Master Thesis on Sound and Music Computing

Universitat Pompeu Fabra

Characterizing Difficulty Levels of Keyboard Music Scores

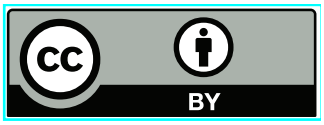
Alia Morsi

Supervisor: Xavier Serra

August 31

Copyright ©2020 by Alia Morsi

This work is licensed under a Creative Commons "Attribution 4.0 International" license.



Contents

1 Introduction **1**

1.1 Project Context 1

1.2 Potential of Technology in Music Education 1

1.3 Project Scope 2

1.4 Difficulty as a Concept 3

1.5 Skill Level Score Categorizations 4

1.6 Research Questions 4

1.7 Contributions 6

1.8 Structure of the Report 6

2 Related Work **8**

2.1 Approaches for Difficulty Analysis 8

2.1.1 Probabilistic Approaches 8

2.1.2 Approaches with Feature Extraction 11

2.2 Score Features 12

2.2.1 Skill/Difficulty Level Characterization 12

2.2.2 Performance Expressiveness 16

3 Methodology **19**

3.1 Data & Resources 19

3.1.1 Trinity College London (TCL) Music Scores 19

3.2 Feature Extraction Approach 21

3.2.1 Feature Selection & Importance 24

3.2.2 Qualitative Analysis . . . 24
3.3 Probabilistic Difficulty Approach . . . 24
3.3.1 Comparisons . . . 25

4 Baseline Feature Set & Results **26**
4.1 Baseline Features . . . 26
4.2 Dimensionality Reduction on Song Level . . . 27
4.2.1 Rock and Pop Keys . . . 28
4.2.2 Classical Piano . . . 29
4.3 Feature Selection & Importance . . . 30
4.3.1 Rock & Pop . . . 31
4.3.2 Classical . . . 33
4.3.3 Do the Feature Results Make Sense? . . . 34

5 Pedagogical Resources Describing Difficulty Levels **36**
5.1 TCL R&P Keys curriculum Own-Choice song Parameters . . . 37
5.1.1 Parameter Categories . . . 37
5.2 TCL R&P Keys Songbook Review . . . 42
5.2.1 Parameter Values for R&P Grades . . . 43

6 Proposed Feature Set & Results **46**
6.1 Proposed Feature List . . . 46
6.1.1 Parameter Coverage of the Baseline Set . . . 47
6.1.2 Proposed Feature List . . . 48
6.2 Dimensionality Reduction on Song Level Features . . . 51
6.2.1 Rock and Pop Keys . . . 52
6.2.2 Classical Piano . . . 54
6.3 Feature Correlation . . . 54

7 Probabilistic Difficulty Results **58**

7.1 Song Level Analysis 58
7.1.1 Rock & Pop Keyboards Curriculum 59
7.1.2 Classical Curriculum 63

8 Discussion & Conclusions **66**
8.1 Discussion 66
8.1.1 Limitations of the Feature Sets 66
8.1.2 Open Ended Feature Suggestions 68
8.1.3 Future Work 69
8.2 Conclusions 70

List of Figures **71**

List of Tables **74**

Bibliography **75**

A Additional Figures **77**

Acknowledgement

I still remember the moment when I found out, by pure coincidence, about the Sound and Music Computing Master's program. After 2 years, much knowledge, and many friendships, I am happy to deliver this Master Thesis, and would like to take a moment to recognize people whom are much appreciated.

I would like express my sincere gratitude to Xavier Serra, for his guidance, patience, trust, and for providing me the opportunity to be part of the ASPLab team from the very beginning. Being part of the lab has been a growing experience in many ways, and I could not imagine my master's journey without it. I would also like to thank Jyoti, Seva, & Rafael Caro, for all the valuable feedback throughout our meetings.

Also, I'd like to specially thank Sergio Giraldo, who has generously extended his time and advice in the early phases of this Master Thesis journey.

Very importantly, I'd like to extend a very big thank you to Ahmed Teirelbar, who I have been very lucky to have as a great friend and mentor since 2013.

To my colleagues, thanks to Jorge Marcos, who in addition to being a great friend, has helped me keep the motivation alive by being an excellent virtual writing companion throughout the earlier lock-down phases. And, thanks Tiange for being great company in those final stressful days. Also, thank you Miguel Casado for your joyful Andalous vibes and helpfulness inside and outside the MTG.

Thanks to Chiara Tiberto, my dearest friend who without which, moving to Barcelona would have been 1000 times lonelier, more difficult, and boring. For sure.

And, last but not least, thanks to my Mom, Dad, and Brother, who have supported me immensely despite all the difficulties imposed by our distance. I am very grateful for everything.

Thanks to all the Master Students, those from 2018-2019, and 2019-2020. It has been a great pleasure.

Abstract

The proliferation online music scores for various instruments and musical styles can be very positive for music learners, who would now witness an increase in score availability for a variety of styles, which could give them autonomy over what styles and songs to invest in learning. However, with the increase of data comes the question of accessibility; how can we aid learners, often without the availability of teachers, to identify good candidates of music scores to learn next given their current skill level? To solve this problem computationally, there needs to be a method by which difficulty can be measured computationally, and effectively. In this master thesis, our goal is to re-visit computational music score difficulty analysis, which as a research problem has been sidelined for some years. We apply the 2 approaches that have been used within the research community on datasets made available to us through the Trinity College London (TCL) examination board. One of these approaches is based on symbolic feature extraction, and the other is based on probabilistic cost. We discuss the strengths and weaknesses of each quantitatively and qualitatively. Moreover, we devote an entire chapter to reviewing through textual content provided by TCL such as information within their songbooks and syllabuses to serve as a foundation for new feature suggestions. 20 features are suggested in addition to the baseline set, and we examine their usefulness by checking if they can characterize difficulty better than the baseline set alone. Despite the new feature having some positive impact, there is still great room for improvement. Finally, after comparing the feature extraction and the probabilistic difficulty approaches empirically, we conclude there is no definitive answer on which is currently more robust. Each approach has its strengths and weaknesses, which are discussed thoroughly, and perhaps the best next step is to combine both approaches.

Keywords: Automatic Difficulty Analysis; Symbolic Analysis; Symbolic Features, Probabilistic Difficulty.

Chapter 1

Introduction

1.1 Project Context

The context within which this project was envisioned is a recent collaboration between the Music Technology Group (MTG) and Trinity College London (TCL). But from a broader sense it is a continuation of the research direction with the aims of supporting Music Education through technology. Previously, the MTG's Audio Signal Processing (ASP) Lab was involved in the TECSOME project (Technologies for Supporting Online Music Education), part of which Music Critic[1] was created. This master thesis is tied with the same end goals of technological enhancement of music education. More specifically, it is concerned with the characterization of music scores based on the skill level required to execute them, one of the first projects within the TCL collaboration which aims to work on projects of technical value to the Music Information Retrieval (MIR) Community and are at the same time, directly applicable to solve current challenges faced by music education institutions.

1.2 Potential of Technology in Music Education

There is a demand for the technological enhancement of music education, evident through the emergence of several applications that aim to support interactive in-

[1] https://musiccritic.upf.edu/

strument learning through engaging experiences[2]. But, despite the existence of these learning contexts, there still remain a vast set of untapped opportunities for improvement of music e-learning experiences. Some improvements could address psychological aspects, requiring studies in human-computer interaction relating to the user experience. Other improvements could address the pedagogical succession of exercises, in terms of their complexity and their target style with respect to a student's level and goals. Exercises should be chosen in a way that maximizes learning while bearing in mind a student's cognitive capacity and what types of exercises are best combined within a learning session. These are only some of many possible open ended questions in how can one expedite progress in music instrument learning. The answers to these questions lie outside the scope of music technology alone. Thus, to be of educational relevance, technologies that claim to aid music education must take into account the plethora of music education research relating to the aforementioned angles (and more), in behavioural research in education, music education, and psychology.

1.3 Project Scope

The aspect that we would like to address in this master thesis is how could we computationally characterize the difficulty music scores, such that we could automatically categorize them by their difficulty. This is useful because especially for intermediate to advanced music learners, it isn't straightforward to identify what music scores would be most suitable to learn next. And, although students can follow the recommended classifications in the syllabuses of different examination boards or music score repositories such as those mentioned in section 1.5, often students want to go beyond these selections, or play music scores of unpublished songs. So, automatically categorizing or quantifying music score difficulty would enable students to have more autonomy over their selection of repertoire, and therefore independence in setting their learning goals in terms of their preferred styles. This problem is espe-

[2]Examples include: www.yousician.com, www.smartmusic.com, www.flowkey.com and www.skoove.com

cially relevant due to the plethora of existing music scores available online[3] [4], where a student may find it confusing to select scores matching their instrument level, and might find it frustrating to struggle through learning a difficult choice that was unsuitable. This master thesis is focused on the keyboard and piano instruments, although this problem is certainly applicable for all musical instruments.

In Chapter 2, we will go through prior work in music score difficulty analysis. There are probabilistic approaches for quantifying difficulty that have obtained interesting results when applied to classical piano scores [1], which we will apply to the music scores available to us. But, we are principally interested in approaches that can help us understand what is it that characterizes a set of scores belonging to the same difficulty level, which makes feature extraction based approaches very relevant since features can be tailored to detect/quantify explicit characteristics or occurrences.

1.4 Difficulty as a Concept

Before moving forward, it is important to reflect on what exactly is meant by difficulty, and whether it can be modelled effectively despite its inherent subjectivity. There are many angles to difficulty. Some music scores might be difficult to read, but not so difficult to learn aurally. Some songs may be difficult to execute due to requiring difficult hand movements, others might be difficult due to the need for emotional sensitivity despite not requiring difficult hand movements. Some songs might require stamina (in terms of hands and emotions) to execute throughout. So, recognizing the different possible 'difficulties' that could exist within a song, one of the primary goals of this master thesis is to try and recognize the importance of these 'difficulties', and how they affect ranking music scores in a pedagogical sense. Regarding objectivity, If we have a set of music scores to rank according to relative difficulty, to what extent would the rankings of participants agree in the order of songs? Would the extent of agreement depend on the expertise of the participants, or the musical styles/genres to which they have more expertise? Would that af-

[3] https://imslp.org/
[4] https://musescore.com/

fect the skills they emphasize the most in their rankings? While we do not have an answer to these questions, which are certainly valid and critical questions, we attempt to circumvent the implications of this subjectivity for this master thesis through using only music scores belonging to the TCL curricula. This is done so that different philosophies of raking song difficulties, or inconsistencies in mapping between different categorization taxonomies, does not confuse a process with an already high potential for subjective pitfalls. The music scores provided by TCL are ranked according to difficulty through belonging to one of 9 grade levels, and refining the selection of songs belonging to a grade level is conducted through an iterative process with many actors (publishers, music experts, and teachers). This provides a degree of security that the final ranking of TCL curricula reflects the voted opinion of many participants of differing backgrounds.

1.5 Skill Level Score Categorizations

Although we mentioned that we will only use music scores from TCL, there are more sources available for level/difficulty categorization. Other than categorizations provided by educational institutions such as TCL, ABRSM, and GCSE, several online sheet music portals (Sheet Music Plus[5], 8notes, and pianostreet) and publishers (Henle) categorize music scores into difficulty levels in order to help students find appropriate sheet music. The majority of such portals are only concerned with classical music. Each of these categorizations is independent, and depends on the philosophy of the publisher/vendor.

1.6 Research Questions

There are mainly 3 questions guiding this work:

- What are the current approaches to characterize keyboard music scores by difficulty?

[5] https://www.sheetmusicplus.com

This will be addressed in Chapter 2 as we go over 2 different classes of approaches that attempt to measure the difficulty of piano music scores. Then, acknowledging that the multifaceted nature of music score difficulty has direct implications on the questions through which we try to understand it, we decided to only be using music scores from the TCL curricula, despite their limited quantities. This is to minimize the clashes between different categorization systems potentially giving different emphases on respective 'difficulties' in each of their skill levels. Our focus is mainly on the TCL Rock and Pop (R&P) keyboards curriculum, but we will also use music scores from the TCL classical curriculum. More information on the curricula is found in 3.1. With that in mind, this brings us to the second question:

- Would the reviewed difficulty characterization approaches succeed in categorizing the R&P Keyboard scores or the classical piano scores according to their indicated grade in the TCL grade level system?

This will be answered separately for each approach in chapters 4 and 7. The benefit of this is 2 fold. First, it allows us to evaluate the reviewed approaches on new data. Also, it provides empirical evidence to help answer the question of what mix of characteristics qualify a music score to a TCL grade level, finally posing the last question:

- Can we use this knowledge, and educational resources provided by TCL, to improve the reviewed difficulty characterization so that eventually, we could assign an unknown music score into one of the TCL exam categories?

A rigid answer to this question was not reached, but chapters 5 and 6 address the question by first thoroughly reviewing TCL resources defining difficulty levels, and performing another attempt for difficulty characterization for the feature extraction approach. Moreover, this question guides the discussions on the limitations and potential improvements which we leave as future work.

1.7 Contributions

In addition to revisiting a research problem that has been marginalized for some time, the main contributions of this work are: 1) Applying both the feature extraction approach and the probabilistic difficulty approaches the TCL classical piano and the TCL Rock & Pop Keyboards curricula, which due to their stylistic differences would make room for interesting comparisons. 2) Providing a discussion on the limitations of each of the approaches with respect to quantifying difficulty in each of the styles through examining both sets of music scores and interpreting the results. 3) Proposing updates to some of the features from prior work, and suggesting the addition of new features to improve difficulty characterization based on the results of 2) and on analyzing TCL educational material, while adopting a more organized framework for viewing score difficulty parameters.

1.8 Structure of the Report

Chapter 3.1 provides a short overview of the available data and processing tools. Chapter 2 provides a review of relevant work, which includes work on skill level and difficulty analysis in general, and then other work on feature extraction applied for end goals other than difficulty analysis, such as expressivity analysis, music genre classification, or music structure analysis. In chapter 3, we describe our methodology, and apply both the feature extraction and the statistical approaches reviewed in chapter 2 to the Classical Piano and the Rock & Pop Keyboard curricula. This is followed by taking a closer look at the respective curricula, and performing an analysis aiming to discuss on why some why some features showed variance across the curricula Grades, while others didn't.

Then, Chapter 5 builds on the discussions in the previous chapter, and analyzes input from several difficulty analysis sources, including the R&P keyboard syllabus, and the textual descriptions in the songbooks. The aim is to update the feature list from earlier works into what we believe would be representative of the difficulty level across the R&P keyboard curriculum especially. Moreover, we compare them

with the most prominent features found based on the works reviewed in Chapter 2.2. Chapter 8 discusses the results from all experiments described in this document, and concludes with a discussion of future work and limitations in the directions that we took.

Chapter 2

Related Work

In this section, we cover prior research related to the difficulty analysis of music scores. Primarily, there have been two approaches: those based on probabilistic cost (section 2.1.1), and others based on feature extraction (section 2.1.2), both using symbolic music data. Since the larger part of our analysis and contribution is in the latter approach, section 2.2 reviews other research in symbolic processing as well, whether or not it is concerned with skill/difficulty level analysis in particular. Symbolic processing was more often conducted for other MIR goals such as expressivity analysis, musicological analysis, cover song detection, and others. It is important to note that difficulty analysis of symbolic data is, to date, quite under explored as a research topic. But despite this, we will try to leverage what has been done previously as much as possible.

2.1 Approaches for Difficulty Analysis

2.1.1 Probabilistic Approaches

Recent work by Nakamura and Yoshii. [1] concerns difficulty analysis, with the goal of reducing ensemble scores into piano scores with performance difficulty levels that can be tuned. From a high level sense, they model this problem as the optimization of the probability $P(R|E)$ of a score reduction R, conditioned on an ensemble score

E, where constraints on score difficulty can be given. To be able to do that, they develop a quantitative measure of score performance difficulty, which is what we are interested in the most.

