound quality grading. However, the algorithms were still capable of onset detection for noisy stems. Full details of all 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 can be found in Appendix 12.5.

6.3. Teacher Grading Policy

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

Technical Control (TC) is song dependent, e.g. for "bjean" the main TF is "Articulation and Coordination". Exam Grades were chosen to fit within the typical ranges of a music academy system:

- Excellent (100%
- Very good (88%),
- Good with occasional lapses.(80%),
- Generally reliable.(63%)
- Unreliable (37%).

The Onset and Offset Grade sections are common to all songs, with slight variations on how the questions was phrased. On the Technical Control side, the table below summarizes the topics covered by each question.

Table 11: Technical Control Question Topics

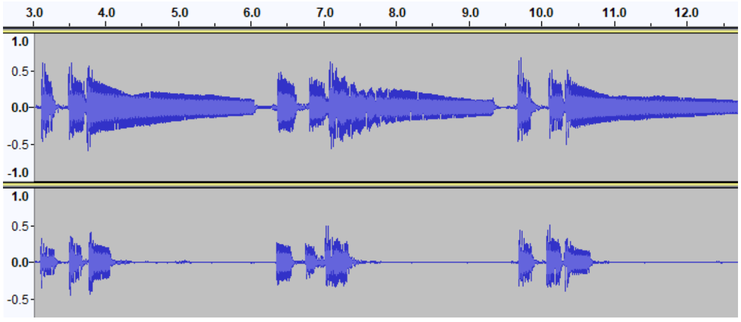
	Question 1	Question 2	Question 3
Yellow	Repeated Notes	Syncopation	Sound Quality
Bjean	Articulation	Dynamics	Sound Quality
Just	Accented syncopation	Dynamics	Sound Quality
Brown	Groove	Sound Quality	-----
Road	Syncopation	Articulation	Sound Quality
wotm	Syncopation	Sound Quality	-----

Comment sections on each question were added to reflect how Music Academies correct exams.

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 27: Contrasting comments of "wotm" performances

Comments also proved useful to understand the rationale behind why a particular recording was graded a certain way. 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. Each song has customized assessment policy that depend on the Technical Control Questions in Table 11. These questions are derived from the unique Song Description that is provided for each song in the Grade Books. I will take two songs as examples and will go into detail in the Song Descriptions and will explain the nuances of that the technical control criteria is that the Teacher is listening for.

Just Looking (just) examines “Accented Syncopation”. 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.

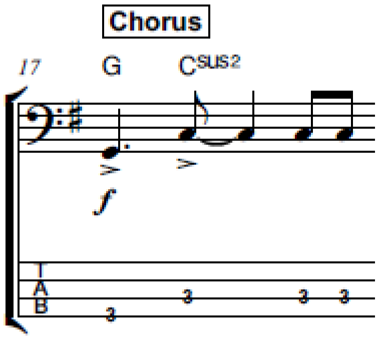


Figure 28: Syncopation example “just”

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 column in 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”.

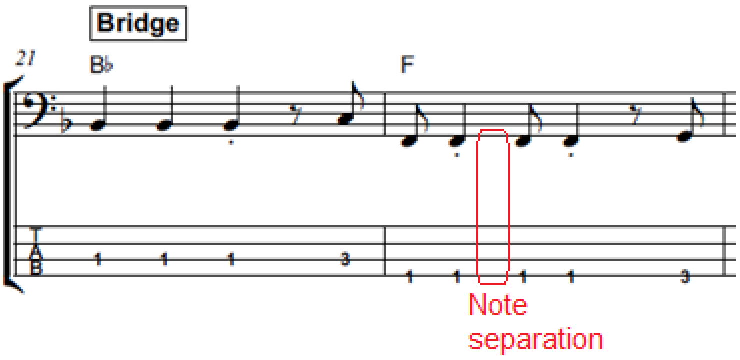


Figure 29: Syncopation example “wotm”

In “wotm” you can see the syncopated notes in the second bar above. The first crotchet in bar 22 sits on a half-beat slot. The effectiveness of syncopation is how precise this note sits in the right location and with the right separation with the subsequent quaver. The separation is slightly larger because the crotchet is dotted.

Because of these varying technical demands, the questions that have been designed for the teacher are phrased differently for the Technical Control questions.

6.4. Teacher Grading Results Analysis

After grading was complete on the Teacher Google Forms the CSV files were downloaded. Details in the Appendix 12.6 are provided on the Google Form Histograms, the python scripts developed to parse the text output of the CSV files to map Radio Box Choices to the numeric values (100%, 88%, 80%, 63 %, 37%) that can be loaded into the Data Tables to be used for training the Model.

The student recordings were analyzed for their onsets and offset measurements and compared against the reference stems. The Appendix section 12.1 gives details on how to reproduce the data with python scripts that is used for generating the accuracy results contained in the “StudentStatistics_<song>.csv file (e.g. StudentStatistics_yellow. Each student performance has an associated Deviations file for Onsets and Offsets, (yellow_devs_student1.csv… yellow_devs_student8) and raw data on the detected onsets and offsets (yellow1_onsets.csv… yellow2_onsets.csv)

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 core feature selection was the PRF accuracy measures. In addition, there are three Onset/Offset Deviations Statistics values available as input features:

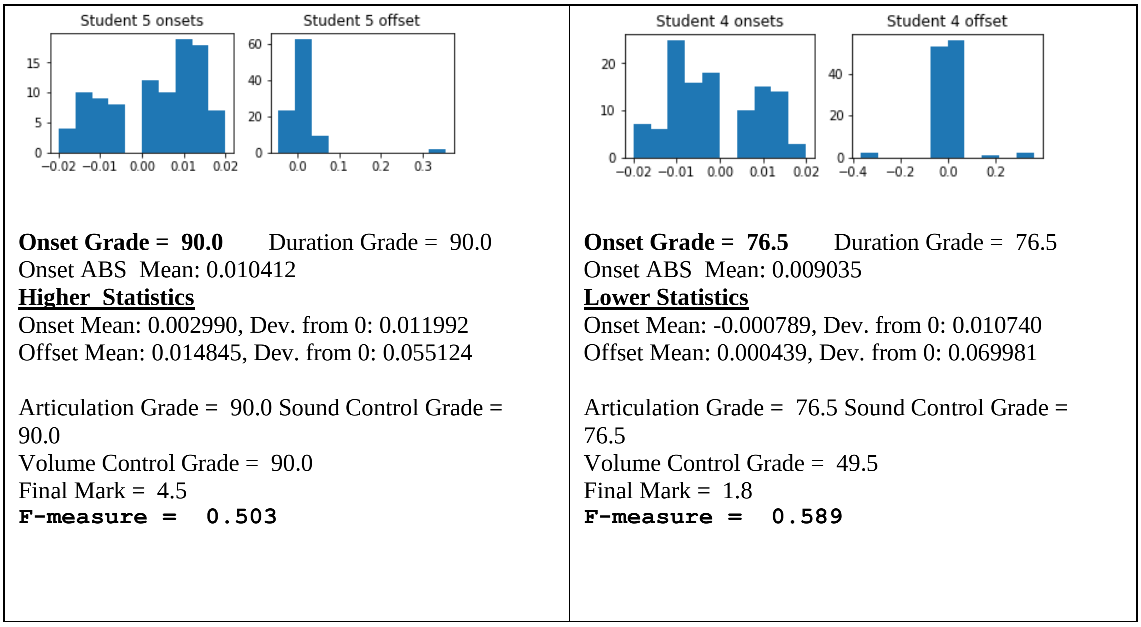
1. Mean Deviations
2. Absolute Mean Deviations
3. Standard Deviation

Ideally, after collecting all the deviation data from the ground truth, all the generated statistics (Mean, Standard Deviation) could be used as Feature Inputs to obtain the best Prediction possible.

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.

6.5. Yellow

While the F measure of the reference stem was close to 100% the highest graded Student 5 only scored a F-measure of 0.503 while 2nd highest graded Student 4 scored and F-measure of 0.589. Student 4 produced lower deviation statistics than Student 5 but this was not reflected in the grades allocated by the Teacher.



```
Onset Grade =  90.0      Duration Grade =  90.0
Onset ABS  Mean: 0.010412
Higher  Statistics
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
F-measure =   0.503
```

```
Onset Grade =  76.5      Duration Grade =  76.5
Onset ABS  Mean: 0.009035
Lower Statistics
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
F-measure =   0.589
```

Figure 30: Best two “yellow” students

Sound Quality

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.

Student 1[28] 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.

Full details of the plots for Predicted Grades vs Actual Grades for Onsets and Duration are in the Appendix 12.3

6.6. Billie Jean

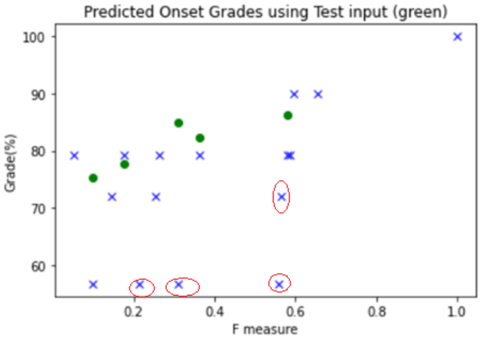
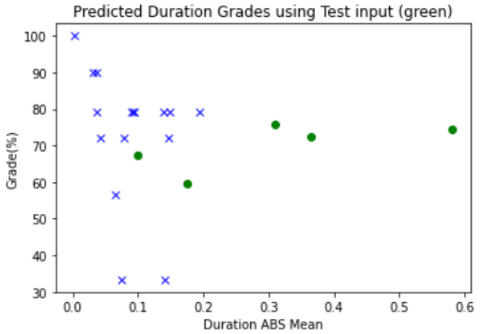
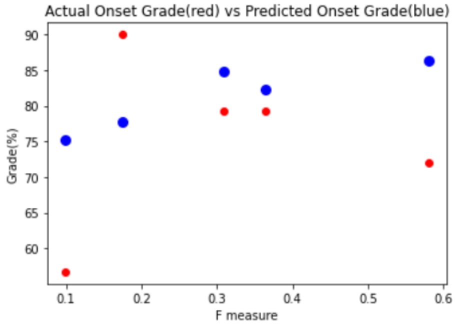
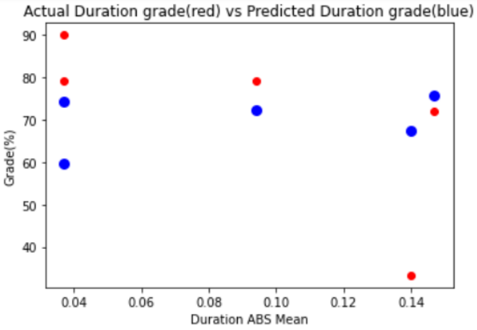
The song was divided into three sections: 1st verse (Muted), Bridge (Non-Muted), Chorus (Muted). “Bjean” having the most recordings, is discussed here with detailed plots on Prediction, while Grade 0, Grade 2 and Grade 3 songs are referred to the Appendix 12.3. The first column in table 12 below contain plots for the Onset grade predictions and the second column shows the plots for the Duration grade predictions. The Test Size is 30% of the total set. The linear regression formula is as follows:

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3)
regressor = LinearRegression()
regressor.fit(X_train,y_train)
y_pred=regressor.predict(X_test)
```

The blue “x” marks are plots of F-measures against actual grade:(y_train+y_test). The green dots are plots of the x_test subset of F-measure against a predicted grade or y_pred.

[28] https://github.com/cvf-bcn-gituser/bass-critic/blob/main/tclEvaluation/audio/yellow/yellow1.wav
Find other recordings yellow4 and yellow5 in the same folder.

Table 12:Billie Jean Predicted Grades

Onset Grade	Duration Grade
	
	

`   Actual Grade  Predicted Grade`
`0      79.199997           73.288`
`1      79.199997           70.397`
`2      56.700001           72.640`
`3      56.700001           64.343`
`4      79.199997           81.822`
`Mean Absolute Error: 8.184 %`
`Root Mean Squared Error: 9.293 %`
	Actual Grade Predicted Grade
	`0      79.199997           72.346`
	`1      90.000000           59.667`
	`2      72.000000           75.751`
	`3      33.299999           67.407`
	`4      79.199997           74.384`
	`Mean Absolute Error: 15.972 %`
	`Root Mean Squared Error: 20.82%`

In the first row, the Onset predictions using the test set in green show a reasonable linear interpolation, if you consider blue ‘x” points with red circles to be outliers. The Duration predictions from the Test Input are way off. There is a green point with Absolute Mean Deviation of 0.6 with a grade above 70%. An explanation might be that the Offset Deviation Histogram of the reference is not as low in its deviation statistics.

The second row of plots (Actual Grade vs Predicted Grade) and summary tables show more errors in predicting duration grades. The onset grade plots in Column 1 should be showing a linear proportional relationship between the F-measure and grade. The Duration curve plots in Column 2 should show an inverse relationship between Absolute Mean Duration and grade. One observation might be that chosen granularity of the actual grades into five numeric values is too wide.

Outliers due to Bad Mixing

Let us return to the four outliers marked in red circles. They are 50% grades with F measure values 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 teacher's 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.

Therefore, this outlier could be explained by the fact that the step was incorrectly synchronized with the Minus 1 track.

Outliers due to Imposter Notes

Student 6 (see Student Statistics section in the Appendix 12.8) 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) in figure 37 student performance show similar statistics, partly because of the "filter effect" [29] of the minimum window size.

[29] Student performances with low accuracies will mean that less onset and offset measurements get inside the 20ms window for checking, so less deviation calculations are made.

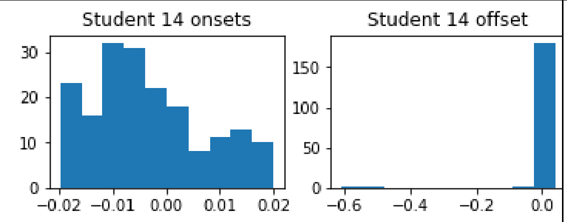
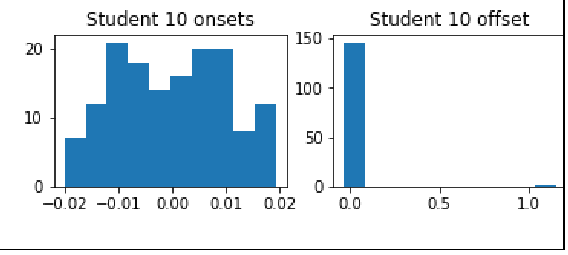
High Grade: Student 14 157 deviations	Worst Grade: Student 10 62 deviations
	

Figure 31:Best vs Worst bjean students

6.7. Just Looking

Poor Duration results

Even though there are muted notes in the chorus, the SOP algorithm is used for the entire song, given its superior precision accuracy on the reference stem. This grade 1 song has wider outliers than “yellow” (grade 0).

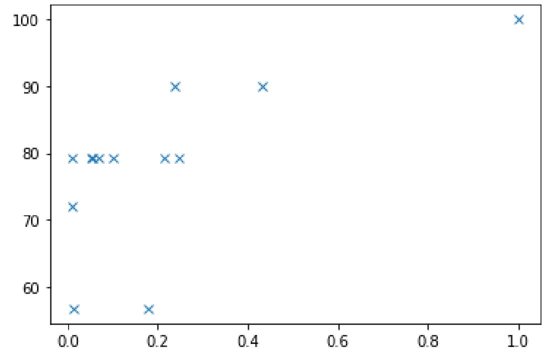
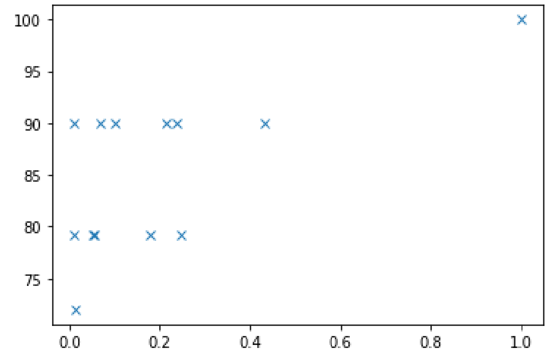
Onsets	Duration
	

Figure 32: Just Looking: Grades % (Y axis) Precision (X-axis)

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

6.8. Brown Eyed Girl

For “brown”, there was no way to improve accuracy by blending the algorithms. 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. These reference figures make Student Evaluations very unreliable.

Short Difference between Notes

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

Table 13:First 8 onsets of “brown”

Onset Mark	Difference with previous onset
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

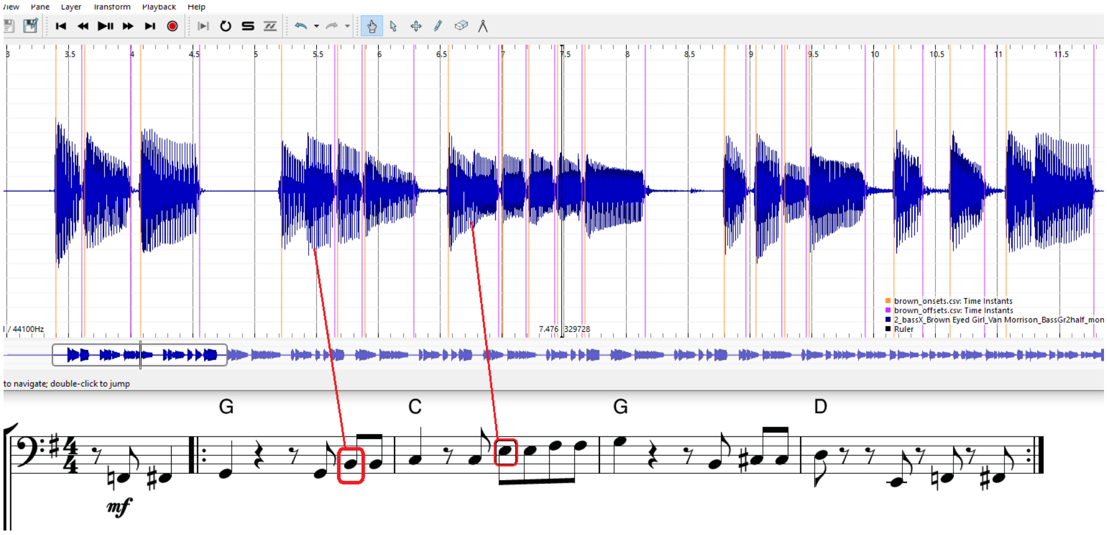

View
Pane
Layer
Transform
Playback
Help
3
3.5
4
4.5
5
5.5
6
6.5
7
7.5
8
8.5
9
9.5
10
10.5
11
11.5
/ 44100Hz
7.476 329728
brown_onsets.csv: Time Instants
brown_offsets.csv: Time Instants
2_bassX_Brown Eyed Girl_Van Morrison_BassGr2half_mon
Ruler
to navigate; double-click to jump
G
C
G
D
mf

Figure 33: Brown: Missed Onset pattern

Sound Archipelagos

In bar 2 (the first G chord bar) the four notes B, B, C# and D are all detected as 4 sound islands. In another occurrence of the exact same note sequence audio blow, not only is there a missing onset of the B note but there are two false alarm sound islands (13.4 seconds). This note-pattern tends to throw false alarms, thus reducing precision below an acceptable amount.

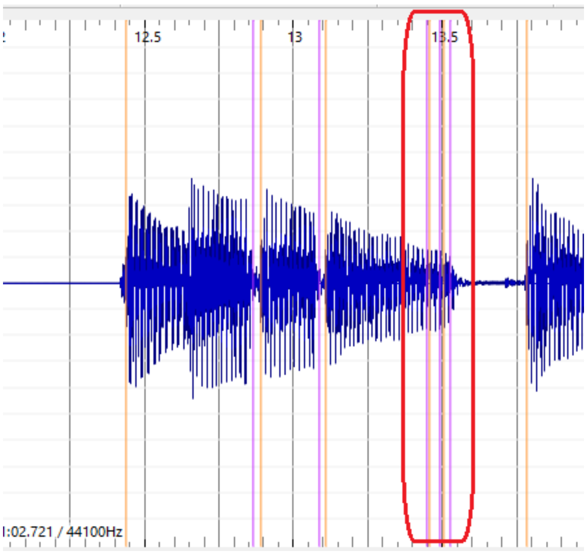


Figure 34: Brown: False Onsets

Mismatch on Grade Prediction

It is not surprising therefore that big outliers can happen : Student 4 who scored 90% grade for onsets, but a precision around 10%.

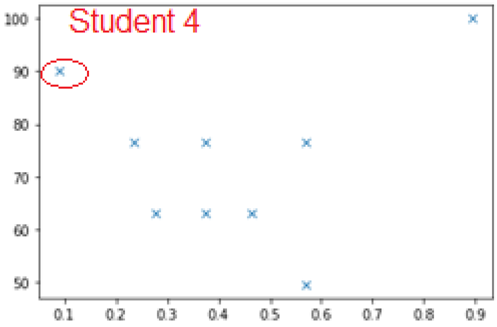


Figure 35: Brown: Onset Grades vs Precision

To try to understand this we look at the waveforms. At 36.5 secs Student 4 correctly hits 6 offsets but the GT only hits 5 onsets. There is almost a 10ms drift at 39.5 second timestamp.

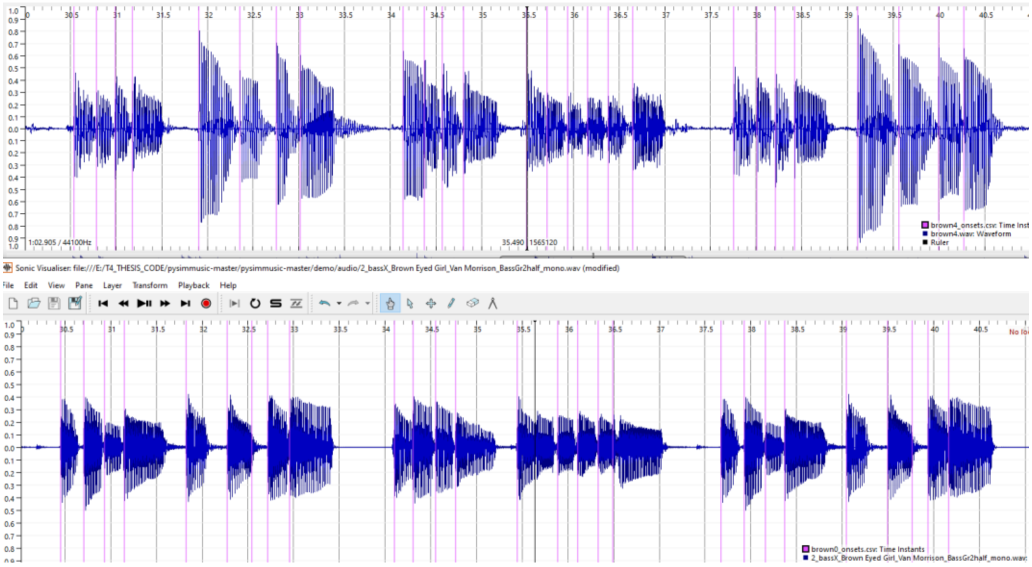

1:02.905 / 44100Hz
35.490 1565120
brown4_onsets.csv: Time Inst
brown4.wav: Waveform
Ruler
Sonic Visualiser: file:///E:/T4_THESIS_CODE/pysimmusic-master/pysimmusic-master/demo/audio/2_bassX_Brown Eyed Girl_Van Morrison_BassGr2half_mono.wav (modified)
File
Edit
View
Pane
Layer
Transform
Playback
Help
No lo
brown0_onsets.csv: Time Instants
2_bassX_Brown Eyed Girl_Van Morrison_BassGr2half_mono.wav:

Figure 36: Brown: Reference and Student 4 stems with onsets detected

As a result of having a very low PRF, the IEC algorithm returned zero offset deviations for Student 4) because the evaluation window was too narrow to measure the difference. To conclude, the given the very short note gaps, this song is the most difficult to predict Duration Grades (see appendix 12.3 for fill statistics)

6.9. Roadrunner

Improved after “Tenuto” section removed.

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 as feature inputs. As expected for a grade 3 song, the number of outliers was higher compared to lower grades. See Appendix 12.3 for full statistics on grades.

6.10. Walking on the Moon

The PRF using the IEC algorithm alone for WOTM was 0.701, 0.932, 0.8. After segmenting the verse into the SOP algorithm and the bridge into the IEC algorithm, the new PRF was 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?

Effective Duration

The Essentia library function “Effective Duration” considers perceptual factors [29] 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

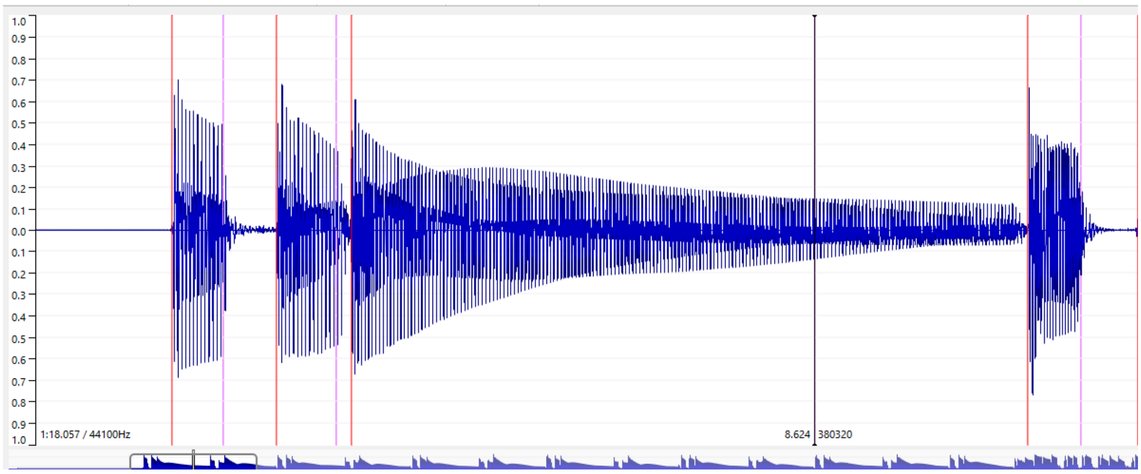


Figure 37: WOTM: Offsets using Effective Duration (T= 0.05)

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. See Appendix 12.3 for full statistics on grades.

6.11. Technical Control Grades

The Actual vs Predicted Grade for Technical Control Element 1 for “bjean”, “brown” and “wotm” are discussed briefly here.

Table 14: Technical Control Grade Prediction Errors

Song.	N	Algorithm	Actual Grades	Precicted Grades	MAE Onset TF1	RMS Error Onset TF1
bjean	15	IEC/SOP	72.0 90.0 79.2 33.3	75.6 76.3 70.3 67.7	13.56	18.15