Score performance difficulty is based on the probabilistic cost of what they refer to as a piano-score model, which describes the probability of a pitch sequence occurring in the score. let the pitch sequence from 1 to N be described by

$$p_{1:N} = (p_n)_{n=1}^{N} \tag{2.1}$$

and, if we are to assume that the probability $P(p_{1:N})$ of the pitch sequence $p_{1:N}$ is first order Markov, then the probability of this pitch sequence can generally be described by

$$P(p_{1:N}) = P(p_1) \prod_{n=2}^{N} P(p_n|p_{n-1}) \tag{2.2}$$

To simplify the idea, let's assume that we have the freedom to define both $P(p)$ and $P(p_{1:N})$ with whichever probability distributions we see fit. Then, then if we define $P(p)$ and $P(p_{1:N})$ to give higher probabilities to simpler choices in the music score, the probabilistic cost of a pitch sequence $p_{1:N}$ would reflect the its difficulty. For example, if the probability distribution of $P(p_n)$ a uniform one, then the probability of any pitch would be the same, and therefore applying equation 2.2 would yield the same probability for all sequences of the same length. In terms of representing difficulty, the assumption of a uniform distribution is not very realistic, but the uniform distribution representation of $P(p_n)$ is one of the three variants used in [1], and is referred to as the No-information model. There are 2 more variants used, all of which are first order Markov. A more sophisticated variant, called the Gaussian model, centers the probability distribution of p_n according to pitch p_{n-1}, which somewhat sensible because closer pitch motions would have a higher probability (and therefore simpler), but has many pitfalls. Finally, the most sophisticated model is one builds directly on the Gaussian model but which takes finger predictions into account. This is referred to as the Fingering model.

However, whichever piano-score model is chosen, the probabilistic difficulty can be calculated as the negative logarithm of the probability of such pitch sequence, as shown below

$$D(t) = -lnP(p(t))/\Delta t \tag{2.3}$$

where $p(t)$ is essentially $p_{k_1:k_n}$ with all pitches p_k occur over Δt. $P(p(t))$ could be any probabilistic model of representing a pitch sequence $p(t)$ over Δt. Finally, in more recent work by Nakamura et al. [2], they employ second order piano-score models in their research.

Evaluation

To evaluate the effectiveness of probabilistic difficulty quantification, difficulty values are computed for a group of scores, and these values are empirically evaluated by using performance MIDI data for these scores. This is done by observing the relationship between difficulty values of score portions and the number of performance errors done by the performers in these score fragments. As shown in figure 1, there is indeed a correlation between score portions with higher difficulty values and the number of performance errors in them, which means that the difficulty is able to predict performance errors, with varying degrees depending on the strength of the piano-score model used in the difficulty calculation. The performance errors considered are one of 3 types: pitch errors (incorrect note performed in place of the correct one indicated in the score), extra notes (note performed without a corresponding score note), and missing notes (a note in the score that wasn't played either by a correct or incorrect note attempt). Difficulty values for the left and right hand parts were calculated on a set of 30 music scores for which there are 90 performances, as each score is performed by 3 players. As shown by figure. The performance errors were manually annotated by error type. By comparing the performance difficulty values (using the 3 pitch sequence models considered) at each onset time t with the rate of performance errors in the time range δt around that onset, they obtain the results in figure 1, showing the relation between difficulty value D_B and the rate of performance errors of the 3 models in question. The above work is certainly unique

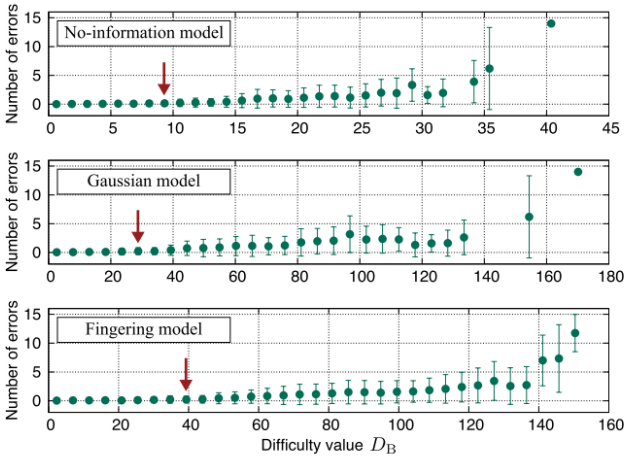


Figure 1: Results of evaluating probabilistic difficulty. No-information, Gaussian, and Fingering models each refer to a particular pitch sequence model. The red arrows highlight the difficulty value in each model after which the error rate starts to increase. Source [1]

in that it measures difficulty using a quantitative measure other than symbolic feature analysis, the approach taken by the majority of other related works. Moreover, it is one of the first attempts of finding an objectively measuring difficulty of music scores through their performance errors. Moreover, the relative success of the fingering model has conceptual implications on difficulty measurement, because it highlights the importance of considering difficulty from the hand perspective. Despite this approach not taking the difficulties of intricate rhythmic combinations into account, or the relationship between the right and left hand rhythms (the pitch sequence of each hand is processed separately), the inclusion of Δt in equation 2.3 implicitly takes tempo into account.

2.1.2 Approaches with Feature Extraction

Other difficulty computation approaches are based on the extraction of score features, which work by concretely defining features that represent difficulty parameters more explicitly. Sébastien et al [3] identify seven dimensions with which they characterize technical difficulties in instrumental performance, and evaluate the relevance of these features by observing whether they are able to group the music scores in the euclidean space. Their data is a set of 50 piano scores from a music conservatory.

The majority of the scores are for intermediate to advanced level students, although some are for beginner students and some are for professional players. The feature list itself is discussed in 2.2.1. Their scores are in MusicXML. After computing the feature values for each score, they use Principal Component Analysis (PCA) and use the first 2 principal components to obtain a 2D projection of the features of each piece. To evaluate their features, they cluster 2D points for the songs, and examine whether the clusters are indeed representative of similar difficulty levels. But first, to get a sense of a number of clusters n that is reasonable, they apply hierarchical clustering using the Ward method, which resulted in a cutoff of $n = 3$. This cutoff is then applied to a k means clustering algorithm, and a qualitative analysis is conducted on the generated clusters to see whether they group music scores of similar difficulty. They concluded that 2 of the clusters indeed group music scores of similar difficulty, but the 3rd cluster groups music scores that have a bass and chord left hand parts, such as ragtime and cakewalk styles.

Chiu and Chen [4] formulate the problem of score difficulty level classification as a regression problem, where their main contribution is a new set of features, some of which are variants of the features in [3], that should be more effective for categorizing difficulty levels. To evaluate their feature set, they extract the features from 2 distinct datasets with manually annotated difficulty values, and calculate the classification accuracy on several distinct feature subsets. Despite the best classification results being quite low (39.9% and 38.8% respectively), we are more interested in their choice of features, and how they compare with those of [3]. Related work by Song and Lee [5] studies music score difficulty through feature extraction.In section 2.2.1, the details of the difficulty features in [3] and [4] are explained.

2.2 Score Features

2.2.1 Skill/Difficulty Level Characterization

The multifaceted nature of difficulty makes its characterization quite difficult. Influential factors affecting the perception of difficulty can vary across several aspects.

For example, polyphony, harmony, melody, rhythm, hand-movements, are all factors that come into play when discussing difficulty. Moreover, there are many situations where the difficulty of the piece is not necessarily consistent throughout. Within the same piece, score difficulty can vary a lot between sections. A method which is able to effectively measure difficulty level has not yet been achieved.

Chiu and Chen [4] aggregate a set of features based on the work of Sébastien et al. [3], in addition to a set of other features as well. Given that the list in [4] is more comprehensive than that of [3], we will explain the features based on the definitions in the former. They used a set of simple features, which they refer to as 'traditional symbolic music features', and more compound features which we will detail in each of the sections below. The traditional features are: 1) Average Pitch Value, 2) Average Note Duration, 3) Deviation of Pitch Value, 4) Deviation of Note Duration, 5) Pitch Range, 6) Shortest Note Duration, 7) Beat Per Measure, 8) Tempo, and 9) Key Signature. While self explanatory in terms of what they entail, there are implementation details that are not clarified in the original paper. For example, it is not clear what the difference between tempo and BPM is. If tempo refers to the tempo marking (adagio, allegro, etc.), is the feature used as a categorical feature, or is it converted into another numeric feature? The same question applies for the Key Signature feature. In the sections below, each of the more compound features is explained.

Playing Speed

This feature is a representation of playing speed. It is calculated by the following equation:

$$\frac{60}{tempo} \times \frac{1}{N} \sum_{i=1}^{n} b(n) \tag{2.4}$$

$60 \div tempo$ is the number of beats per second. $\frac{1}{N} \sum_{i=1}^{n} b(n)$ is the average beat value for the window considered. So, their multiplication can be verbalized as the duration in seconds of the average beat value present in the measure. it's a combination of

the duration value of the notes and the bpm of the piece.

Pitch Entropy

This feature is a representation of the variability of pitch choices over a chosen time/score unit. It can be represented using the following equation:

$$-\sum_{i=1}^{n} p(x_i)\, log_2 p(x_i) \tag{2.5}$$

where $p(x_i)$ is the probability of this pitch occurring given the score window considered, which means that for a considered portion, $\sum_{i=1}^{n} p(x_i)$ would always add up to 1, because the distribution represents the pitches in this segment. In a general sense, entropy tells us how unpredictable a probability distribution is by averaging the entropy of each point in the distribution. This captures, on average, how predictable or unpredictable the notes within every score are window are.

Hand Displacement Rate

It is a representation of the hand displacements in consecutive note (or chord) events, and is calculated as follows:

$$\frac{1}{2N} \sum_{i=1}^{n} DC(d_i) \tag{2.6}$$

d_i is the displacement between different notes in semitones. $DC(d_i)$ can take one of 3 values: $DC(d_i) = 1$ when $d_i > 7$ semitones, $DC(d_i) = 2$ when $d_i > 12$ semitones, and $DC(d_i) = 0$ otherwise. In the case of chords, each note in the chord would be considered separately, meaning that a movement from a note n to a 3 note chord would be calculated as the movement from n to each of the notes in the chord.

Distinct Stroke Rate

A stroke is defined as "a note stroke that occurs while a note or a rest occurs". This feature is calculated using the following equation:

$$1 - \frac{|R \cap L|}{|R \cup L|} \tag{2.7}$$

Where R is a set of the strokes in the right hand and L is a set of the strokes in the left hand. This calculation assumes that that $R \cup L$ would never yield zero. Given the above definition, a distinct stroke in a hand would occur when there is a note event in one hand, while in the other there is a held note or tie. A rest is considered as a stroke. This might make perceptual sense in terms of difficulty. When there are many distinct strokes in a measure, the intersection would be much less than the union, so the ratio $\frac{|R \cap L|}{|R \cup L|}$ would be very small, and the overall calculation would be close to 1.

Hand Stretch

This is calculated as the difference between the average pitch on the right hand and the average pitch in the left hand.

Polyphonic Rate

Proportion of chords in a given excerpt. The calculation proposed by the author does not differentiate between dyads, triads, or larger groups.

Altered Note Rate

Proportion of notes with new accidental information. It is described as the ratio between the number of altered notes over the total number of notes.

Fingering Complexity

Although they don't implement if for lack of data annotated with finger values, they show an example of how finger complexity could be calculated based on the

guidelines defined by Parncutt and Slaboda [6], who, as a part of their work, lay out a set of thorough rules to assign difficulties to fingering choices.

Although the feature extraction and the probabilistic difficulty approaches are presented as two separate measures of difficulty, Nakamura et al. [1] draw a link between both approaches by explaining that their piano-score models implicitly take into account a similar set of features as those in [3] and [4], where the simplest model takes into account note density, playing speed, and pitch entropy, the middle one additionally taking hand stretch and hand displacement rate into account, and the most complex one taking fingering information into account.

2.2.2 Performance Expressiveness

More recent works involving symbolic feature extraction is that of Giraldo and Ramirez [7], and Bantula et al.[8], both in the context of performance expressivity analysis. In [8], the research focus is expressivity analysis in the context of jazz ensemble performances. For example, what is the effect of the piano accompaniment in the performance of a guitar melody, and vice versa. To achieve their goals, they compute a descriptors from the MIDI performances corresponding to a lead sheet (chord and melody indication regarding a piece). Through that, they can calculate what they refer to as 'performance actions', which they consider to represent how the performer interprets the chord. They are:

- Density: the number of chord onsets used to represent one notated chord.
- Weight: the number of notes utilized to execute the chord.
- Range: the semitone distance between highest and lowest note to perform the chord.

In Giraldo and Ramirez [7], the focus is to model expressivity for solo guitar performances of jazz standards. In doing so, they devise a set of 30 features calculated from MIDI transcriptions from guitar performances. Features are classified into 3 main types: 1) Nominal Features: which are features of each note, such as duration,

onset, pitch, chroma, and energy. 2) Neighbour Features: which are features that relate the current note with its previous and next neighbours, such as the intervals between them, and the duration of the neighbouring notes. 3) Contextual Features: Features that contextualize the current note, such as the measure, tempo, key, mode, chord, metrical strength, position in the phrase, etc.

Symbolic Rhythm Extraction

Approaches to rhythm extraction can be applied to a range of symbolic data, including percussion lines of a score and polyphonic music lines. Some approaches consider MIDI note velocity among the inputs to the algorithms that determine what is prominent in the rhythmic patterns, but this is out of our scope. The work in [9] is concerned with how one could extract rhythmic texture from polyphonic scores, and compared the performances of different approaches through their effect on song identification, and music similarity experiments for 2 stylistically distinct corpora with very different rhythmic characters. Their rhythm extraction approach reduces the polyphony to a simple combined monophonic track by considering all onsets of the song. The variations in their approaches stem from the need to form a reliable representation of rhythm in situation where note onsets are played simultaneously across different voices, but with different durations. How to choose which durations most represent this polyphonic set of onsets? For that, they consider 5 approaches to select a note duration to represent the prominent duration in a group of simultaneously played notes: 1) the longest, 2) the shortest, 3) value by notes less than a threshold k, 4) duration value by notes greater than a threshold k, and finally 5) notes on.

Another work concerning rhythmic pattern summarization is that of Coca and Zhao [10], although they use the percussion lines of music scores rather than polyphonic lines or melodic lines. They propose a system for extracting the rhythmic patterns of percussion in a song and for rhythmic summarization in a set of songs of the same genre, artist, and combinations of genres and artists based on community detection in complex networks, based on 'rhythmic cells' of an input song's percussion lines.

They define a rhythmic cell as the units from which the predefined rhythmic patterns defining a song are composed. constructed 2 separate networks: one considering individual rhythmic figures as nodes, and another using rhythmic cells as the nodes of the network. For the network constructed with rhythmic cells, several community detection techniques were applied to determine their effect on forming communities within the network. While this is not directly related with feature extraction, but rhythmic pattern summarization is important to identify the extent of rhythmic diversity within songs, which so far has not been addressed effectively in any of the score difficulty quantification approaches.

Chapter 3

Methodology

The main contributions of this master thesis are: 1) Applying both the feature extraction approach and the probabilistic difficulty approaches (explained in sections 2.1.2 and 2.1.1 respectively) on the TCL classical piano and TCL Rock & Pop Keyboards curricula. 2) Discussing the limitations of each of the approaches through a qualitative analysis of the results. 3) Thoroughly examining educational content from TCL to learn from their thought framework in difficulty parameters, and re-position the baseline feature set accordingly. 4) Proposing new features to augment the baseline set for the feature extraction approach. In this chapter, the data and resources are shown and the conducted experiments are outlined. Then, we show the steps guiding the qualitative analysis on which our discussions on the strengths and weaknesses of either approach are based.

3.1 Data & Resources

3.1.1 Trinity College London (TCL) Music Scores

Our primary source of data is the Keyboards R&P music scores, since this is the difficulty classification scheme that this research is mostly concerned with. We will also use scores from the classical piano curriculum for making comparisons. Whether classical or R&P, a curriculum consists of 9 song books, each corresponding to one

of nine TCL grade level categories. (with 0 being the easiest and 8 being the most difficult) The TCL machine readable music scores are provided in Sibelius format[1], which is a proprietary format tied to the music notation software Sibelius. Sibelius files can be converted to MusicXML or MIDI quite reliably for most score instances, although there have been are cases of conversion errors when using special symbols.