brown	8	IEC	79.2 79.2 72.0	76.8 77.2 76.9	3	3.3
wotm	11	SOP/IEC	72.0 72.0 79.2 56.7	80.0 82.6 74.9 78.8	11.26	13.07

A Mean Absolute Error of 13.56 % in predicting the Articulation or TF (Technical focus 1) Grades using the PRF and Mean Duration Deviation as inputs. “Brown” yielded a low error prediction, but caution must be exercised in drawing a conclusion, considering the onset and duration grade error were over 10%. However, it does highlight the case that maybe the teacher should combine the Technical Focus Areas 1 and 2 into one joint grade.

7. Conclusion

7.1. Student Recordings and Teacher Gradings

Five submissions were collected by researchers at the MTG, and I submitted the remaining 48 recordings.

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, especially for the songs like "Roadrunner" ("road") that were difficult to execute. 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" began, 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 professional reference 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.

7.2. Musical and Technical issues with Recording Experiment

Some of the bassists who declined 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.

Another restriction was that Music Critic is not supported on Tablets and Smart Phones, and one professional bassist reported that he had no 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, there 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 marked between 1-8 and the Technical Control grades should be discarded, until such time that significant progress is made on accurately measuring duration.

7.3. Improving analysis and prediction strategies

The semi-automatic strategy of customizing the algorithms according to the songs was a compromise and not the original intention. It was proposed to use as input a Three-Column “rhythm” csv-format annotation files. The threes columns are: Onset Time, Muted, Offset Time”. The Onset and Offset time represent the annotations of the reference stem recording while the “Muted” property is set to Y to denoted “muted” or “N” to denote “not muted”. The idea was that the IEC algorithm would choose the “next onset” as the offset, but it did not yield significant improvements. An improvement could be made to switch between the calling of one algorithm or another depending on the CSV column contents.

Although I made additional submitted recordings and gradings, the overall quantity of data meant a limited choice of Machine Learning algorithms. The feature extraction experiments by Abesser used large single note data sets and so support vector machines were used for prediction. I found that in some cases, adding 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. Adding the Standard Deviation statistic as an input did not help in most cases.

No further research was added to Abesser’s classification of plucking styles [17], but the three column “rhythm” csv file, would be a good continuation point for adding more functionality and parameters to the algorithms. For example, the ground truth input files would 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 example of how to extend the existing methodology to customize algorithms for certain technical focus elements.

For the sound quality grade, no grade prediction was attempted as the audio features that the algorithms extracted doesn’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 is useful in trying to understand how robust the onset detection is in the presence of noise.

7.4. Improving Duration measurements

The main challenge was the offset on long notes, typically composed of tied quarter, eight notes. If a particular song yielded better results using SOP, then this meant that there was no validity check on whether the 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 more than double that of Onset detection. A more rigorous measurement approach would incorporate the Effective Duration into the offset algorithms as discussed in results section for “wotm”. The matching window size should also be made narrower.

The repeated quavers in the verse for “yellow” leaves no gaps for duration so it is meaningless here. This applies to many of the TCL songs, apart from the ones mentioned here (e.g., the AC/DC song the Grade 1 book [26]). Theoretically it would be possible to play shorter notes than quavers, but that is 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: when all notes are equal or when the third note is absent. Between these two extremes: an ideal energy statistic could be chosen for tenuto, considering the limit of human perception.

7.5. 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 and partitioning into different “tendencies”: late/early onsets and short/long duration playing styles.
- Continually annotating more reference songs.
- Source Separation to obtain more annotated data.

- Consideration of Rockschool Music Syllabus[30]

Finally, the results have shown us that the precision, recall and f-measure metrics are good indicators of good performance, but they need more robustness to close the gap between the Student Performance and the reference. It is only possible to compare the student performance deviation statistic data with another student performance of similar PRF measure. If the overall mean value of the onset deviation is negative that indicates early tendency and likewise if the mean of the duration deviation is negative this indicates a tendency to play shorter notes. Earliness and lateness (and duration deviation) must be looked at in the context of the musical properties of a song.

Data-driven alternatives to ASP methods were not tested in depth in this thesis, but some late tests on the Billie Jean student data in the Appendix 12.9 show that the Recall Rate competes well with Signal Processing techniques. The advantage of non-real-time performance assessments is that they are not constrained by the computational cost requirements that training sets and statistical methods demand. The evaluation results given by the pre-trained models in the U-Net architecture were low [19] but perhaps new models based on the training the TCL dataset would yield better results.

To revisit the goals of the State of the Art and the title of the thesis “automatic assessment of timing and rhythm”, the algorithms presented in this thesis can help improve how assessment is performed. Since the algorithms are song customised, they are better described as semi-automatic assessment of onset and duration. It introduced and tested some ideas that consider technical control assessment, but there is still much work to do build on the research of the expressive and plucking styles of playing bass.

[30] "RockSchool," [Online]. Available: https://www.rslawards.com/country/espana/.

8. List of figures

Figure 1: 4 crotchets, 4 quavers and 4 semi-quavers. 3
Figure 2: Yousician example with Vocals 3
Figure 3: Onset Detection Stages 5
Figure 4: Peak Picking implementation 8
Figure 5: Processing Flow chart of the instrument level extraction process 13
Figure 6: Yousician Detecting slides, hammer-ons, pull-offs on 16th notes. 14
Figure 7: Extract from IDMT Dataset. File 002.xml 18
Figure 8: Note duration of Staccato notes 19
Figure 9: Audio 002.wav (IDMT DATASET) with onsets using HFC method 20
Figure 10: Billie Jean: Midi Rendering vs Human recording (shorter notes) 21