TCL R&P Keyboards Curriculum

The TCL R&P examination category is relatively new compared to other TCL exams. We use the 2018-2020 curriculum for Keyboards. The music scores used are arrangements created specifically for the TCL curriculum, and over the course of the nine grades many styles are covered including synth-pop, rock, blues, ballads, jazz, pop, and reggae. Most of the songs accompaniment style arrangements (with a few exceptions), which makes their challenges different from solo piano scores. The main resources that we utilized are:

- 9 Songbooks, each containing 8 songs. This gives a total of 72 songs, 8 in each grade level. In addition to the machine readable scores, each song has text commentaries describing elements in the song or practice tips.
- The 2018-2020 syllabus [11]. This describes the exam structure, and a summative view of what is expected from the student during the exam. But most importantly, there is a section with musical parameters and descriptions of songs within each grade, which are meant to serve as guidelines according to which students can choose songs from outside the official TCL songbook for their exam.

TCL Classical Piano Curriculum

2 TCL classical piano curricula (2012-2015, and 2015-2017) were used. These songs are compositions for solo piano, from both curricula, there are around 17-24 songs per grade (the number varies), making a total of 198 music scores. Teacher notes

[1] https://www.avid.com/sibelius

are provided with notable features of the songs and practice hints. Since the main use case in this research is the R&P curriculum, the textual material relating to the classical piano curricula were not used.

Symbolic Data Formats

MusicXML and MIDI

In terms of processing libraries and applications to view and edit, Music XML[2] and Music Instrument Digital Interface (MIDI)[3] are arguably the most covered formats in terms of libraries and support, and we have used them both. Music Instrument Digital Interface (MIDI) is a protocol for communicating note events. It is certainly lower level than Music XML, since score information is stored as a set of note events, rather than the more nested structures of the latter. MIDI files typically have the extension .mid. MIDI is used in performance analysis contexts that rely on symbolic data. It very suitable because it would preserve micro-timings without quantizing them to the nearest grid element. It would also preserve expressive information stored as note velocities. Capturing MIDI from keyboard instruments is very straightforward and reliable, unlike other instruments. However, in symbolic analysis contexts that concern music scores rather than performance data, musicxml could often be the preferred format. Music XML is an XML based file format for representing music notation. The feature extraction experiments will utilize Music XML, and the probabilistic difficulty experiments will utilize MIDI. However, both formats were obtained by conversion through the Sibellius software. The music21 library [12] is used for implementing all the features necessary for this research.

3.2 Feature Extraction Approach

Through visualizing the ability of the proposed features to separate songs of different grades when they are represented as points in a 2D space, we can get a sense of the effectiveness of the features and launch further qualitative analysis to identify

[2]https://www.w3.org/2017/12/musicxml31/
[3]https://www.midi.org/specifications

candidates for feature improvements, implement them, and evaluate them similarly. It is an iterative process. The pipeline of the feature extraction approach is shown in figure 2. In this section, we will go through the details of each step.

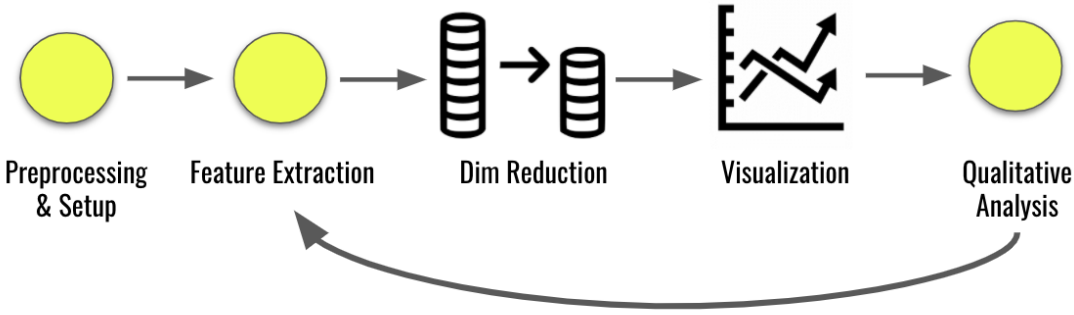


Figure 2: Basic pipeline of the steps involved in feature extraction

Pre-processing & Setup

In this step, we converted the music scores from Sibellius to musicxml, and created a simple visualization setup to observe features values according their locations in the score layout, which was only useful for debugging throughout the implementation of the features.

Feature Extraction

This step was executed twice. The first time when baselining, where the feature extraction functions based on the definitions from [4] and [3], thoroughly described in section 2.2.1, were implemented. The second time was to implement the proposed feature set resultant from interpreting the baselining, reviewing the TCL textual material and songbooks, and proposing a new feature set. The functions were written in python and all the musical processing used the music21 library. For each feature, one value is calculated to represent a song, and this is either a standard deviation of a set of values, the minimum value, the maximum value, or the average value, or some other combination (like the case for the Pitch Entropy feature described in 2.2.1). In section 8.1 we will discuss the advantages and pitfalls of this.

Dimensionality Reduction

We use Principal Component Analysis (PCA) [13], and t-Distributed Stochastic Neighbor Embedding (T-SNE) [14] in order to reduce the dimensionality to 2D and be able to visually examine the effect of adding or modifying the feature set. We employed two fundamentally distinct dimensionality reduction approaches so that any assumptions of linearity by PCA would be avoided by T-SNE, so only failure of both plots to separate the songs of different grades would mean that the feature set needs improvement. Although T-SNE is a more robust dimensionality reduction approach, PCA has the advantage of being more interpretable. With PCA we are able to identify the extent of the contribution of each feature in each of the generated principal components through observing the feature loadings, which allows us to understand the correlations and orthogonalities between features to some extent, and gives insights on the extent of usefulness for features in characterizing complexity. This is appropriate in the context of initial evaluations and data understanding.

Visualizations

After dimensionality reduction, we create 2 kinds of plots to launch qualitative analysis for the development of a better feature set:

- Scatter plot with each song shown a point in a 2D space. This is done for PCA and T-SNE.
- Heat-map of the influence of each feature on the 2 principal components. Applicable for PCA only.

The first allows us to observe whether songs within the same grade lie close together, and if boundaries exists between songs of differing grades in the Euclidean space, and the second allows us to understand the amount of information provided by each feature.

3.2.1 Feature Selection & Importance

This is done through two approaches:

- Feature Correlation: this is used examine the strength of correlations between the features and the grade variable, matrix between the features themselves to detect feature redundancies. Spearman Correlation was used rather than Pearson correlation in order to relax the linearity assumptions by assuming monotonicity.
- Linear Regression: through fitting a linear regression model that predicts the grade level, we can use the weights of each feature in the model to understand their relative importance of the different features.

3.2.2 Qualitative Analysis

Finally, based on the results of feature importance, it is worth taking a deeper look at why some features were not effective or important by trying to look at the music scores and identify cases where the features did not yield the expected results, or successfully capture the intended concept. Qualitative Analysis, in addition to the resources and analysis in chapter 5, is the main foundation on which we shall propose improvements to the baseline feature set.

3.3 Probabilistic Difficulty Approach

To apply the probabilistic score difficulty approaches to our data, we used the code and results of the fingering prediction model by the researchers in [2], which allows non commercial academic use only. The piano-score model used is a version of the fingering model (see 2.1.1) which is second order Markov.

First, we compute the average value of probabilistic difficulty over full songs. And while we expect averaging over full songs lose important many nuances, it still makes a good starting point to get a sense of the effectiveness of this particular approach

on our data. This is done for both Rock & Pop and Classical curricula, we conduct the comparisons across grades and comparisons within grades. Since the provided difficulty computation code expects MIDI scores, we converted the music scores from Sibelius to MIDI.

3.3.1 Comparisons

The purpose of these comparisons is to understand the cases where the same probabilistic difficulty value is observed in songs across different grades. Such songs serve as empirical evidence to understand the strengths and the weaknesses of the probabilistic difficulty approach as defined by its authors. The steps to do so are:

- Compute a box plot across all grades of the difficulty values averaged over songs.
- Find a range of difficulty values with a span across many grades.
- Qualitatively analyze the songs within the chosen difficulty range according by reviewing the music scores with emphasis on trying to explain the reasons for the score differences/similarities. Moreover, in some cases, observe more granular values for the probabilistic difficulty across the music score, in cases the score value has been pushed down/up significantly because of the averaging.

We will not provide an exhaustive comparisons for all scores and all grades. But, following the steps above, we will try to get a good sense of the workings of the probabilistic difficulty approach using empirical evidence.

Chapter 4

Baseline Feature Set & Results

In this chapter, we reiterate the baseline feature list, and then show the results of applying the methodology described in chapter 3 relating to the feature extraction approach of difficulty characterization.

4.1 Baseline Features

The features elaborated in section 4 categorized by musical characteristics are:

- Basic Pitch Info: Average Pitch Value, Deviation of Pitch Value, Pitch Range, Pitch Entropy, & Altered Note Rate.
- Rhythmic: Average Note Duration, Deviation of Note Duration, Shortest Note Duration.
- Tempo: BPM, & Playing Speed.
- Hand Motion: Hand Displacement Rate, Hand Stretch, Polyphonic Rate.
- Harmonic: Key signature
- Coordination: Distinct Stroke Rate

This list comprises the 8 'traditional' features described in 2.2.1 (excluding tempo), and the 7 'compound' features (excluding fingering complexity). The only feature in the above list that was not explained in the state of the art chapter is key signature, and we implement it as follows: If the score has one key signature, then key signature difficulty is a ratio between the number of accidentals in the key signature, and the maximum number of possible accidentals. Then, for every new key signature encountered, the overall difficulty number is incremented with a ratio representing the distance between the new and old key signatures. The calculation can be represented by the following equation:

$$\frac{|A(ks_0)|}{MaxAccidentals} + \sum_{i=1}^{n-1} \frac{A(ks_{i+1}) - A(ks_i)}{MaxDistance} \tag{4.1}$$

$A(x)$ is a function that returns a signed integer for accidentals in the key signature. It is positive for sharps and negative for flats. $MaxAccidentals$ and $MaxDistance$ are constants with values 7 and 15 respectively. Moreover, our implementation of the hand displacement feature is slightly different than that explained by [4]. In the case of hand displacements that involve chords, instead of interpreting them as a group of displacements involving each of the chord notes separately, we will take an average of all the pitches of a chord, so that any displacement, no matter how many notes are in the source or the destination, is a displacement between 2 numbers, where each number represent the average of the pitches played in that moment. We believe this is a sensible simplification because in terms of physical reality of hand displacements.

4.2 Dimensionality Reduction on Song Level

Here we show the results of running PCA and T-SNE on the musicxml scores from the Rock & Pop Keys Curriculum and the Classical piano curriculum of TCL respectively. Although dimensionality reduction itself is not an end goal, it still is an effective tool that allows us to visually examine whether the features can separate the songs according to TCL grades.

4.2.1 Rock and Pop Keys

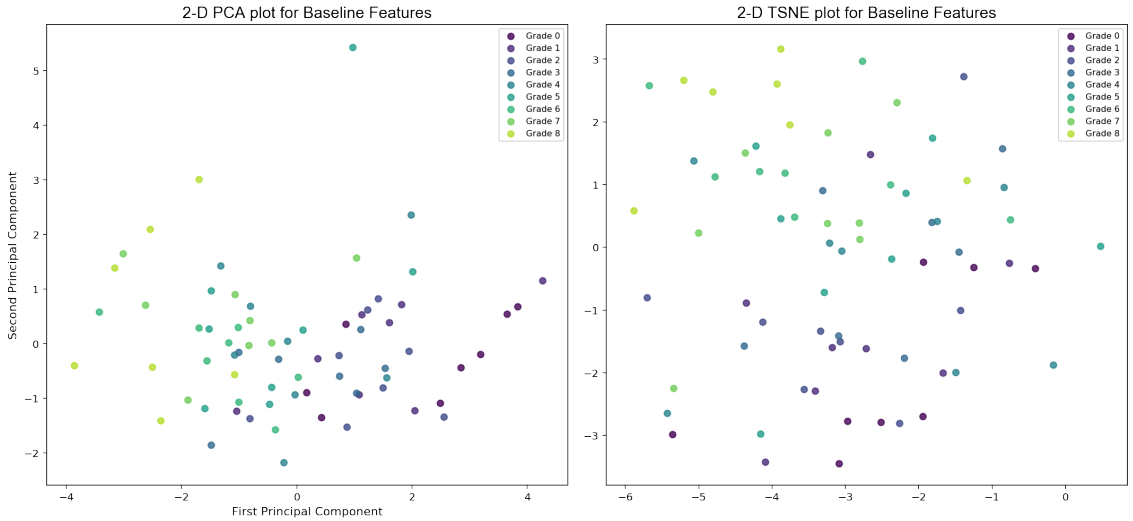


Figure 3: Results of both PCA (left) and T-SNE (right) dimensionality reduction on the R&P Keys curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs

In figure 3, both PCA and T-SNE results are shown. Both were generated by using the scikit-learn library [15]. First, the feature vectors are scaled using the Standard-Scaler algorithm, and then the respective dimensionality reduction is applied.Both algorithms are used with their default parameters in scikit-learn.

In the PCA analysis plot, the first principal component (PC), shown on the x-axis, is the axis of highest variance. Therefore it is a sensible observation that indeed, it is the more effective axis in capturing the gradual progression across the grades. There is not much information retained by the y axis, which represents the second PC, since all grade colors are overlapping in the y axis regions. The 2-D T-SNE plot shows a slight diagonal progression of grades, where the lower grades are more towards the bottom right and the higher grades towards the top left. Nevertheless, in both plots, the results are very unsatisfactory, because certainly no boundaries (with reasonable error rate) can be drawn between the different grade colors in either plot. But despite this, there is more to be observed from PCA analysis in specific. First, we can get a sense of the importance of each feature within each PC through the factor loadings corresponding to that feature in the eigenvectors, as shown in the heatmap of figure 4. Moreover, the variance ratios between the

principal components can inform us of the relative importance of each PC. In case of the baseline set, the first 3 components account for 0.481 of the variance, with individual variance ratios 0.215, 0.166, and 0.099 respectively. Therefore, it is not a case where the majority of the variance is accounted by the first 2 or 3 components only. This means that as a dimensionality reduction approach, the effectiveness of PCA is limited for our data, since it could not return a small number of axes that retain the majority of the variance in the data.

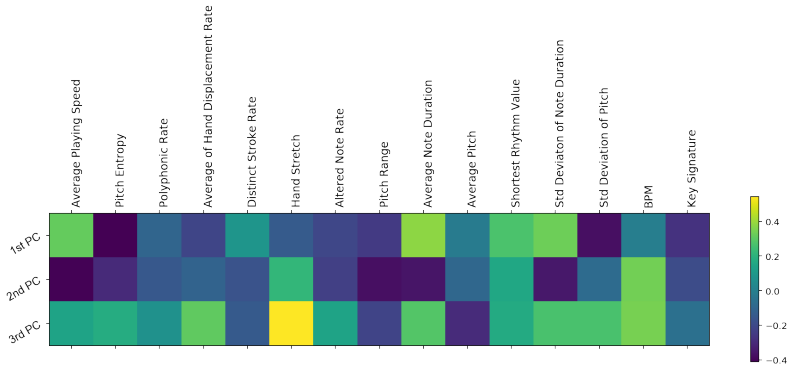


Figure 4: Component Coefficients of the baseline features on the the R&P Keys curriculum for the first 3 Principal Components

The heatmap in figure 4 shows the component coefficients, or factor loadings, of the first 3 principal components. In each PC, the features with the strongest factor loadings are indicated with the lighter colors of the spectrum. In the first PC, the most influential features from the baseline set are to a large extent the features of the rhythmic and tempo categories. The second PC is mainly influenced by BPM, Hand Stretch, and Shortest Rhythm Value features.

4.2.2 Classical Piano

Below are the same plots generated for the TCL classical curriculum. As mentioned in chapter 3, it would be interesting to observe whether these features would be more (or less) effective between the R&P and Classical styles, given the baseline feature set was created with classical piano in mind.