Figure 11: Strategy to evaluate, test and find the best accuracy 25
Figure 12: End to End test of Music Critic with Billie Jean GT 27
Figure 13: Normalized Energy of Audio 002.wav sample (IDMT DATASET) 29
Figure 14: Sound Islands of Audio 002.wav (IDMT). FS = 1024, HS = 512. 29
Figure 15: IEC onsets for “wotm” verse 32
Figure 16: Section of “wotm”verse showing missed IEC onsets. 32
Figure 17: Calculated IEC Sound Islands for WOTM bridge, P = 0.937 33
Figure 18: PitchMelodia derived Onsets for WOTM verse , P = 0.97 33
Figure 19: PitchMelodia derived Onsets for WOTM bridge 34
Figure 20: Sample of bjean_rhythm.csv verse (left) and bridge(right) 34
Figure 21: IEC Deviations, IDMT Dataset with 002.wav 36
Figure 22: IDMT Dataset Onset deviations (SOP) algorithm with 002.wav 37
Figure 23: Billie Jean Reference onsets deviation 37
Figure 24: Student Portal (copyrighted score removed) 39
Figure 25: Transition states on student portal 40
Figure 26: Recording Processing Stages 40
Figure 27: Contrasting comments of “wotm” performances 42
Figure 28: Syncopation example “just” 43
Figure 29: Syncopation example “wotm” 43
Figure 30: Best two “yellow” students 45
Figure 31:Best vs Worst bjean students 49
Figure 32: Just Looking: Grades % (Y axis) Precision (X-axis) 49
Figure 33: Brown: Missed Onset pattern 51
Figure 34: Brown: False Onsets 51
Figure 35: Brown: Onset Grades vs Precision 52
Figure 36: Brown: Reference and Student 4 stems with onsets detected 52
Figure 37: WOTM: Offsets using Effective Duration (T= 0.05) 54
Figure 38: WOTM: Offsets using Effective Duration (T= 0.05) 70
Figure 39: Student Portal Section 2 77
Figure 40: Grading Histogram Onsets: Billie Jean 79
Figure 41: Grading Histogram Offsets: Billie Jean 80

9. List of tables

Table 1: Some ODF methods from Essentia and Madmom 6
Table 2: Abesser Signal Processing vs MELODIA 10
Table 3: MELODA Algorithm steps 11
Table 4: Abessers-Müller Data Driven algorithms vs Salamon/Gomez 12
Table 5: Algorithms for Onset/Offset detection 15
Table 6: TCL dataset 20
Table 7: Tech. Control Parameters for 6 TCL songs 22
Table 8: Benchmarking Onset detection algorithms using IDMT Dataset 30
Table 9: Selected comparison: IEC vs MELODIA 31
Table 10: Accuracy Metrics for 6 TCL references using IEC-SOP 35
Table 11: Technical Control Question Topics 42
Table 12:Billie Jean Predicted Grades 47
Table 13:First 8 onsets of "brown" 50
Table 14: Technical Control Grade Prediction Errors 55
Table 15:Yellow Predicted Grades 71
Table 16:Brown Predicted Grades 72
Table 17:Brown Reference Histograms 72
Table 18:Road Predicted Grades 74
Table 19:Road Reference Histograms 75
Table 20:Wotm Predicted Grades 76
Table 21: R&P Grades Actual vs Predicted 81
Table 22: Student Statistic extract "bjean": ONSET focus 82
Table 23: Student Statistic extract "bjean": OFFSET focus 83
Table 24: Abesser Bass U-Net: PRF_measure_IDMT_sample.csv 84
Table 25: Abesser Bass U-Net on Billie Jean Student Recordings 84
Table 26: Abesser Bass U-Net on Billie Jean Student Recordings 85
Table 27: Bock Online Onset Detector 85

10. List of symbols

Equation 1 Local Energy Equation 4
Equation 2 Energy in a Frequency Band 5
Equation 3 Onset Detection Function Equation 5
Equation 4 Peak Picking Formula 7
Equation 5 Peak Picking Formula (adaptive) 7
Equation 6 Normalized Energy Function 29

11. Bibliography

[1] N. McCormick, U2 by U2: Bono, The Edge, Adam Clayton, Larry Mullen Jr, New York: HarperCollins, (2009).

[2] C. Dittmar, E. Cano, J. Abeßer and S. Grollmisch, "Music Information Retrieval Meets Music Education," in *Multimodal Music Processing*, Dagstuhl, Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik, p95-120, (2012)..

[3] P. Pfeiffer, Bass Guitar For Dummies, 3rd edition, Hoboken, NJ.: John Wiley and Sons, (2014).

[4] B. H., S. Giraldo and R. Ramirez, "Jazz ensemble expressive performance," in *17th International Society for Music Information*, (2016).

[5] TCL, Trinity R&P Bass Syllabus from 2018, London: Trinity College London, (2017).

[6] F. E. MUSIEK and G. D. CHERMAK, "Psychophysical and behavioral peripheral and central auditory tests," in *Handbook of Clinical Neurology, Chapter 18*, vol. 129 (3rd series), Elsevier, (2015).

[7] J. Bello, C. Duxbury, M. Davies and M. S. M. I. Sandler, "On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain," in *IEEE Signal Processing Letters*, (2004).

[8] G. C. P. Tzanetakis, "Multifeature Audio Segmentation for Browsing and Annotation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, W99 1-4, (1999).

[9] S. Dixon, "Onset detection revisited," in *International Conference on Digital Audio Effects (DAFx'06)*, pp. 133-137, (2006).

[10] 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).

[11] S. Böck and G. Widmer, "Maximum Filter Vibrato Suppression for Onset Detection," in *Proceedings of the 16th International Conference on Digital Audio Effects*, (2013).

[12] E. Gómez and J. Salamon, "Melody extraction from polyphonic music signals using pitch contour characteristics," in *IEEE Transactions on Audio, Speech, and Language Processing,*, (2012).

[13] J. P. Bello and M. B. Sandler, "Phase-based note onset detection for music signals," in *ICASSP-88*, (2003)

[14] E. Terhardt, "Psychoacoustic evaluation of musical sounds," *Perception and Psychoacoustics,* vol. 23 (6), 483-492, (1978).

[15] 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).

[16] Mirex, "music-ir.org," MIREX, [Online]. Available: https://www.music-ir.org/mirex/wiki/2020:Singing_Transcription_from_Polyphonic_Music.

[17] 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).

[18] Fraunhofer. [Online]. Available: https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/bass.html.

[19] J. Abeßer and M. Müller, "Jazz Bass Transcription Using a U-Net Architecture," *Electronics,* vol. 10, 12 March (2021).

[20] M. Goto, "Predominant-F0 estimation for detecting melody and bass lines in," in *Speech Communication*, (2004).

[21] 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).

[22] J. Abeßer, K. Frieler, M. Pfleiderer 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).

[23] J. Abeßer, H. Lukashevich and G. Schuller, "Feature-based extraction of plucking and expression styles of the electric bass guitar.," in *IEEE International Conference on Acoustics,Speech and Signal Processing*, (2010).

[24] 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).

[25] 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).

[26] Trinity College London, "R&P Bass Grade 1," in *R&P Bass Grade 1*, Londoin, (2017)

[27] C. Kehling, J. Abeßer, C. Dittmar and G. Schuller, "Automatic Tablature Transcription of Electric Guitar Recordings by estimation of score and instrument-related parameters," in *DAFx*, (2014).

[28] V. Eremenko, A. Morsi, J. Narang and X. Serra, "Performance assessment technologies for the support of musical instrument learning," in *12th International Conference on Computer Supported Education*, (2020).

[29] G. Peeters, ""A large set of audio features for sound description (similarity and classification) in the CUIDADO project," CUIDADO I.S.T., (2004).

12. Appendices

12.1. Code for generating data.

README

https://github.com/cvf-bcn-gituser/bass-critic/blob/main/README.md

Follow guidelines for in the requirements.txt

In addition, install the following