Interestingly, for both plots, the separation of grades is worse for classical than it is for R&P, despite our initial belief that the results for the classical curriculum

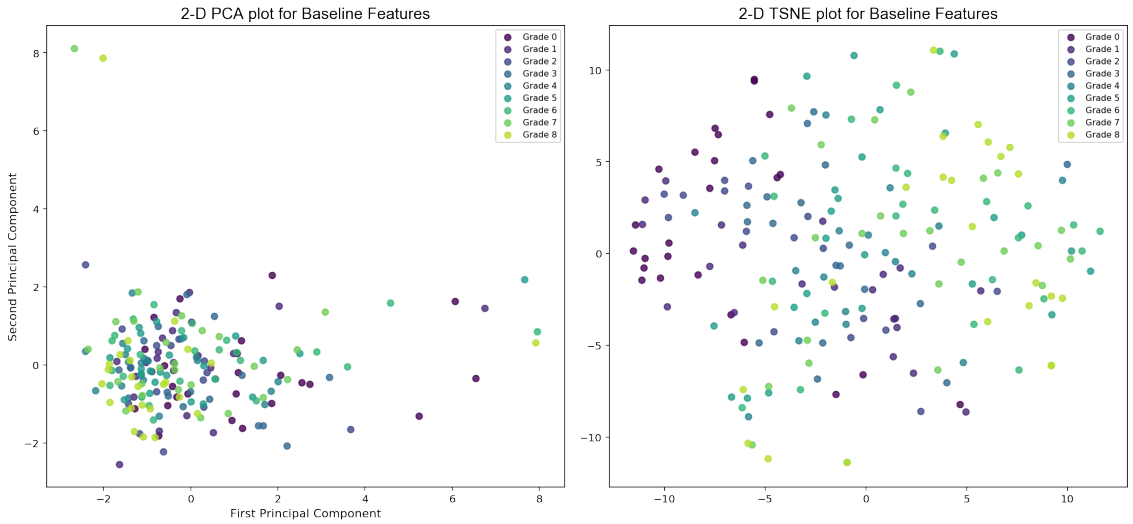


Figure 5: Left: PCA on Baseline Set. Right: T-SNE on Baseline Set

Figure 6: Results of both PCA (left) and T-SNE (right) on the TCL Classical Piano curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs

would be better. Perhaps the songs of the TCL classical piano curriculum are more diverse or are different stylistically than the music scores considered for the work in [3] and [4], on which the baseline feature set was based, but this hypothesis was not tested because their music scores are not available online. In terms of PC variance, the first 3 components account for 0.519 of the total variance, which is not much higher than the case for the R&P curriculum analysis. The variance ratios are 0.232, 0.185, 0.102 respectively. Similarly, the heat map of the first 3 components is shown in figure 7. It does slightly resemble the heat map for the R&P curriculum, but since the 1st and 2nd PC were found ineffective in separating the songs per grade, interpreting the heat map is not very beneficial and is skipped.

4.3 Feature Selection & Importance

Through feature correlation, we will examine the strength of correlations with the grade variable, and through that we will try to rationalize the observed correlations through qualitatively analyzing the music scores. Moreover, we identify potentially redundant features. As already mentioned in 3.2.1, Spearman correlation is used. The full correlation matrices are present in the appendix A, but in this section we

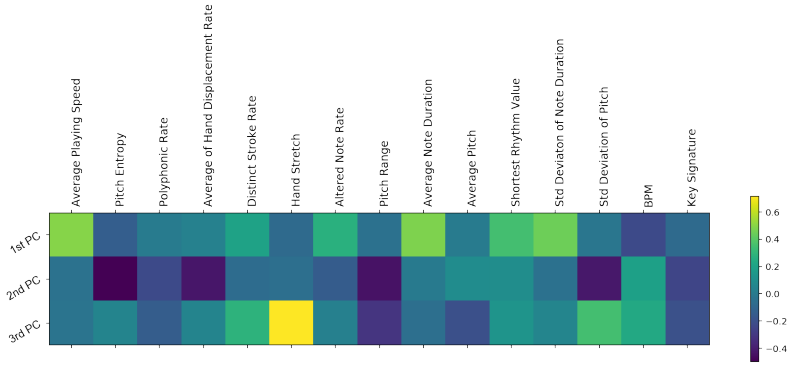


Figure 7: Component Coefficients of the baseline features on the first 3 principal components for the Classical Curricula

will highlight which features are highly correlated with the TCL grade label, and discuss some of the poorly correlated ones.

4.3.1 Rock & Pop

Feature Correlation

The full correlation matrix can be found in figure 29 of the appendix. Figure 8 shows the correlations with grade, and we can see that the top 5 positive correlated features are Pitch Entropy (0.818799), Pitch Range (0.710148), Average of Hand Displacement Rate (0.591861), Std Deviation of Pitch (0.518679), and Key Signature (0.449846). In addition, Shortest Note Rhythm Value (-0.416030) and Average Playing Speed (-0.408915) are the negatively correlated variables with absolute values > 4.

For feature pairwise correlations, Average Playing Speed is strongly correlated with BPM, Average Note Duration, and Std. Deviation of Note Duration (BPM being inversely correlated). This is sensible because based on the definition in chapter 2.2.1, Average Playing Speed is calculated from both of them. So since the pairwise analysis of the features with the grade variable showed an inverse correlation with the Average Playing Speed feature, it would be sensible to drop BPM, Average Note Duration, and Std. Deviation of Note Duration in further analysis related to R&P. Moreover, Pitch Entropy is highly correlated with Standard Deviation of Pitch, Pitch Range, and to a lesser extent the Key Signature feature, which also is

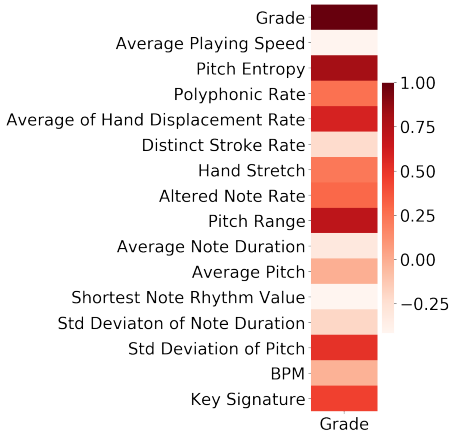


Figure 8: Results of Spearman Correlation between the grade label and the baseline features for the R&P Curriculum

a sensible observation given that conceptually, Pitch Entropy should encompass all three features. Therefore, following the same reasoning, those three features would be excluded if Pitch Entropy is used, despite 2 of them being among the highest correlated with the Grade feature. And indeed, after rerunning the PCA and T-SNE plots for the R&P curriculum (figure 33 in appendix), the resulting filtered baseline set of just Average Playing Speed, Pitch Entropy, Hand Displacement Rate, and Smallest Note Duration, a very slight improvement can be seen.

Linear Regression

To see how the features given most weights in a regression model for predicting the grade of a music score would compare to the observations made from correlation analysis, a regression model was fit on the feature set using all the data points (no train/test split), since our goal is not to build a regression model, and our data points are too few. The scikit-learn library is used for linear regression, and first the features are scaled using the Standard Scaler function, as was done prior to dimensionality reduction. In sorted order, table 1 shows the regression weights that were obtained.

Similar to the features highlighted by the correlation analysis, Pitch Entropy, Pitch Range, Average of Hand Displacement Rate, Average Playing Speed, Shortest Rhythm Value, and Standard Deviation of Pitch are influential features. Features that were

Feature	Regression Weight
Pitch Entropy	1.277820
Avg. of Hand Displacement Rate	0.724053
Hand Stretch	0.670398
Pitch Range	0.596266
Std. Deviation of Note Duration	0.310811
Average Pitch	0.226390
Altered Note Rate	0.178369
Polyphonic Rate	0.131461
BPM	0.047763
Key Signature	-0.102063
Average Note Duration	-0.195540
Std. Deviation of Pitch	-0.489692
Average Playing Speed	-0.500347
Shortest Rhythm Value	-0.509111
Distinct Stroke Rate	-0.574872

Table 1: Linear regression weights for each of the baseline features

assigned relatively high absolute weights despite their low correlation are Hand Stretch and Distinct Stroke Rate, which is interesting since the Distinct Stroke Rate, as will be shown in section 4.3.3, did not capture coordination difficulty very well and therefore is quite a noisy feature. In contrast, features that were assigned lower weights despite their high correlation contribution is the Key Signature. In all cases, these weights are just an rough indication of feature importance, since due to the small number of data points compared to the dimensionality of the features, the linear regression model is not quite accurate.

4.3.2 Classical

Feature Correlation

The full correlation matrix is found in figure 30 in the Appendix. But providing a similar analysis approach to that done for the R&P curriculum, figure 9 shows the correlation of each of the features with the grade label.

Compared to the R&P case, less features displayed an absolute correlation coefficient > 0.4, which we considered our threshold for something meaningful (despite being loose). These features are: Pitch Entropy (0.765137), Pitch Range (0.612504), Aver-

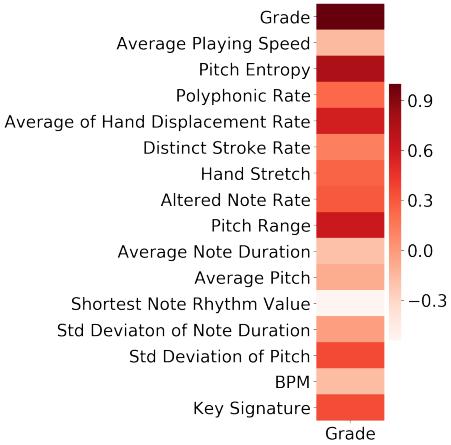
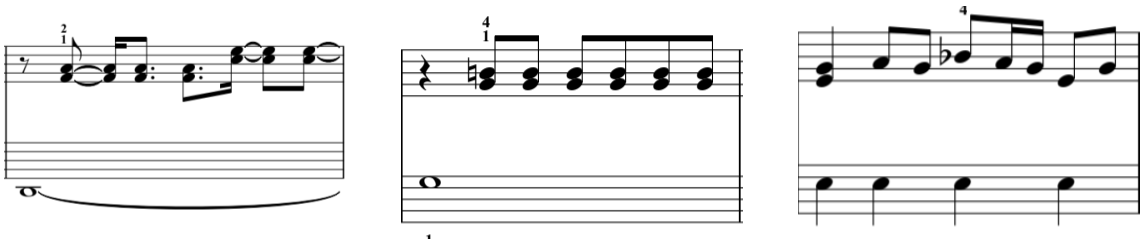


Figure 9: Component Coefficients of the baseline features on the first 3 principal components for the Classical Curricula

age of Hand Displacement (0.583008), and Shortest Note Rhythm Value (-0.541095). They all overlap with the features subset with high correlations for R&P. However, due to our knowledge from the PCA and T-SNE plots in figure 6 that this feature set was very unsuccessful in characterizing the difficulty levels, we will not proceed with further correlation or regression analysis.

4.3.3 Do the Feature Results Make Sense?

Why did some features features perform badly for the R&P scores? This could be due to the feature not effectively capturing the score parameters they should, or due to this score parameter not being significant in the TCL difficulty progression context, or perhaps something else. The features which yielded low absolute values for correlation with grade are Altered Note Rate (0.302866), Polyphonic Rate (0.263362), Hand Stretch (0.230354), Average Pitch (-0.008644), and Distinct Stroke Rate (-0.219721). An obvious limitation of in the Polyphonic Rate implementation is that it treats 2 note, 3 note, 4 note, and 5 note chords as chords, and this is not realistic. Therefore, this will be changed in the new feature set. It was unexpected to find that Altered Note Rate was not successful in capturing the progression across grades, but looking at the results clarified why. The song Golden Brown in grade 5, has a near 0 altered note rate because it has no accidental changes from the key signature (although the key signature has 5 flats). Same for a song like Sinnerman



(a) Grade 3 (b) Grade 0 (c) Grade 2

Figure 10: Snippets from the TCL R&P Songbooks through which we demonstrate why the distinct stroke rate feature is not an effective measure of LH and RH corrdination

which is in grade 7. Contrarily, the song Just Kissed My Baby in grade 2 has an value of 0.12, which is a quite high value (noting that the feature value ranges from order 0.001-0.1) Perhaps this feature would be more meaningful if merged with Key Signature, because when used alone it introduces noise. Concerning Average Pitch and Hand Stretch, it is understandable when thinking retrospectively that applying them on a whole score does yield interesting results. These are examples of features that should be applied on more granular units such as individual phrases or measures, because areas of pitch extremities would be deflated within the average. Lastly, the distinct stroke rate, which we believed would capture coordination difficulties between the right and left hand parts, turned out to have several pitfalls. For example, it shows a very high result in the first few bars of Feel, which belongs to grade 3 (figure 10a), whereas it's not actually difficult to coordinate. For the same reason, the song 96 Tears in grade 0 (figure 10b) will have a high Distinct Stroke Rate (DSR). This is because the union of the RH and LH strokes is large, but the intersection is small. In all examples with a sustained note in one hand, and many notes in the other, DSR will be high, although there is no difficulty in coordination between the parts. The last example, which is Born to Be Wild in Grade 2 (figure 10c) , would have a lower DSR than the earlier two examples, despite being more challenging to coordinate.

Chapter 5

Pedagogical Resources Describing Difficulty Levels

Observing the results of chapter 4, especially figures 3 and 6, and the results of correlation analysis and feature selection, it is clear that there is a large room for improvement such that the chosen feature set can better characterize songs according to their TCL grades. Although that we cannot disregard the possibility that there is a discrepancy in the song difficulties within grades themselves, which means that perhaps the detected overlaps between feature values across songs of different grades may in-fact be a reflection of reality, the qualitative analysis of section 4.3.3 show that perhaps the features fail to effectively capture the cases they were meant to represent, suggesting a need for improvements.

So, to improve our feature set, we conduct a review of the relevant commentaries in the TCL books to narrow down the parameters that could capture difficulty levels from score information, and to potentially inspire new ideas for features which were perhaps not accounted for in [3] and [4]. For TCL R&P, these commentaries are:

- The own choice song parameters included in the syllabus (section 5.1)
- The textual information relating to each song that is available in the songbooks (section 5.2)

Moreover, we will also review the music scores themselves to get a better sense of what the songs of each grade look like, and decide whether the music score contents reflect the song parameters as per the comments. Such resources will be used as input to suggest new features and form a new feature list that could represent difficulty levels better than the baseline set. Then, after implementing them and running them on the available scores for R&P and Classical, the same analysis conducted for the baseline set will be repeated here, to understand whether some of the new features are better correlated score difficulty, or if they are more impactful when used to build a linear regression or as was done in section 4.3.

5.1 TCL R&P Keys curriculum Own-Choice song Parameters

In this section, we review the own-choice song parameters from the 2018-2020 syllabus [11], which are criteria provided to enable students wishing to choose a song from outside the official curriculum to make a suitable choice with respect to the grade of their exam. This allows us to highlight what could possibly be good candidates for new features, and to give us a framework of categories through which we can think of the difficulty of keyboard and piano music scores in a more general sense.

5.1.1 Parameter Categories

The score parameters are divided into 12 categories, and they are explained below along with comments discussing their strengths, weaknesses, and clarity. Some categories are only relevant for earlier grades (e.g rhythmic values and dynamics) , and others for more advanced ones (melodic writing & intervals). The grades for which each category is relevant is mentioned after its explanation.

Duration: Length of the song. Shows more variation from grades 0 - 5, and less variation between grades 6 - 8. Although duration would be a differentiating feature in the TCL R&P context, distinctions based on it would not be a profound reflection

of musical difficulty, and would fail in case we are comparing song fragments of similar lengths but different difficulties.

Rhythmic Values: Covers elements such as size of note durations, rest durations, dotted notes, ties, triplets, duplets, and swung rhythms. If the use of particular sizes of note durations or rest durations was associated to particular grades, it would have been straightforward to write a feature based on the description of this criterion (although this would have made the difficulty characterization overfit for the TCL grades). The descriptions are not sharply defined, which perhaps makes them more realistic. Because, despite our basic intuition leaning towards associating smaller rhythmic values to greater difficulty, in reality the distinction between what is difficult and what is simple is more multifaceted than this. In [11], the term 'intuitive rhythms' is used to describe this parameter for grade 0 songs, and, while ambiguous, reflects the truth that an important part of difficulty is affected by perception. According to the own-song choice parameters, the full range of rhythmic values should have been reached starting grade 5, technically this feature would only differentiate between songs in grades 0 to 4.