```
pip3 install ipywidgets
pip3 install lxml
pip3 install plotly
pip3 install ipython
```

Python Code

https://github.com/cvf-bcn-gituser/bass-critic/tree/main/tclEvaluation

The main program for producing all the Student Statistics is

osets_song.py

This file gives you a menu to choose one of the 6 Rock and Pop songs for generating onsets, offset, deviations for onsets and offsets, PRF statistics.
It parses the CSV text files with the Grades from the Bass teacher and combines the summary statistics along with the given Grades.

These generated "StudentStatisics" files and the generated deviation files can be uploaded to the google drive for generating the Histogram Plots and the Student Statistics.

Abesser_test.py
This file generates the PRF accuracies for the 4 chosen songs from the IDMT datatset

using both SOP and IEC algorithms.

The MIR evaluation window in the script can be set to desired values (e.g. 20ms)

12.2. Reproduction of Predicted Grades Plots

The python scripts produce data that are used for plots. To reproduce these plots, you need to upload the CSV files in the data folder below to your Google Drive.

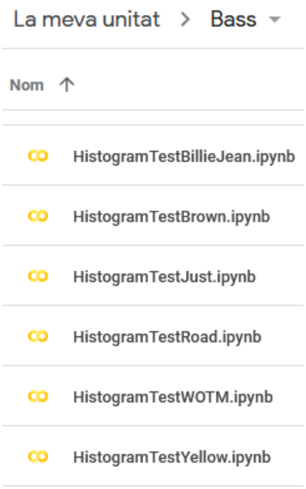
https://github.com/cvf-bcn-gituser/bass-critic/tree/main/tclEvaluation/data

STEPS:

1. Create a folder in your Google drive called "Bass" and subfolder "data"
   Checkout all the notebooks from the following link:

   https://github.com/cvf-bcn-gituser/bass-critic/tree/main/notebooks_for_plots

2. Upload those notebooks to the "Bass" folder in your Google Drive



La meva unitat > Bass

Nom

HistogramTestBillieJean.ipynb

HistogramTestBrown.ipynb

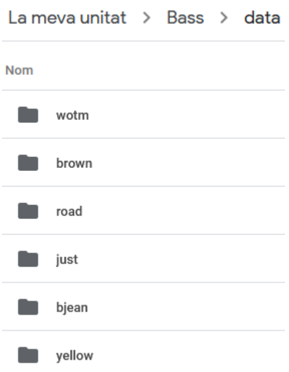
HistogramTestJust.ipynb

HistogramTestRoad.ipynb

HistogramTestWOTM.ipynb

HistogramTestYellow.ipynb

3. In this sub-folder create sub folder "data" and inside this the folders "wotm", "brown" etc. as shown:



La meva unitat > Bass > data

Nom

wotm

brown

road

just

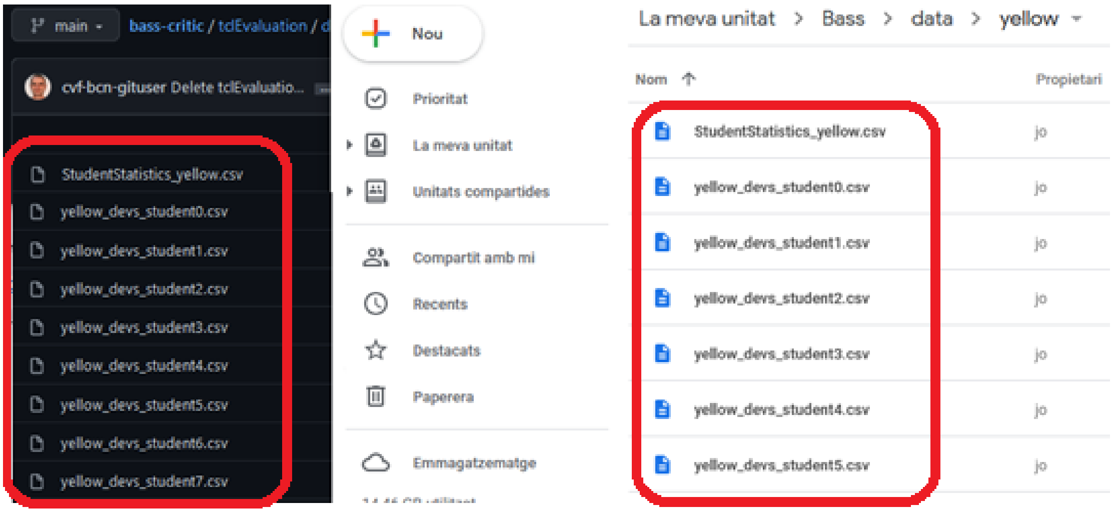
bjean

yellow

4. From git hub, go to the contents of each the data folder, starting with "yellow"
   https://github.com/cvf-bcn-gituser/bass-critic/tree/main/tclEvaluation/data/yellow

5. Download the “StudentStatisics” file and deviation files (yellow_devs_student0….8.csv) to your local machine, then upload them to the corresponding folder (“yellow”) in google drive as follows:

Figure 38: WOTM: Offsets using Effective Duration (T= 0.05)



StudentStatistics_yellow.csv
yellow_devs_student0.csv
yellow_devs_student1.csv
yellow_devs_student2.csv
yellow_devs_student3.csv
yellow_devs_student4.csv
yellow_devs_student5.csv
yellow_devs_student6.csv
yellow_devs_student7.csv

Nom	Propietari
StudentStatistics_yellow.csv	jo
yellow_devs_student0.csv	jo
yellow_devs_student1.csv	jo
yellow_devs_student2.csv	jo
yellow_devs_student3.csv	jo
yellow_devs_student4.csv	jo
yellow_devs_student5.csv	jo

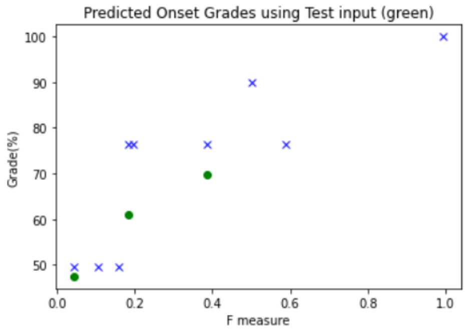
6. Repeat the process of copying data files from github links https://github.com/cvf-bcn-gituser/bass-critic/tree/main/tclEvaluation/data to google Drive for the remaining songs : “bjean,”just”,”brown”,road,”wotm”
7. Open the Notebook “HistogramTestYellow.ipynb” and execute.
8. Do this for all the songs to obtain the Deviation Histogram for each student and the Predicted Grades vs Actual Grades and the Mean Absolute Error and Root Mean Squared Error.

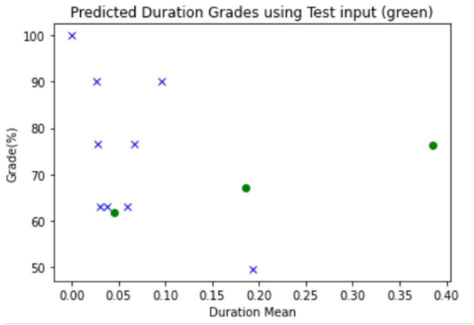
12.3. Extra Plots for Predicted Grades

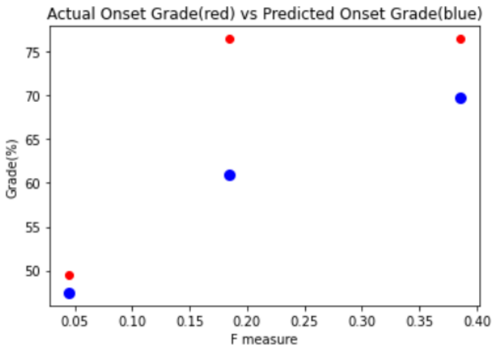
Billie Jean was already included in the results chapter. Here are the others.

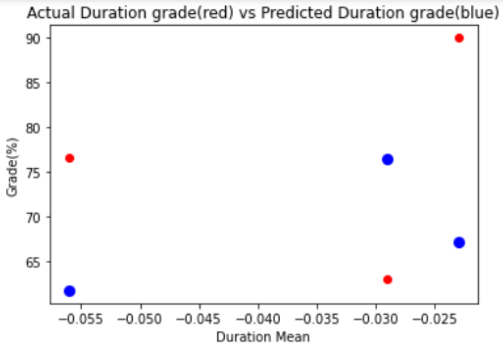
Yellow (Grade 0)

Table 15:Yellow Predicted Grades

<table>
<tr><td>Onset Grade</td><td>Duration Grade</td></tr>
<tr><td>


</td><td>


</td></tr>
<tr><td>


</td><td>


</td></tr>
<tr><td>
    Actual   Predicted
0   76.5   60.924
1   76.5   69.761
2   49.5   47.442
Mean Absolute Error: 8.124
Root Mean Squared Error: 9.87
</td><td>
Actual   Predicted
0   90.0   67.061
1   63.0   76.366
2   76.5   61.687
Mean Absolute Error: 17.039
Root Mean Squared Error: 17.552
</td></tr>
</table>

For the next two plot series for “brown” and “road” the predictions get more difficult for Onsets. The Histograms of the reference recordings for these two songs are shown.

Brown Eyed Girl (Grade 2)

Table 16:Brown Predicted Grades

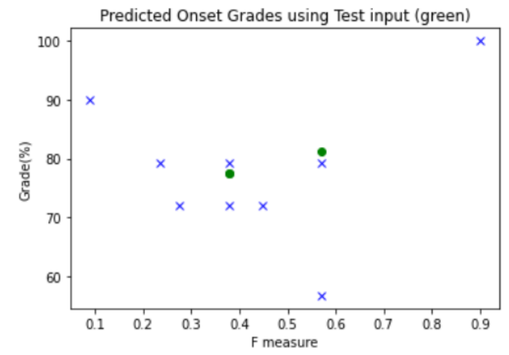
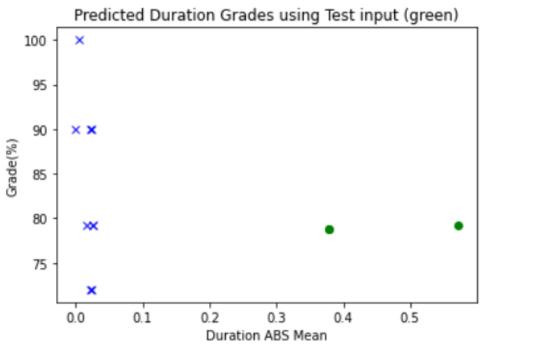
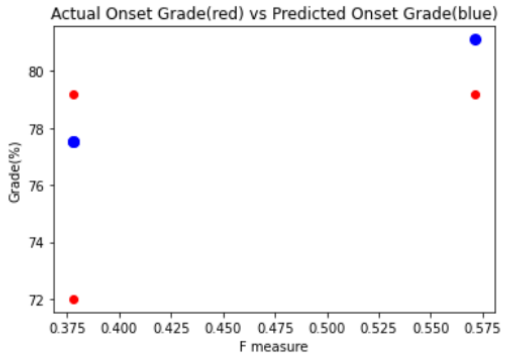
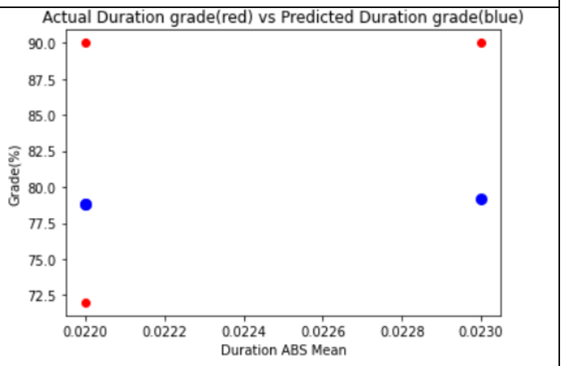
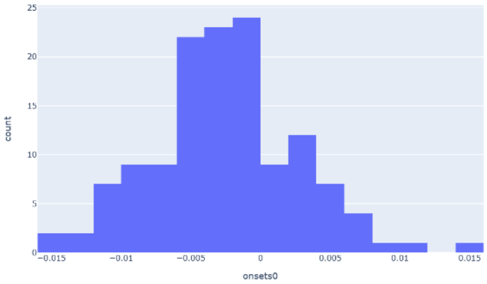
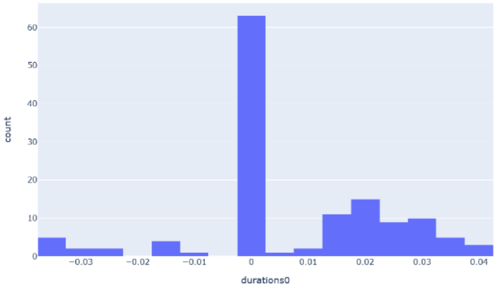
Onset Grade	Duration Grade
	
	
Actual Predicted 0 79.199997 77.9311 79.199997 79.709 2 72.000000 77.931 Mean Absolute Error: 2.57 Root Mean Squared Error: 3.514	Actual Predicted 0 72.0 78.778 1 90.0 79.201 2 90.0 78.778 Mean Absolute Error: 9.6 Root Mean Squared Error: 9.806

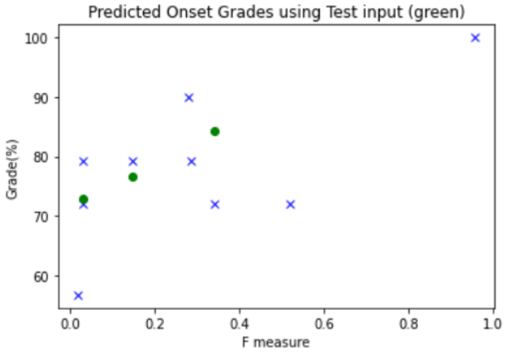
Table 17:Brown Reference Histograms

<table>
<tr><th>Onset Deviations</th><th>Duration Deviations</th></tr>
<tr><td> </td><td> </td></tr>
</table>

Road Runner (Grade 3)

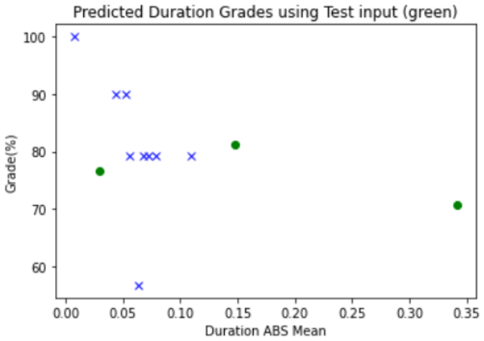
Table 18:Road Predicted Grades

<table>
<tr><th>Onset Grade</th><th>Duration Grade</th></tr>
<tr><td>



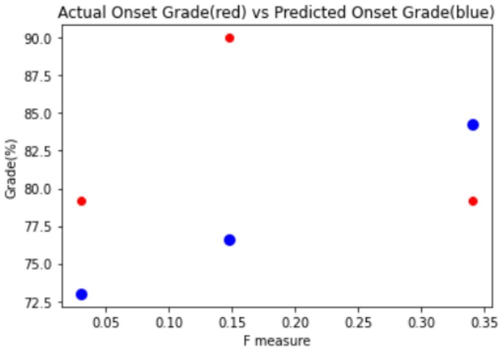
Predicted Onset Grades using Test input (green)

</td><td>



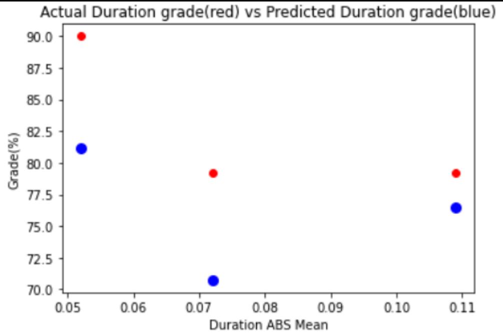
Predicted Duration Grades using Test input (green)

</td></tr>
<tr><td>



Actual Onset Grade(red) vs Predicted Onset Grade(blue)

</td><td>



Actual Duration grade(red) vs Predicted Duration grade(blue)

</td></tr>
<tr><td>

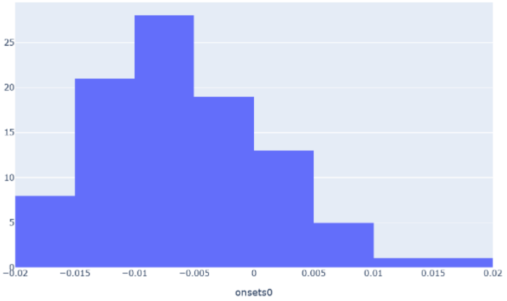
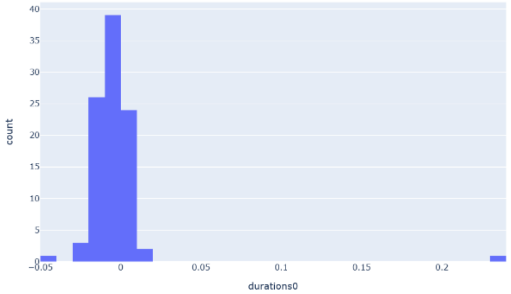
```
      Actual   Predicted