Syncopation: While many musicians would agree on what a syncopation is, the case is not the same for determining what a simple or a difficult syncopation would be. This is certainly reflected in the descriptions of the syncopation parameter, which are a bit vague. Hints of how syncopation difficulty progresses along the TCL grades include syncopation frequency, the duration of the syncopated note (e.g quaver, semiquaver). However, some ambiguous descriptions are: 'the appropriateness of the syncopation to the music', 'an increase in the importance of syncopation' in more advanced songs, or even more vaguely, 'an increase in the complexity of syncopation' for more difficult songs. It is clearly an important parameter, and again, like the case for rhythmic values, perhaps the lack of sharp definitions is a realistic way to accommodate for perceptual factors affecting the difficulty according to the student. The syncopation category is used to distinguish songs up until grade 8.

Time Signatures: Relevant for grades 0 to 4, as there is no indicated increase in difficulty starting grade 5. Relevant attributes are: changes in time signature within a song, and regular vs irregular time signatures. However, although some of the descriptions for this parameter are very clear, like the case of grade 3 where it is written that they should have 6/8 or 12/8 time, none of the songs of the grade 3 curriculum actually have these time signatures.

Tempos: BPM mark of song. This parameter reaches full range by grade 3. This would also be a tricky feature to use, because high values would mean that more dexterity is needed to perform the song, but low values do not necessarily mean ease.

Dynamics: This parameter is only significant up until Grade 4, after which the full dynamic range is expected to be reached. Things included are the dynamic states (p, mp, f, etc), sharpness of contrasts, and hairpins.

Range: Despite the initial assumption that this category is mainly about a song's pitch range, this category is in a broader sense concerned with song elements that affect hand positions, which certainly include pitch range but are not exclusive to it. As shown in table 2, the description in the initial grade includes: whether the hands are close together or far apart, movements between keyboard parts, changes of hand position, finger extensions. Unfortunately, it is not clear what is meant by fixed hand positions in the context of the piano or keyboard instruments, and consequently it is not clear which hand position changes or movements are more difficult than others. Does this concern the size displacements? or shape of hand according to the layout of the keyboard's black and white keys? This parameter is relevant to differentiate songs up to grade 5.

Keys: Described as a ceiling for number of sharps or flats that could exist in the songs of a grade, and changes between key signatures. Reaches the maximum ceiling in Grade 5.

Grade	Description
Initial	Hands together, moving between different parts of the keyboard but with time to move; some simple changes of hand position
Grade 1	More frequent movements between hand positions, some finger extension beyond standard positions
Grade 2	Hand beginning to travel beyond fixed hand positions, contained within a range two octaves either side of middle C, occasional stretch to octave.
Grade 3	Extending further into the ledger lines below bass claf stave
Grade 4	Approaching full range
Grade 5	Full use of range

Table 2: Descriptions of the Range parameter across the different grades for the R&P Keys own choice parameters. Source [11]

Part writing: Although this parameter includes a lot of useful information to characterize difficulty, it conflates many different elements together, almost seeming like it is a parameter that includes high level statements about possible descriptions for song characteristics at each grade. In other words, it shows what the songs of a grade would look like from a very general sense. It covers legato/staccato playing, relationship between the RH and LH parts, sparsity or density of textures, chord density, levels of harmonic complexity, repetitions. Table 3 shows the part writing descriptions as per the syllabus, and it shows the usefulness of the information given but also the difficulty of converting them into concrete measures.

Melodic Writing & Intervals: This is the equivalent of the aforementioned part writing parameter, but for higher grades. It is only used to describe songs of grades 6-8. The descriptions are shown in table 4. While the expected number of notes within a chord can be turned into features, different 'textures' and 'keyboard roles', and 'shaped phrases', are not very clear.

Improvisation: In contrast to several parameters lacking importance past Grade 5, improvisation is one of few parameters that appears more prominently in advanced grades. There are 3 types of improvisations: solo, accompaniments, and fills. Unless example transcriptions of solos or improvisations are given in the music score, there is no way by which we can measure the difficulty of the improvisation segment, although metadata about the length of improvised passages could be indicative.

Grade	Description
Initial	Legato playing; simple melodic exchange between the hands, otherwise, very basic LH and spare texture; occasional chromaticism, for example where this fits with a blues scale; two-note chords in RH and occasionally in LH.
Grade 1	Occasional three-note chords, more use of repeated notes and two-note chords, repeated RH accompaniment patterns, still a simple LH.
Grade 2	More clearly defined legato and accented/staccato contrasts, one hand can play up to four-note chords but mostly two- or three-note chords, more silent/empty bars as appropriate, hands becoming more independent with more complex LH.
Grade 3	Faster repeated notes, more irregular accompaniment patterns, greater independence of hands, more textural variety within songs if musically appropriate, melody over sustained notes within one hand.
Grade 4	Octave stretch, thirds in one hand, passing running passages between hands.
Grade 5	More challenging passage-work in both hands, together or moving independently, repeated 6ths, fast repeated notes, layered accompaniments requiring more textural sensitivity.

Table 3: Descriptions of the Part Writing parameter across the different grades for the R&P Keys own choice parameters. Source [11]

Grade	Description
Grade 6	Greater frequency of four-note chords, flowing semiquaver passages, denser chordal accompaniments with a stretch up to one octave in both hands.
Grade 7	Octaves in both hands, including melodic writing and shaped phrases; extended parts featuring lots of different textures and keyboard roles.
Grade 8	Five-note chords (in one hand); parts show a high level of keyboard versatility, with most parts containing an aspect of improvising, melodic or accompaniment passage-work, layered accompaniments, etc.

Table 4: Descriptions of the Melodic Writing parameter across the different grades for the R&P Keys own choice parameters. Source [11]

Other Directions / Techniques: only present occasionally, includes things such as the presence of glissandi and grace notes in Grade 2, ornaments and notated pedal use in Grade 4.

From now onward, we will use these 12 categories as guiding factors, and we will always try to relate difficulty related features to one of them. In terms of the descriptions of what each of the categories entails for the songs of different grades, it should be understood that there could exist discrepancies between some of the descriptions and the actual contents of the songs. So despite being very useful information, it should not be taken in a canonical fashion, and more analytical work is due to test how much the own choice song parameters reflect the contents of the curricula.

5.2 TCL R&P Keys Songbook Review

In the TCL R&P Keyboards Songbook, each of the songs is preceded by information on what a student should emphasize or take note of when performing or practicing the song. This information is typically around 2 paragraphs, so it is not extremely detailed. Nevertheless, below we extract as many hints from such textual information, and we use those in conjunction with our observations from the music scores themselves to document the parameters describing the actual curriculum contents. To be consistent, we will fit the information we encounter into some of the parameter categories described in section 5.1.1. For simplicity, articulations will be placed under the dynamics category, although we understand that under many frameworks they would be two separate things. This was done because there are some markings (such as sforzando, rinforzando) where the indication has implications in terms of dynamics and in terms of articulation.

Nevertheless, even articulation entropy does capture whether the articulations change abruptly (put image of three little birds) or whether they are consecutive regions (put whichever piece I find that has these consecutive changes).

5.2.1 Parameter Values for R&P Grades

Table 5 shows the gathered information from the songbooks of grades 0, 1, an 2, and table 6 shows this information for grades 3, 4, and 5. We have not created such tables for grades 6, 7, and 8 because there are no new difficulty inducing elements added, and instead there is an increased frequency and co-occurrence of elements from the tables of earlier grades, the presence of improvisation (sometimes in long passages), and more performance annotations in the music score itself. Since the tables are dense, we will only be commenting on information in them that warrants additional comments/discussions for clarity. In grade 0, one of the comments for a song is that it 'involves diversity in rhythmic style and playing technique', which causes us to question what could those constitute in the context of grade 0. Through observing the songbooks, we assume a playing style could be a combination of rhythmic pattern, dynamics and accents, fingers moving up/down, phrases crossing left hand and right hand (which should be played smoothly), arpeggiations/broken chords, etc. And while this diversity is important for measuring difficulty, there is no simple metric available to quantify the extent of it.

Beyond the first 3 introductory grades, the comments around the songs change from highlighting warnings or pointers for students to take note of, to offering suggestions stylistic and performative suggestions for the student to achieve the style of the song, and even tips about how to approach the song in the later grades. The idea of textural sensitivity becomes more prominent. For example, the student needs to be more aware of the differing requirements of executing different textures successfully, and which notes should pierce through the overall texture at every moment. In more advanced grades, there are cases where the right hand plays both the melody and chords, or when there are 2 parts within the same hand and they need to be clearly heard as separate. Moreover, related to textural sensitivity, students need to understand when power or emotion rather than aggression are the required meaning of a crescendo (or lightness). Whether it is accompaniment style and requires taking the vocal part into consideration.

(a) Grade 0 Notable Parameters

Rhythmic Values	Range	Dynamics
Ties between bars. Offbeat rhythm (Three Little Birds) Triplets (Gimme Some Lovin')	Thumb crossings. (96 Tears) Hand shifts. (Get Lucky, Something To Talk About) Hand positions in chord changes. (Hello)	Alternating staccato and tenuto. (Three Little Birds) Dynamic contrasts. (Hello) Slurs.
Part Writing		Syncopation
Split staff writing. (96 Tears) Switching between rhythmic styles & playing techniques. (Something To Talk About) Smoothly playing sets of moving 2 note intervals. (Hello) Phrases divided between left and right hands (Are Friends Electric?) Number of chord changes (Are Friends Electric?)		Tie between upbeat & downbeat (Blue Monday)

(b) Grade 1 Notable Parameters

Syncopation	Other	Range	Dynamics
(Gold on the Ceiling, Love is the drug)	Time signature change (Hey Jude)	Hand Position Changes/ Hand Shifts (Crazy, Le Freak, Love is the Drug) Thumb Crossing (Hey Jude)	Crescendo (Crazy, Hey Jude)
Rhythmic Values		Part Writing	
Ties between bars while maintaining left-hand rhythm (Gold on the Ceiling) Tie between bars (Crazy, Gold on the Ceiling, Hold On, Le Freak, Love is the Drug, Mustang Sally) Small note/rest divisions (Le Freak, Two Weeks) Dotted quaver (Love is the Drug)		Right-hand thirds (Mustang Sally) Right-hand Repeated Triads (Two Weeks) Contrast between held chords left-hand and articulated right-hand notes (Le Freak) Phrase evenly divided between left and right hands (Love is the drug) Occasional independence between left and right-hand rhythm (Mustang Sally) Chord changes (Crazy, Hey Jude) Chord complexity (Crazy) Rhythmic Variety (Love is the Drug)	

(c) Grade 2 Notable Parameters

Rhythmic Values	Part Writing	Dynamics
Smallest note and rest unit (Born to be Wild) Ties between off and on beat (In My Place)	4 note chord (Born To Be Wild) Right Hand Thirds (Chandelier) Split staff writing (Chandelier) Arpeggiated Chords (one note and 2 notes) (Chandelier) Octave Leap (LH. Chandelier) Rhythmic Variety across the song (Born to be Wild)	Dynamic Contrasts (In My Place, Chandelier) Crescendo, diminuendo (In My Place, Chandelier) Articulations (Miss you)

Table 5: Notable parameter values observed from R&P Grade 0, 1, and 2 songbook information and the music scores

(a) Grade 4 Notable Parameters

Syncopation	Other	Range	Dynamics
Syncopation at end of measure (Dancing in the Moonlight) Syncopation (Town called Malice)	Time signature change (Back in the USSR, Reelin' in the Years) Grace notes (Dancing in the Moonlight, Reelin' in the Years, Town Called Malice) Pedals (Great Gig in the Sky)	Hand Shifts (I Never Loved a Man) Large pitch range (Knock on Wood) Rising & Falling hand movement (Great Gig in the Sky)	Shifting Accidentals (Back in the USSR, Town Called Malice) Dynamic Variety (Dancing in the Moonlight, Great Gig in the Sky) Crescendos on repeated quavers (I Never Loved a Man) Crescendos & Accents (Town called Malice)
Rhythmic Values		Part Writing	
Small rhythmic units in motif (Feel) Swing 3 time (I Never Loved a Man) Swing (Reelin' in the Years) Ties (Almost all)		Chord Shifts (Back in the USSR) Off-beat playing (Feel, Dancing in the Moonlight, Knock on Wood, Reelin' in the Years) Right and Left sometimes not in sync (Knock on Wood, Reelin' in the Years) Arpeggio (Great Gig in the Sky) Transcribed Solo (Dancing in the Moonlight) Movement in intervals (Feel) Split-Staff writing (Town Called Malice) Four-note chords (Town Called Malice)	

(b) Grade 4 Notable Parameters

Syncopation	Other	Range	Dynamics
Semiquaver syncopation (I Don't Like Mondays) Syncopation to create anticipation feel (Retrograde) Regular Syncopation (The Lovecats)	Glissandos (I Don't Like Mondays) Pedals (I Don't Like Mondays, Retrograde, Vienna) Improvisation (I Don't Like Mondays) Grace Notes (I Don't Like Mondays, Oh! You Pretty Things, Something Got Me Started) Tremolo (I Heard it through the Grapevine) Time Signature Changes (Oh! You Pretty Things) Right hand Improvisation (Something Got Me Started)	Spread out chords (I Don't Like Mondays) Octave intervals (I Don't Like Mondays, Retrograde, Something Got Me Started, Vienna) Legato fifths (I Heard it through the Grapevine) Light 3-note chords (The Lovecats) Fast chromatic rising motions (The Lovecats) Low bass line (Vienna)	Articulations (I Don't Like Mondays, Freedom '90, I Heard it through the Grapevine, Oh! You Pretty Things, Something Got Me Started, The Lovecats) Crescendos with repeated chords (I Don't Like Mondays, Oh! You Pretty Things, The Lovecats) Dynamic Range (I Heard it through the Grapevine, Retrograde, Vienna)
Rhythmic Values		Part Writing	
Offbeat playing (I Don't Like Mondays, Freedom '90, I Heard it through the Grapevine, Oh! You Pretty Things, Something Got Me Started) Ties within and across bars (almost all songs) Dotted quaver rests (Something Got Me Started) Semiquaver rests (Something Got Me Started) Double dotted crotchet (Retrograde) Triplets (The Lovacats, Vienna)		Textural changes throughout song (I Don't Like Mondays, Oh! You Pretty Things, Vienna) Split-staff writing (Retrograde, Vienna) Four-note chords (I Don't Like Mondays) Right hand supporting left hand (Freedom '90) Coordination difficulty between left and right hands (I Heard it through the Grapevine) Unusual Chord Progressions (Oh! You Pretty Things) Arpeggiations (Vienna)	

(c) Grade 5 Notable Parameters

Syncopation	Other	Range	Dynamics
Semiquaver syncopation (Ghost Town) Quaver syncopation (If I Ain't Got You, Shake a Tailfeather)	Time signature change (Golden Brown, Take me To Church) Right hand chordal improvisation (Golden Brown) Both hand rhythm accompaniment improvisation (Golden Brown, Shake a Tailfeather, A Little Help from My Friends) Solo Improvisation (Trampled Underfoot) Grace Notes (Shake a Tailfeather, She's a Rainbow, Trampled Underfoot, A Little Help From My Friends) Pedals (She's a Rainbow) Glissando (Shake a Tailfeather)	Octave Jumps (If I Ain't Got You, Take me to Church, Trampled Underfoot) Octave Stretch (If I Ain't Got You, Shake a Tailfeather, Take me to Church) High pitched crescendo rising (If I Ain't Got You) Rising Chords (If I Ain't Got You) Moving Intervals (She's a Rainbow)	Crescendo (Ghost Town, Shake a Tailfeather, Take me to Church) Articulations (Ghost Town, Trampled Underfoot) Dynamic Range (If I Ain't Got You) Crescendo over repeated chords (If I Ain't Got You, Shake a Tailfeather) Dynamic Contrasts (Take me To Church, A Little Help From My Friends)
Rhythmic Values		Part Writing	
Semiquaver Rests (Ghost Town) Dotted quaver rests (Ghost Town) Triplets (If I Ain't Got You, Take me to Church) Shifts from swing to straight (Take me To Church, If I Ain't Got You)		Off-beat playing (Ghost Town) Left and right hand independence (Ghost Town, Golden Brown, Shake a Tailfeather) Rhythmic pattern variety (Ghost Town) Split-staff writing (She's a Rainbow, Shake a Tailfeather, Take me to Church, A Little Help From My Friends) Broken Chords (If I Ain't Got You) 4 note chords (almost all songs)	

Table 6: Notable parameter values observed from R&P Grades 3, 4, and 5 songbook information and the music scores

Chapter 6

Proposed Feature Set & Results

After reviewing sources of difficulty related content in Chapter 5, and identifying examples of potential new features in section 6.1, a new set of features is proposed. In this chapter, we apply the same methodology described in chapter 3 to the new features, in order to compare its results with those of the baseline set, and to understand which of the new features are beneficial for difficulty characterization, and which aren't.