0  79.199997     76.868

1  79.199997     79.540

2  72.000000     92.954

Mean Absolute Error: 7.875

Root Mean Squared Error: 12.174
```

</td><td>

```
      Actual   Predicted

0  79.199997     76.521

1  90.000000     81.158

2  79.199997     70.737

Mean Absolute Error: 6.661

Root Mean Squared Error: 7.234
```

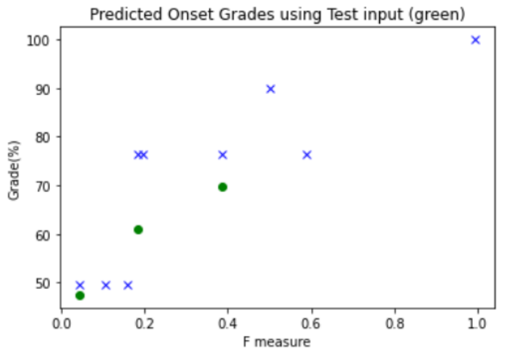
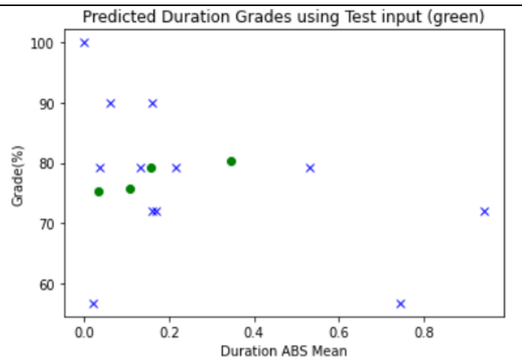
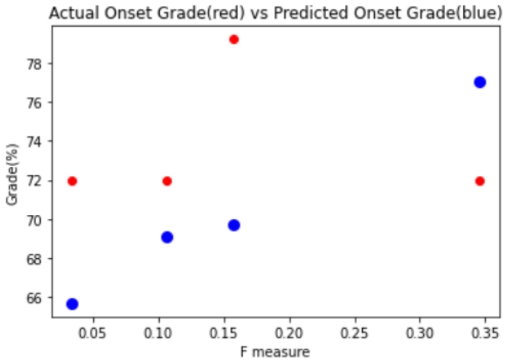
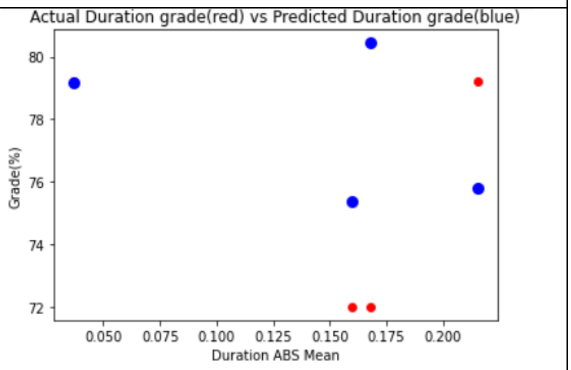
</td></tr>
</table>

Table 19:Road Reference Histograms

Onset Deviations	Duration Deviations
	

Walking On the Moon (Grade 3)

Table 20:Wotm Predicted Grades

Onset Grade	Duration Grade
	
	
Actual Predicted 0 72.000000 77.003 1 79.199997 69.733 2 72.000000 69.079 3 72.000000 65.686 Mean Absolute Error: 5.926 Root Mean Squared Error: 6.385	Actual Predicted 0 72.000000 80.442 1 79.199997 79.164 2 79.199997 75.784 3 72.000000 75.363 Mean Absolute Error: 3.814 Root Mean Squared Error: 4.854

12.4. Student Portal

The following folder contains the PDF generated from the Student Portal Google Form.

https://github.com/cvf-bcn-gituser/bass-critic/tree/main/TeacherGrades

The name of the PDF file is the following:

“Experiment for collecting music performance data for Bass Guitar.pdf”

Section 1 contains the important Latency Test Link

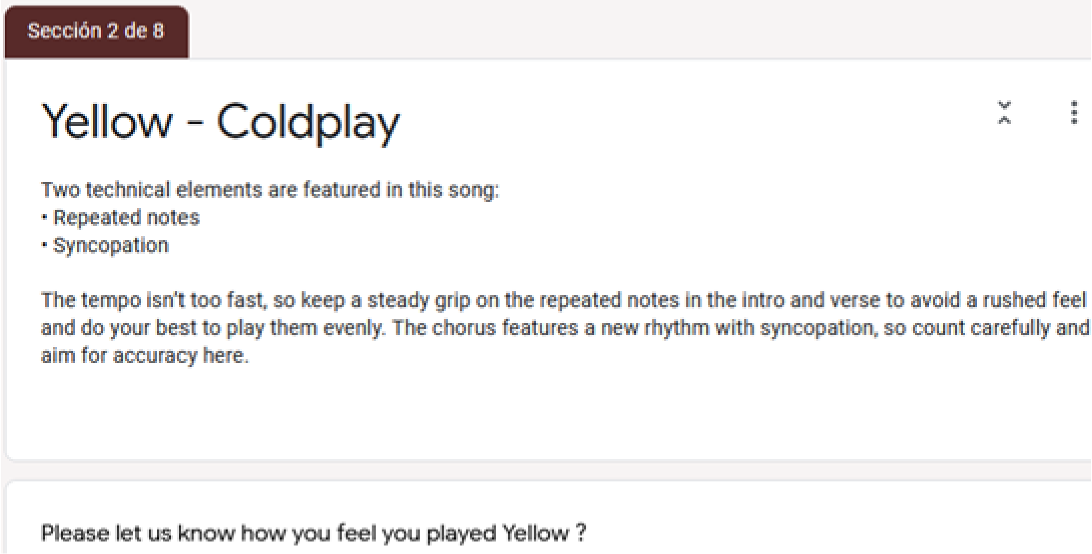
https://musiccritic.upf.edu/training/demo/182

Sections 2 – 8 contain Music Critic links created for each of the 6 R&P songs.

Audio links can be found in the following file:

https://github.com/cvf-bcn-gituser/bass-critic/blob/main/TeacherGrades/Minus1TrackLinks.docx

Each song section has a Technical Description that the student is asked to follow.
In addition they are requested to give their own feedback of the performance.



Sección 2 de 8

Yellow - Coldplay

Two technical elements are featured in this song:
- Repeated notes
- Syncopation

The tempo isn't too fast, so keep a steady grip on the repeated notes in the intro and verse to avoid a rushed feel and do your best to play them evenly. The chorus features a new rhythm with syncopation, so count carefully and aim for accuracy here.

Please let us know how you feel you played Yellow ?

Figure 39: Student Portal Section 2

For full details on the instructions and on how each section is described, please scroll down the PDF “Experiment for collecting music performance data for Bass Guitar”

12.5. Steps for Processing Student Recordings

When all recordings were gathered, these were the steps followed in preparing and mixing the student recordings for Teacher Grading

Steps:

1. Download the student recording from the server.
   It has a name like this: "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 Reference and make a split track to Mono
6. Zoom in on initial onsets and align them manually. Align Student recording with the first onsets of the Reference.
7. Boost the Bass 6db and the Treble 1db and after words add another 1edb to align amplitude with Reference.
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 export as WAV file.
11. Remove the other tracks in Audacity project and export the student stem also as a WAV file.

Any future development that would require the collection of recordings on a large scale would require automating some or preferably all these steps. All the Student recordings can be found in the following link:

https://github.com/cvf-bcn-gituser/bass-critic/tree/main/tclEvaluation/audio

Each of the six folders are named “yellow”, “bjean”, “just”, “brown”, “road”, “wotm”.

12.6. Teacher Portals

The files contain the Questions, Answers and Summary Histogram chart for all the R&P songs. They are PDF exports of the original Google Forms used.
(e.g., The file Grade 0 Yellow.pdf contains the teacher questions for "Yellow". The rest is self-explanatory.

Grade 0 Yellow.pdf
Grade 0 Yellow_Answers.pdf
Grade 0 Yellow_Charts.pdf
Grade 1 Billie Jean.pdf
Grade 1 Billie Jean_Answers.pdf
Grade 1 Billie Jean_Charts.pdf
Grade 1 Just Looking.pdf
Grade 1 Just Looking_Answers.pdf
Grade 1 Just Looking_Charts.pdf
Grade 2 Brown Eyed Girl.pdf
Grade 2 Brown Eyed Girl_Answers.pdf
Grade 2 Brown Eyed Girl_Charts.pdf
Grade 3 Road Runner.pdf
Grade 3 Road Runner_Answers.pdf
Grade 3 Road Runner_Charts.pdf
Grade 3 Walking On the Moon.pdf
Grade 3 Walking On the Moon_Answers.pdf
Grade 3 Walking On the Moon_Charts.pdf

The following extracts are Teacher Histogram gradings for Billie Jean Onsets and Offsets.

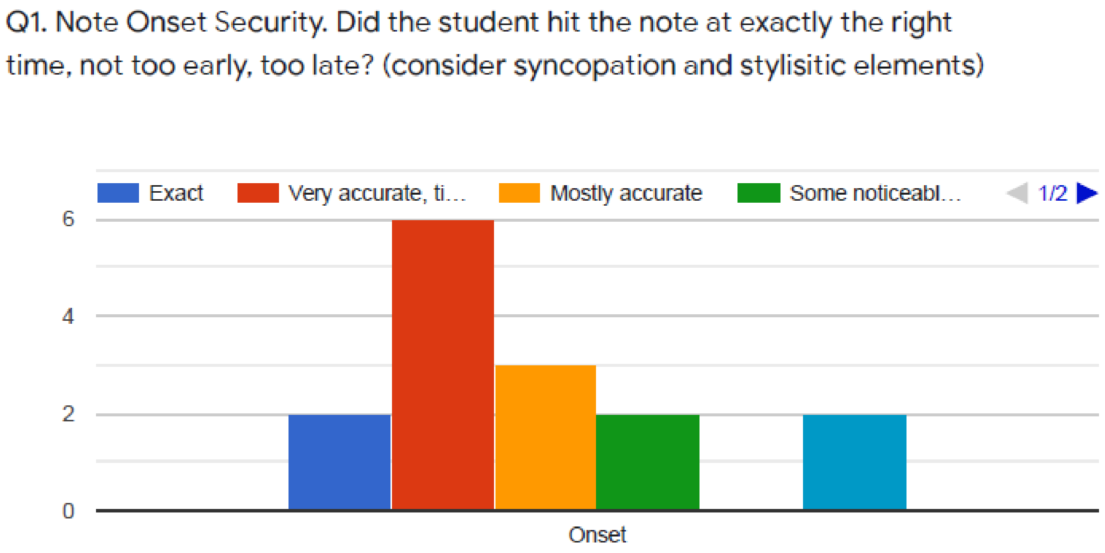
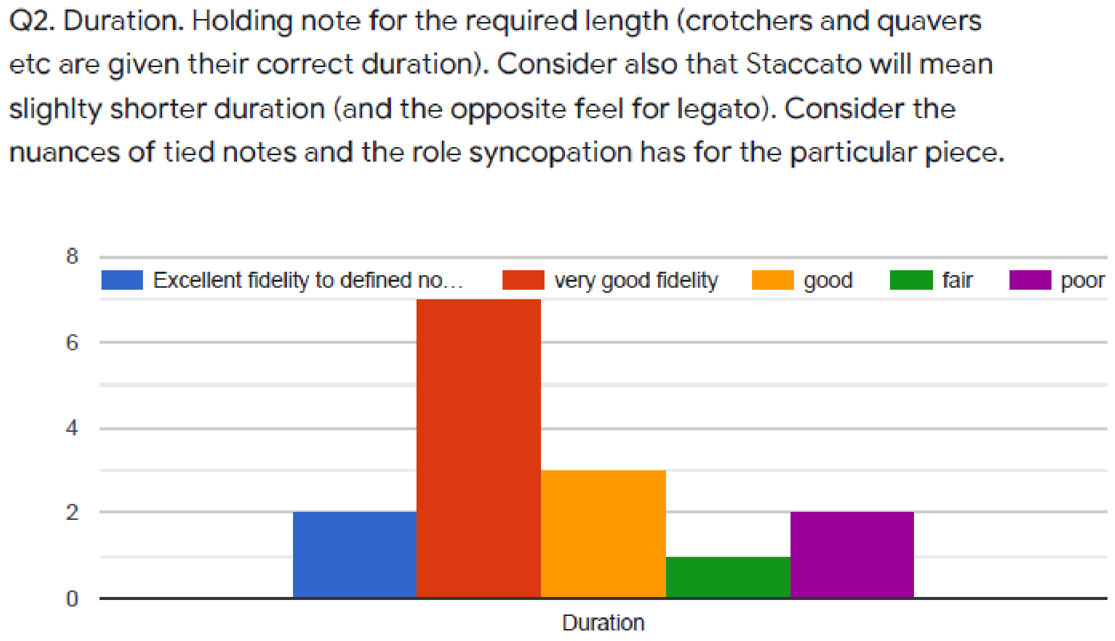