6.1 Proposed Feature List

Adopting the parameter categories of the TCL own choice songs, the proposed features are presented in table 7 according to the same categories, where we use 6 of the total 12 categories. Namely, we add features in the duration, rhythmic values, range, dynamics, part writing, and syncopation categories. Before proposing additional features, the baseline feature set is revisited in light of the aforementioned TCL parameter categories, which we do in section 6.1.1. Then, in section 6.1.2, we suggest features to fill in some of the gaps that weren't covered by the baseline set. Discussions on the feature values can be found in 8, where we discuss limitations of the new feature set and proposals for open ended features which will be left for future work.

6.1.1 Parameter Coverage of the Baseline Set

While some of the baseline feature set described in chapter 4 addresses the straightforward parameters such as tempo and key signature, many pieces of information from the other parameters are not addressed effectively. To identify key missing points, the following list is created where each relevant TCL parameter is linked to features in the baseline set that implicitly capture parts of it.

- Range: some information captured by the average pitch value, std. deviation of pitch, hand stretch, pitch range, and hand displacement rate. However, none of the baseline set covers information about standard positions/finger extensions, which were mentioned in 5. Moreover, the hand stretch feature of the baseline set only represents the distance between the hands, and not the difference within each hand itself.

- Rhythmic Values: Slightly covered by shortest note duration value and std. deviation of note duration features. But nothing to capture swung rhythms, ties, triplets, or duplets.

- Part Writing: Lightly captured by the distinct stroke rate in its attempt to detect portions that are difficult to coordinate between LH and RH, but, as discussed in section 4.3.3, distinct stroke rate failed to represent coordination in some corner cases. Moreover, polyphonic rate and pitch entropy capture partial information about polyphony and pitch diversity respectively, despite the need to add another measure for polyphony that distinguishes between differing chord densities. Nevertheless, there is much that is not explored, such as detecting chromaticism, accompaniment patterns (whether repeated or irregular), articulation contrasts, or layered accompaniments. To provide a good coverage of the part writing category, there needs to be more mid/high level feature usage.

Features	Parameter
Song duration in seconds	Duration
Smallest rest duration Std. Deviation of rest durations Average rest duration Complex Durations Number of ties across bars Number of ties within bars Number of all ties Tuplets (eg. triplets or duplets)	Rhythmic Values
Difference between the highest and lowest notes in LH, Difference between the highest and lowest notes in RH Chord compactness/sparseness	Range
Articulations rate Articulations Entropy Dynamic changes Dynamic range	Dynamics
Average chord density Max chord density Key Signature Offbeat/Onbeat ratio* Split-Staff writing* Note density per time & stats*	Part Writing
Number of syncopations Minimum duration of syncopated note*	Syncopation

Table 7: Proposed features by parameter. Those suffixed with * are not implemented

6.1.2 Proposed Feature List

Table 7 holds suggestions for new features organized according to the parameter categories in which they belong. It is important to note that this is not a comprehensive list of all features that could potentially be used, since we will not attempt to tackle many of the ambiguous parameters, nor parameters relying on Music XML elements that could be inconsistent to represent (such as the pedal marks and other elements belonging to the other directions category). Below, each of the proposed features will be explained in more detail. Some additions are straightforward, such as duration, time signature, and dynamics, while others are not so straightforward and represent parameter categories that encompass several information points.

Parameter: Duration

Song duration in seconds. Although important in the TCL context, the problem is that it doesn't capture profound difficulty from a musical sense. It is straightforward to calculate with the music21 library.

Parameter: Rhythmic Values

Despite the realization that rhythmic difficulty is affected by perceptual elements of intuitiveness of the rhythmic pattern, the features we will use to represent rhythmic difficulty do not capture this. We only use nominal features such as the rhythmic values themselves and statistical deviations of this.

Rest Durations: There are 2 features in that category: smallest rest duration, average rest duration, and the standard deviation of rest durations.

Ties, Complex Durations, and Tuplets: Since ties were explicitly mentioned in the own choice parameters and in the song commentaries. In the introductory grades, ties between bars was mentioned several times, and eventually in the later grades, it becomes a very regular occurrence. Therefore, it might be interesting to count the ties between the bars and ties within bars, potentially as a representation of compoundness in the rhythmic phrases, and hence difficulty. But ultimately, the presence of ties is a reflection of complex durations, so we want to return the number of complex durations as well.

Parameter: Range

This parameter has more potential than is captured by our proposed features (especially in terms of hand positions and finger extensions). The new features are very simple: pitch range within each part separately (the right and left hand parts), and chord spread. is another which will identify the average chord spread throughout the song. It would be calculated by getting the average semitone interval distance between highest and lowest pitch of every chord in a song.

Parameter: Dynamics & Articulations

Previously unrepresented in the baseline set, dynamics & articulations will be represented with 3 features:

Dynamic Changes & Dynamic Range: The dynamic changes feature is meant to capture the average magnitude of dynamic difference between consecutive marks, as shown in the equation below:

$$\frac{\sum_{i=1}^{n-1} f(D_{i+1}) - f(D_i)}{n-1} \tag{6.1}$$

D_i is the ith dynamic mark, n is the total number of dynamic marks in the score, and $f(D_i)$ is a function that converts between the ith dynamic mark and a numeric value. $f(D)$ converts dynamic marks into a number from 1 (ppp) to 8 (fff) reflecting its loudness. rfz, rf, fz, and sfz marks are treated as f. Moreover, compound dynamic marks such as 'fp' will be treated as 2 consecutive marks, although this does not capture the additional difficulty of the abruptness of the contrast, since this measure of dynamics does not take into account the distance between consecutive measures. . Based on the same $f(D)$ in the dynamic changes feature, the dynamic range function just returns the difference between the maximum and minimum dynamic marks.

Articulations Rate & Entropy Given that the articulation palette increases the higher up the grade, we chose 2 features to represent this: the articulations rate and articulation entropy. Articulation rate is agnostic to the type of articulation, it is just a ratio between the number of notes with and without articulation marks. However, since as a measure this does not take into account the diversity within the articulation marks themselves, another feature is needed to capture this, and therefore articulations entropy is used. Inspired by the pitch entropy feature of the baseline feature set, is meant to summarize the extent of articulation predictability within a music score. . It can be defined as:

$$-\sum_{i=1}^{n} p(a_i)\, log_2 p(a_i) \tag{6.2}$$

where a_i is the probability of a given articulation calculated from the number of notes applicable to it within a song.

Parameter: Part Writing

Average Chord Density & Max Chord Density Given that the polyphonic rate feature of the baseline set does not differentiate between differing chord densities, its calculation is updated in the updated features set so that it reflects the number of notes in a chord.

Ratio between Offbeat and Onbeat notes The ratio between on beat and offbeat notes could give a nice feeling of rhythmic difficulty throughout the song.

Parameter: Syncopations

Despite the realization that syncoptions should be classified based on different reflecting their difficulty, based on the TCL own choice song parameter reviews, for this iteration of features, only the total number of syncopations in the score is used. We define a syncopation to be any note that starts on a weak beat, and a main beat is crossed within the total duration of the note (whether it is a normal, tied, or dotted note), whether it is represented using ties or dotted notes.

6.2 Dimensionality Reduction on Song Level Features

In this section, we show the results of running PCA and T-SNE on the Rock and Pop Keys Curriculum and the Classical piano curriculum respectively using the features that were implemented from the proposed feature list. The same pre-processing and algorithms as section 4.2 are applied. As an attempt to see the effectiveness of the newly implemented features, we will create the visualizations first on the new feature set only, then on the aggregated feature set (which includes both the new feature set and the baseline feature set).

6.2.1 Rock and Pop Keys

New feature Set only

The PCA plot with the new feature set shows similar performance to that of the baseline features (figure 3), but the T-SNE one shows an improvement in comparison. Figure 12 shows the contribution of each feature in each of the principal components. The first 3 components account for 0.498 percent of the variance in the principal components with individual variances 0.2681, 0.1204, and 0.1097 respectively.

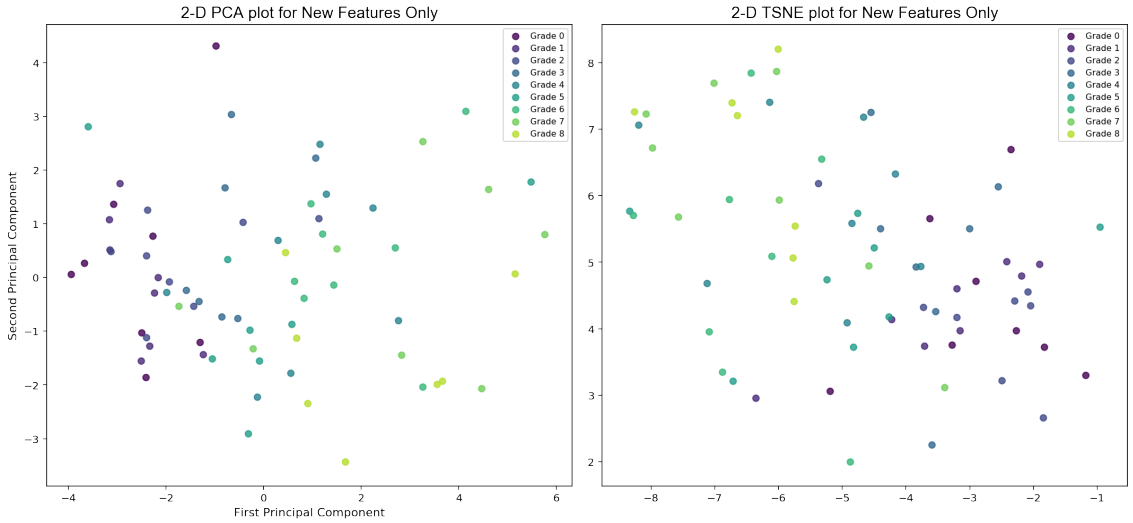


Figure 11: Results of both PCA (left) and T-SNE (right) dimensionality reduction on the new feature set calculated from the R&P Keys curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs

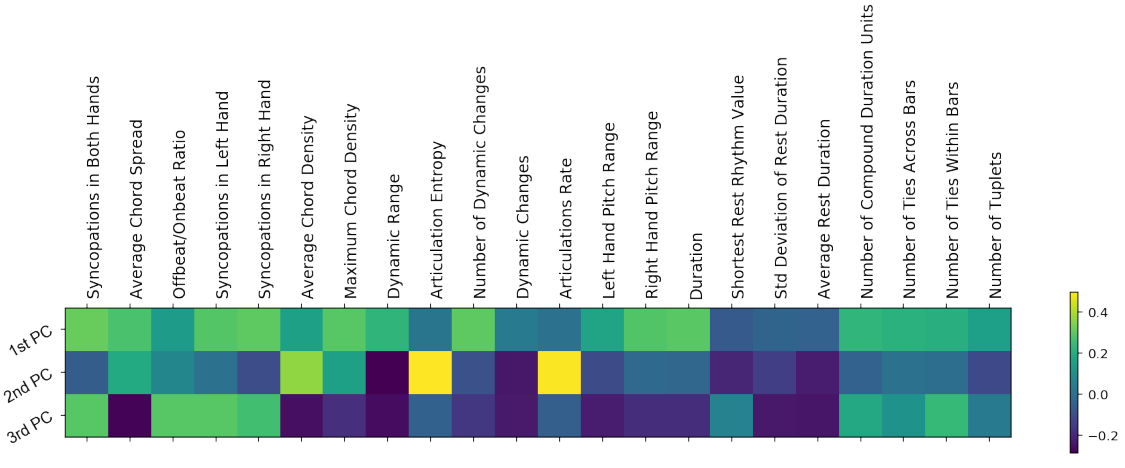


Figure 12

Total Feature Set

Both the PCA and the T-SNE plot for the total feature set (figure 13) show the best result. In the PCA plot, It seems that the majority of the points for grade 0 and grade 1 (the 2 darkest colors) were localized in a small region, but for the higher grades, still only the first principal component is the effective one. The first 3 principal components account for 0.401 of the variance, each having individual variance 0.208, 0.108, and 0.083 respectively.

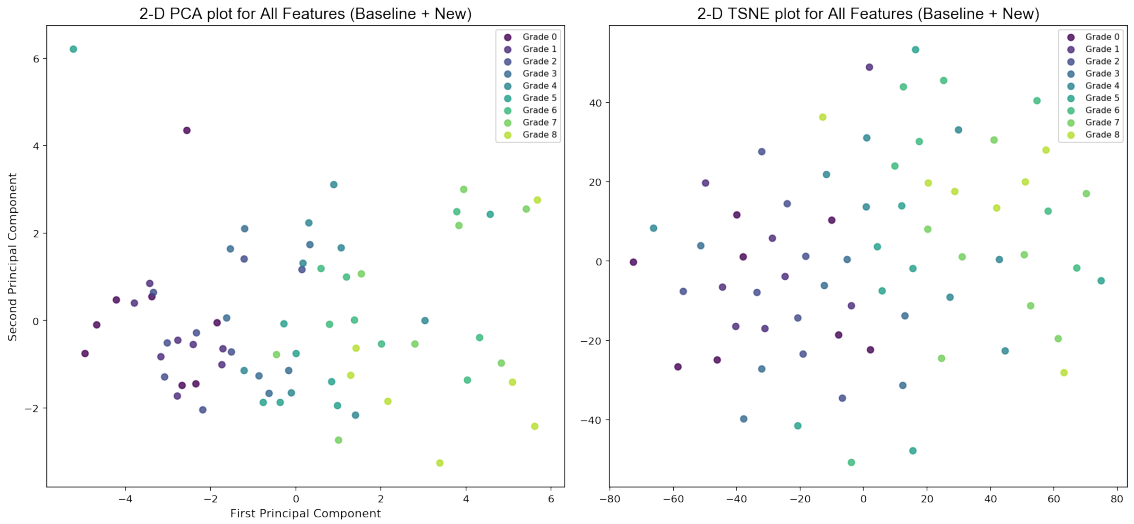


Figure 13: Results of both PCA (left) and T-SNE (right) dimensionality reduction on the total feature set (baseline + new features) set calculated from the R&P Keys curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs

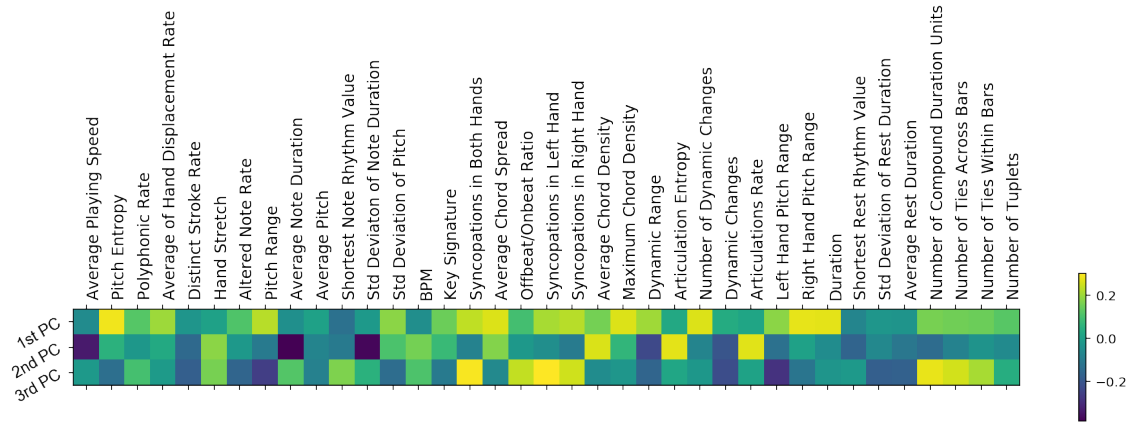


Figure 14

6.2.2 Classical Piano

New Feature Set & Total Feature Set

For the classical piano plots, it is clear that the addition of the new features did improve both the PCA and T-SNE plots, which in comparison to figure 6 that showed very minimal if no separation of the colors across the 2D plane, does show an improved visual separation of the colors (despite still not being robust enough). Figure 15 shows the results calculated with the new feature set only, and figure 17 shows the results calculated with the total feature set (baseline features + new features).

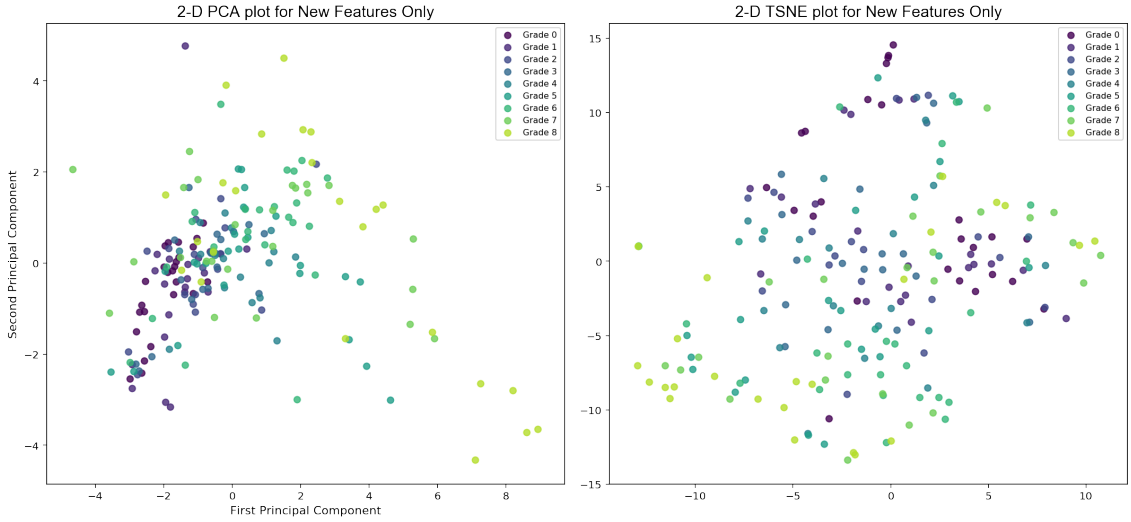


Figure 15: Results of both PCA (left) and T-SNE (right) dimensionality reduction on the new feature set calculated from the Classical Piano curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs

6.3 Feature Correlation

Rock and Pop Keys

The correlation plot for the the total feature set can be found in figure 31 of the appendix. In figure 19, we show the correlations between the grade label and the new features only (since the baseline set was already covered in section 4.3.2)

If we take a threshold > 0.5 this time, then the 7 features in table 8 are the most

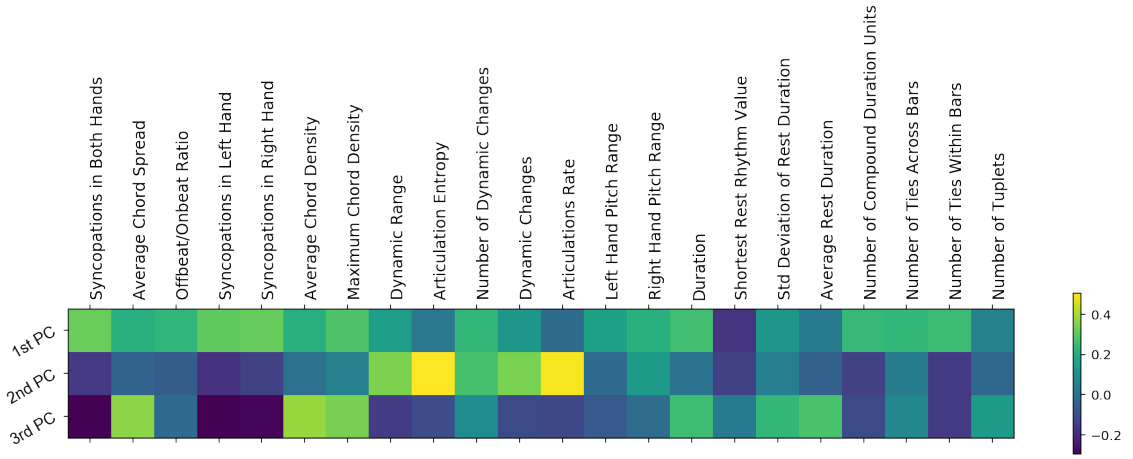


Figure 16

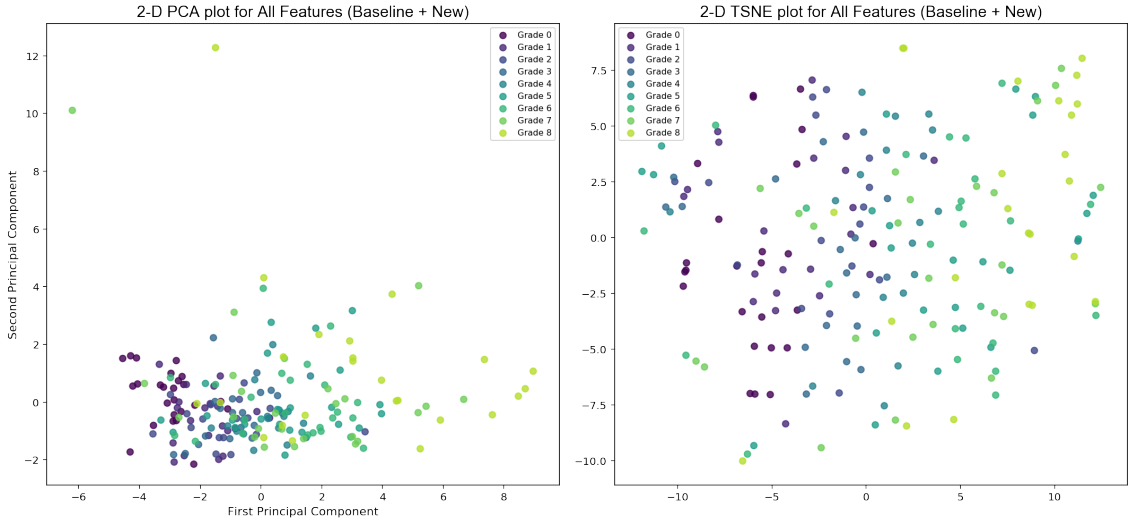


Figure 17: Results of both PCA (left) and T-SNE (right) dimensionality reduction on the new feature set calculated from the Classical Piano curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs

influential.

Classical Piano

Full correlation matrix can be found in figure 32 of appendix. The heatmap of the correlation with the grade feature is shown in fig 20. For the classical case, however, the results are quite worse, which is no surprise given the dimensionality reduction plots shown above. The only features with correlation > 5 are right hand pitch range and left hand pitch range, with correlations 0.562709 and 0.619616 respectively.

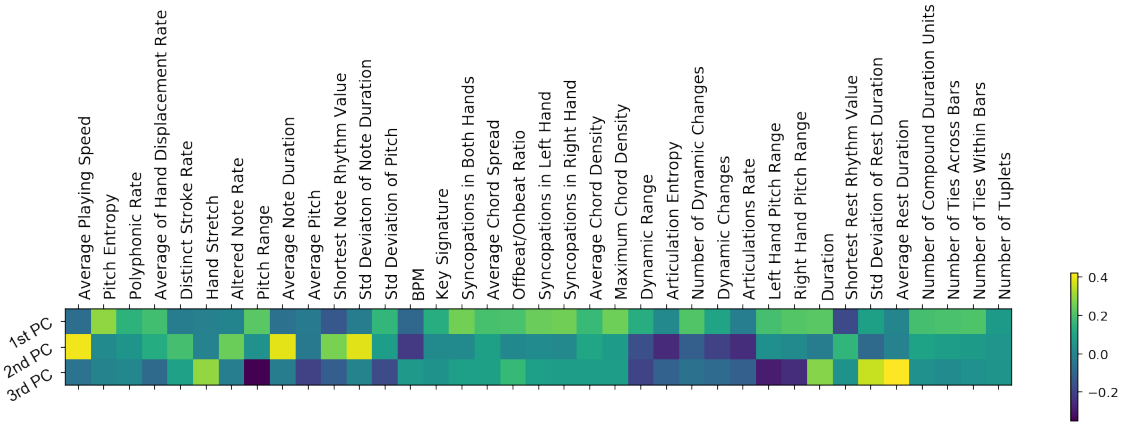


Figure 18

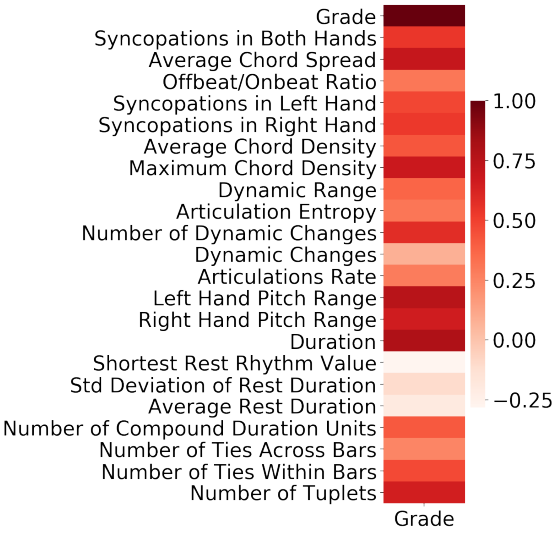


Figure 19: Feature Correlations with Grade for the new feature set for R&P Keys

Feature	Correlation
Syncopations in Right Hand	0.532744
Syncopations in Both Hands	0.540061
Number of Dynamic Changes	0.583760
Number of Tuplets	0.653059
Right Hand Pitch Range	0.662893
Maximum Chord Density	0.683675
Average Chord Spread	0.703144
Left Hand Pitch Range	0.757405
Duration	0.797441

Table 8: Results of Correlation with grade for the new feature set on R&P

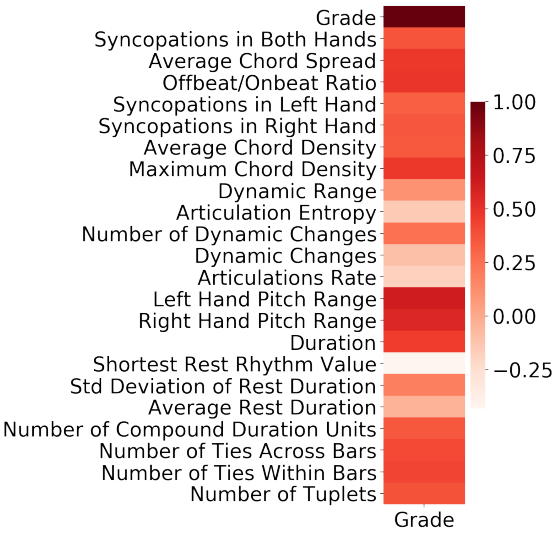


Figure 20: Feature Correlations with Grade for the new feature set for Classical Piano

Chapter 7

Probabilistic Difficulty Results

In this section, we describe the results of the probabilistic difficulty measurement on TCL data, as described in 3.3. The launching point for the probabilistic difficulty comparison is by calculating the overall probabilistic difficulty score for each song, and observing the shape of the trends across grades and within grades. We do this for both R&P and Classical Curricula. The code used to calculate probabilistic difficulty was given to us by the researchers of [2]. An explanation of the probabilistic approach for difficulty computation is found in chapter 2. The code uses a second order Markov model that takes fingering information into account. Especially after the review of TCL difficulty parameters, the value of incorporating fingering information is more prominent because potentially, it implicitly includes information relating to finger extensions hand positions, and hand motions, which so far were not covered at all by either of the baseline or new feature sets.

7.1 Song Level Analysis

Scores are compared first on the song level, by which we mean that we calculate one difficulty score per song. Difficulty is calculated for 1 second intervals, for each hand separately, calculated from the MIDI representation of the song. Then all the results for both hands are averaged into a single difficulty score, which is a combination of the left hand and right hand part difficulties, over all the 1 second intervals in a

song's MIDI file.

7.1.1 Rock & Pop Keyboards Curriculum

In figure 21, a box plot and a swarm plot are combined. The swarm plot is useful to allow the visualization of the points corresponding to the individual songs, which are shown by the black dots.

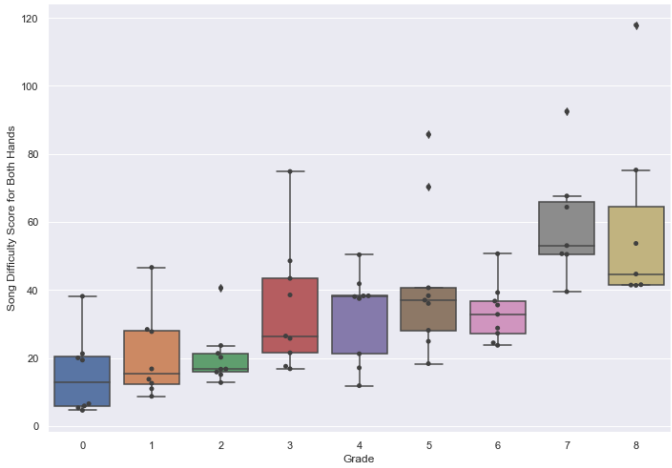


Figure 21: Probabilistic Score Difficulty for songs in the R&P curriculum based on the second order fingering model described in [2]. It is a combination of a box plot and a swarm plot. A black dot represents the overall difficulty score of one song

A distant view at the box-plot seems to indicate a rising trend to an extent. The bottom part of the box plots generally rises (except for grade 4, where the song with the lowest score is lower than that of grade 3). The ceiling of difficulty for grades does not rise as gradually as the floor, although if we take the outliers of the box plot into account, the trend is rising with the exception of grade 2, 4, and 6. However, one clear observation that is somewhat problematic is that the range of difficulty values is very large within each grade, such that there are large overlaps across the different grades. For example, a query of the results shows that a song with an overall both hands difficulty ranging from 21 to 25 can exist (albeit with differing probabilities) between any of grades 0 to grade 6, which is too large of a grade window. It is not completely unreasonable to have some songs in a grade be considered equally or more difficult than a few of the songs in the grade higher than it, but with a maximum of one or maybe 2 grades higher, otherwise this would

Grade	Song Name	Difficulty Score
0	Gimme Some Lovin'	21.2795
2	Born To Be Wild	23.6628
2	Chandelier	21.3773
3	Knock On Wood	21.5266
4	Born To Run	21.2360
5	She's A Rainbow	24.8982
6	Easy	23.7310
6	The House Of The Rising Sun	24.4531

Table 9: Distribution of Songs with Song Level Difficulty Scores Ranging from 21 to 25 in the TCL 2018-2020 R&P Curriculum

reflect an incoherence in the curriculum itself.

In order to gain a better understanding of the strengths and weaknesses of the probabilistic difficulty measure, we conduct a qualitative analysis on some of the music scores of the R&P curriculum and report the findings.

Song Level Difficulty Between 21 and 25

Table 9 lists the songs with song-level difficulty scores in the range 21 to 25 inclusive. Since this list spans many grades, it is worthwhile to observe the music scores of these songs. Through qualitative analysis, we will try and understand why their difficulty scores are similar despite belonging to different grades. As mentioned in the section 2.1.1 and especially equations 2.2 and 2.3, we already expect that high note density per time and large distances between consecutive notes would significantly affect this difficulty scoring. Although they are both important criteria in difficulty, neither rhythmic challenges or coordination difficulty or score variety beyond neighbouring notes are taken into account. These issues will become more evident in the next sections where we qualitatively examine the songs of table 9.

Gimme Some Lovin' Out of the 8 songs in grade 0, this song is ranked the second in terms of overall song difficulty. It is superseded by the song 96 Tears, which has a difficulty score of 38.14. However, there are 2 more songs with similar difficulty scores (Blue Monday and Three Little Birds), meaning that Gimme Some Lovin' is not an outlier in terms of score difficulty within this grade.