Q1. Note Onset Security. Did the student hit the note at exactly the right time, not too early, too late? (consider syncopation and stylisitic elements)

Figure 40: Grading Histogram Onsets: Billie Jean



Q2. Duration. Holding note for the required length (crotchers and quavers etc are given their correct duration). Consider also that Staccato will mean slighlty shorter duration (and the opposite feel for legato). Consider the nuances of tied notes and the role syncopation has for the particular piece.

Figure 41: Grading Histogram Offsets: Billie Jean

The CSV data exports of the original download CSV files are as follows

https://github.com/cvf-bcn-gituser/bass-critic/tree/main/TeacherGrades/Grades_and_Comments_ESP

The CSV data exports of the English translated downloaded CSV files are as follows:

https://github.com/cvf-bcn-gituser/bass-critic/tree/main/TeacherGrades/Grades_and_Comments_ENG

Grades are the same, it is just the comments that are translated.

12.7. Overall Student Performances

Two actual grades of “49.5” produced two different grades of 42.87 and 63.16 respectively. The failed grades of “36” are from Students 9 and 10 have very low PRF scores as shown in the table. Billie Jean is the song with most Student recordings; however, they are more divergent in quality than other songs.

Table 21: R&P Grades Actual vs Predicted

	ONSET		**DURATION**	
Song.	**Actual**	**Predicted**	**Actual**	**Predicted**
yellow	76.5 76.5 49.5	60.92 69.76 47.44	90 63 76.5	79.76 78.3 66.13
bjean	76.50 68.85 49.50 49.50 76.50	68.76 55.13 42.87 63.16 78.93	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	72.86 73.87 90.70 79.14	90.00 79.2 90.00 79.2	78.19 78.90 92.32 83.46
Brown	79.2 79.0 72.0	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

12.8. Billie Jean Student Performances

The Student Statistics extract shows the PRF and ONSET statistics annotations in yellow and the related grades that were allocated by the teacher.

Table 22: Student Statistic extract "bjean": ONSET focus

Stud.	P	R	F	Abs. Mean	Mean	Std. Dev	ONSET GRADE	OVERALL GRADE
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

Low F-measure 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, then you can do a fair comparison of the Statistics. The next table shows the relevant columns related to offset detection statistics and duration grades.

Table 23: Student Statistic extract “bjean”: OFFSET focus

Stud.	P	R	F	A. Mean	Mean	Std. Dev	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

12.9. Performance of Neural Net methods

The following results were obtained from testing Abesser-Muller [19]Bass U-Net on the full 17 songs of IDMT Dataset. To reproduce these results follow instructions on https://github.com/cvf-bcn-gituser/bassunet for setting up and download code.

Run the command: `$python bassunet_evaluate_17_single_tracks.py`

View sample output in file "*PRF_measure_IDMT_sample.csv*"

Table 24: Abesser Bass U-Net: PRF_measure_IDMT_sample.csv

index	precision	recall	f_measure
1	0.647	0.75	0.695
2	0.5	0.583	0.538
3	0.594	0.586	0.59
4	0.456	0.464	0.46
5	0.524	0.5	0.512
6	0.442	0.523	0.479
7	0.315	0.411	0.357
8	0.412	0.412	0.412
9	0.395	0.425	0.41
10	0.405	0.471	0.435
11	0.515	0.583	0.547
12	0.545	0.6	0.571
13	0.306	0.393	0.344
14	0.589	0.635	0.611
15	0.569	0.635	0.6
16	0.513	0.385	0.44
17	0.222	0.25	0.235

The following result was obtained from testing Abesser-Muller [19]Bass U-Net on the trinity refence recordings of R&P repertoire.

Table 25: Abesser Bass U-Net on Billie Jean Student Recordings

	precision	**recall**	**f_measure_value**
Reference	0.815	0.969	0.885

python bassunet_evaluate_trinity.py:

To reproduce the above table Run the command:

```
$python bassunet_evaluate_trinity.py
```

to get *"PRF_BJ_TRINITY_REFERENCE_MEASURE.txt"*

You also get the generated file "PRF_BJ_TRINITY_measure.csv" student's results as follows:

Table 26: Abesser Bass U-Net on Billie Jean Student Recordings

Student	precision	recall	f_measure_value
1	0.643	0.811	0.717
2	0.381	0.535	0.445
3	0.283	0.357	0.316
4	0.325	0.385	0.353
5	0.303	0.371	0.333
6	0.276	0.28	0.278
7	0.562	0.699	0.623
8	0.42	0.524	0.467

The Bock Online Onset Detector returned the following results, with noticeable high Recall.

Table 27: Bock Online Onset Detector

Song	**Precision**	**Recall**	**F-measure**
Yellow	0.733	1.0	0.846
Bjean	0.713	0.997	0.831

To extract onset on the audio, run program[31] in 'single' file mode to process a single audio file and write the detected onsets to STDOUT or the output file.

```
$ OnsetDetectorLL single <ReferenceAudioFile> [-o OnsetOutputFile]
```

[31] https://github.com/CPJKU/madmom/blob/master/bin/OnsetDetectorLL