The tempo of Gimme Some Lovin' is indicated as 130 bpm, and is in 4/4 time. The song is quite repetitive in terms of rhythm and pitches, but since the probabilistic difficulty measure doesn't penalize repetitions beyond a 2 note locality, this does not result in a lower overall score. Figure 22 shows 3 measure examples through which the song is mainly comprised. These measures are either repeated exactly or with some pitch shifts.

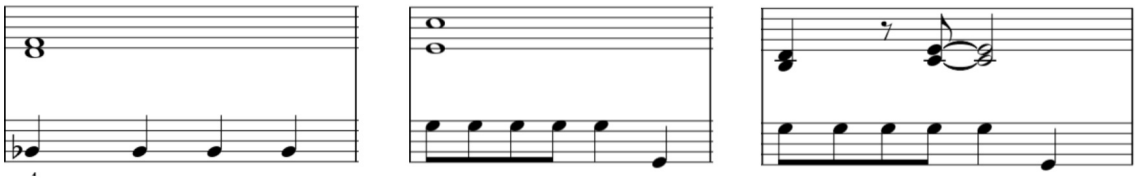


Figure 22: Example Measures in the song Gimme Some Lovin' of Grade 0 in the R&P 2018-2020 Curriculum

Comparing this to the song Something to Talk About, which has the lowest overall difficulty score (4.6162), we can realize the following:

- There are many measures where the right and left hands are not playing simultaneously, or measures where the left hand is only playing a single whole note. This causes a very low left hand difficulty score, and takes down the overall song difficulty significantly.
- The song has more diversity than Gimme Some Lovin', but since this diversity does not get reflected in the overall difficulty score, it is slightly undervalued.
- The song has a slower tempo than Gimme Some Lovin', reducing the per second difficulty of the song.

The MIDI scores of Blue Monday and Three Little Birds each have elements that would score highly using the probabilistic difficulty measure; Both have the same bpm mark, but the former is a bit faster due to the prominence of quavers throughout the song. The latter still has a high note density despite being slower, since it consistently has 3 note chords in the right hand, which would contribute to a high note density, and therefore a relatively high score of probabilistic difficulty.

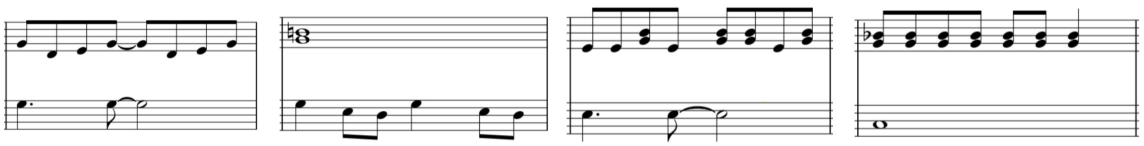
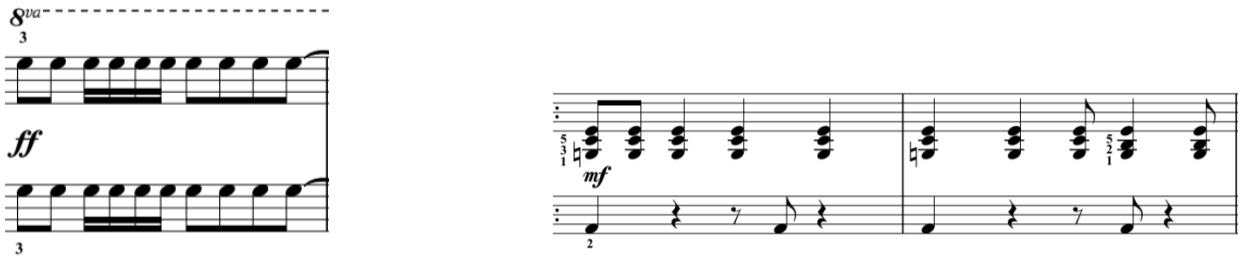


Figure 23: Example measures from the song 96 Tears of Grade 0 in the R&P 2018-2020 curriculum

Is there a notable discrepancy between the actual song difficulties in Grade 0? The answer is, yes, but not to the extent indicated by the differences in their difficulty scores. There are songs that are more challenging than others, and indeed 96 tears is arguably the most difficult song in the grade (See figure 23). But as mentioned before, the difficulty scores favor note density per time, and distances between neighbouring notes, while elements like articulations, diversity within a song, and rhythmic difficulties are not highly influential. This is reflected in the calculated difficulty scores. In the following sections, we will only go through the songs in table 9 that contribute to new observations about the probabilistic difficulty measure.

Knock on Wood The difficulty score given for this song is slightly undervalued because 2 reasons: First, there are major hand shifts that need to be executed with precision, since there two parts played at higher octaves as shown in 24a. However, this would give a large difficulty value at the point of transition only, so calculating a difficulty average for the entire song would certainly bias against these kinds of hand displacements in favor of potentially smaller but more frequent displacements. Moreover, although the pitch variety is not very high, there are subtle rhythmic changes between consecutive measures (similar to the example of figure 24b) that require attention and memory to play correctly. These would not be captured at all in the difficulty scoring.

Easy & House of The Rising Sun Both of these songs certainly undervalued, especially in comparison to the other songs within their grade. For easy in particular, the issue is that the tempo is quite slow, which would reduce the difficulty per second. The song has a high dynamic range, articulation variety, which need to be performed convincingly. These elements are not captured by this difficulty measure. House of



(a) Hand Shift Example

(b) Rhythmic Subtle Variations

Figure 24: Measures from the song knock on wood in grade 3

the rising sun is even more difficult than easy, and what brings its score down is, similarly, the fact that dynamic range is not captured by probabilistic difficulty. So is is the diversity in styles across the song. Lastly, the song is quite lengthy, so despite the existence of very fast and dexterous parts towards the end of the song, the overall averaging would bring the score down. There are many grace notes in the song, and since those would count as very small note transitions, they will not count for the difficulty they deserve.

7.1.2 Classical Curriculum

Figure 25 same plot is generated for the classical curriculum, and it is clear that when applied on the classical curriculum, the probabilistic difficulty does a worse job at characterizing difficulty according to grade than in the R&P case. Following a similar approach to the above, we will try to give brief hints as for why this is the case. Apparently, 26 songs out of 196 fall within a difficulty range between 18 and 23, and within those 26 songs there are 4 songs in grade 0 and one song in grade 8. This is too big of a span. However, in addition to being aware of the fact that perhaps better than overall song averages we should base the difficulty characterization based on more granular probabilistic difficulty results, we will review some of the songs in grade 0, 5, and 6 which have values in that difficulty range.

The four songs shown in figures 26a, 26b, 27a, and 27b are examples of songs that are certainly undervalued using the probabilistic difficulty measure. In Mister Trumpet Man (figure 26a the left hand is relatively easy, which is evident through the difficulty scores for each hand sepeartely (RH: 15.63430, LH: 2.61914). But overall, since the

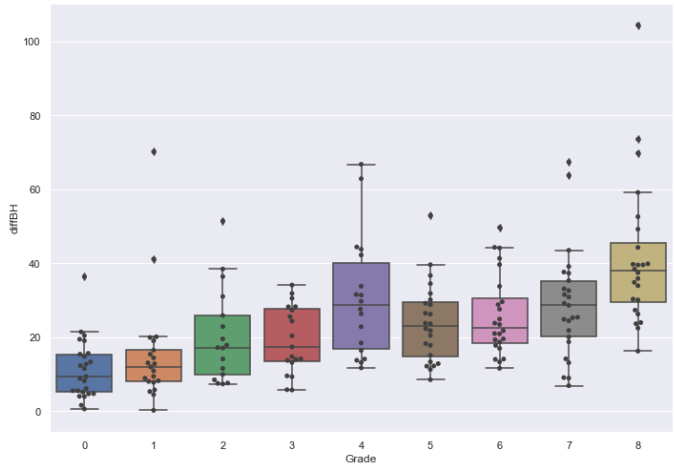
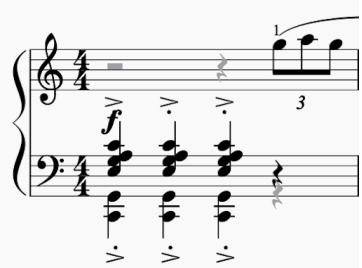


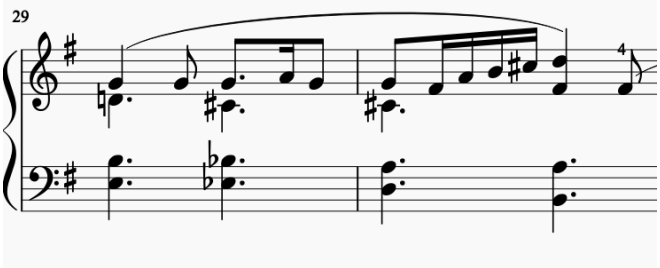
Figure 25: Probabilistic Score Difficulty for songs in the classical curriculum based on the second order fingering model described in [2]. It is a combination of a box plot and a swarm plot. A black dot represents the overall difficulty score of one song

main difficulty is in dynamics, accents, pedals, which are not taken into account by the model. Also, harmonically there are many alterations, but they are all in the form of chromatisms and bluesy notes, which this difficulty model skips because difficulties of neighbouring notes would be the lowest. The case for the song Dreamy is quite similar, although it is more difficult than mister trumpet man and the song never. Certainly the difficulty result for dreamy is undervalued. One point of complexity in it is the coordination between the right and left hand, but this is not taken into account because in general this algorithm works by computing a difficulty score for each time offset and averaging the total. Moreover, there are many triplets are a difficult point and rhythmic intricacies that are not captured, and many accidental alterations but in the context of neighbouring of closeby notes. Also, the note density is low since the right hand is predominantly monophonic, or with 2 notes (the measure shown as an example is one of the few exceptions). Another song that is highly undervalued for similar reasons is sentimental, shown in figure 27b. One important observation is that these songs mostly have similar stylistic traits, which makes it sensible that they suffer from the same pitfalls.

Finally, we show an example of a very overvalued song (Russian Dance in grade 0), and a song that seems to be a bit easy for the grade it belongs to (Valse Lente in grade 6), therefore the relatively low probabilistic difficulty score given to it was

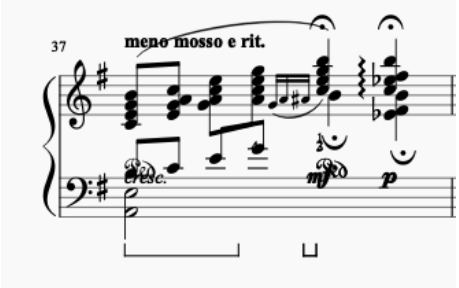


(a) Mister Trumpet Man (Grade 5)

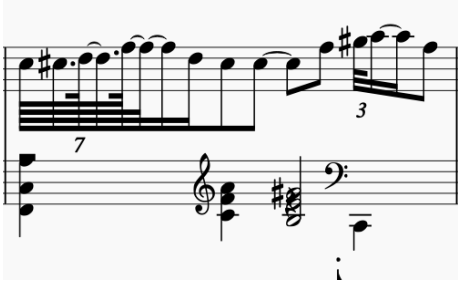


(b) Never (Grade 4)

Figure 26: Example measures from undervalued songs in the classical piano curriculum according to the probabilistic difficulty measure



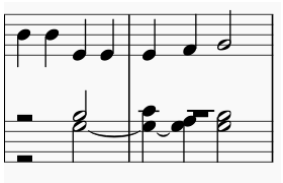
(a) Dreamy (Grade 6)



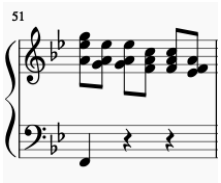
(b) Sentimental (Grade 7)

Figure 27: Example measures from undervalued songs in the classical piano curriculum according to the probabilistic difficulty measure

actually a reflection of reality. Russian Dance (28b). It is overvalued because its tempo is not so slow, and it's very consistent such that in each time window there are at least 2 notes, so this results in no sparsity in the average difficulty calculation. Also, It's extremely short, but since the average normalizes over the number of time windows considered, this doesn't change the overall difficulty score. In Valse Lente (figure 28b) but the low score is due to: challenging chordal changes happening within very short pitch distances as shown in the slide, and also because there are several parts in the score where there it's either left hand only or right hand only.



(a) Russian Dance (Grade 0)



(b) Valse Lente (Grade 6)

Figure 28: More examples from the classical piano curriculum to demonstrate the effectiveness of probabilistic difficulty

Chapter 8

Discussion & Conclusions

In this chapter, a short discussion is provided to sum up the limitations of the experiments conducted. This followed by a set of proposals for future work, and final concluding remarks.

8.1 Discussion

8.1.1 Limitations of the Feature Sets

Although it is difficult to make any definitive conclusions about the effectiveness of the feature set through the dimensionality reduction plots and the correlation plots alone, but one thing worth noting is that in all the plots for R&P, and in the PCA ones specifically, the overlaps happen between 2 or 3 neighbouring grades at most, resulting in plots that show a gradual change of color from the darker (representing the easier grades) to the lighter (representing the more difficult grades), but without the ability to draw any concrete boundaries between points of different color. The results for the classical curriculum were consistently less satisfactory, and while this could make sense in the case of the new feature set because it is entirely built on commentaries from the TCL Rock & Pop Keyboard curriculum, it was surprising for the baseline set which was supposedly developed for classical piano.

Thinking retrospectively, after suggesting features and observing their results, it is

important to note that perhaps low correlations are not necessarily an indication of a feature that is useless, but it is rather an indication that the features should be used/combined differently. For example, when we introduce a feature like articulation entropy as a way to capture the diversity of articulations and therefore to capture a form of difficulty, unless diverse articulations are present in the majority of songs in an increasing sense throughout the grades, this feature would ultimately have a low correlation with the grade label. This is certainly applicable to all features, and by taking a closer look at the features that were more significant using correlations, such as pitch entropy, we can see why it would have a high correlation since pitch variety and more complicated key signatures certainly do increase monotonically across the grade.

This, in addition to some of the features not capturing what they were intended to, such as distinct stroke rate as explained in chapter 4 (which means that coordination between hands is still unrepresented in the feature set), are important reasons for why the results were less than satisfactory. Another feature that warrants a debate about its definition is the offbeat ratio. Although we assume that the presence of a lot of offbeat notes is a measure of difficulty, it is not clear which is more difficult perceptually: regular appearance of offbeat notes in a score, or occasional appearance of offbeat notes in unexpected areas. This feature, which is a ratio between the number of notes played off-beat over the number of notes played on-beat, would return a higher number for the case of regular offbeats as opposed to occasional offbeats. So, offbeats and syncopation are parameters that need a closer examination to fully understand their difficulties and accordingly design more robust features to capture them.

Moreover, pros and cons of the song-level comparisons, due to the average over a full song often being a bad representation, and one feature where this was empirically demonstrated was the shortest rest duration feature, which showed a much higher correlation than the average rest duration feature. Lastly, the number of dimensions we have been considering (minimum 15 in baseline set, and maximum 35 in the total feature set) are too much with respect to the data that we have available, which is

a factor that must have affected the quality of our results for the feature extraction experiments.

Perhaps approaches should not necessarily be either or. It could be a good idea to examine which of the features are better represented through the probabilistic difficulty measure, and the difficulty scores as compound features representing these parameters in more diverse feature set that includes the other elements that can reliably be captured by explicit features.

8.1.2 Open Ended Feature Suggestions

When analyzing the TCL additional material for R&P, several important concepts were highlighted that we did not address in the new feature set. Below, we provide some open ended suggestions for potential features to be explored as future work.

Hand Position Features As covered in the Range parameter of the TCL criteria, there is emphasis on hand positions (standard vs non standard) and finger extensions. In the more introductory grades, the need to learn thumb crossings is made very explicit, since this one of the very first things students must learn to execute smoothly. To implement features for hand positions, we must first annotate the music scores with fingering predictions (which is a feasible step), and define the hand motions that would correspond to each of the features. Taking thumb crossings as an example, it can occur as thumb over or thumb under. However to implement this, a set of hand positions or finger extensions would need to be defined with the help of music teachers.

Playing Styles This is one parameter that is commonly seen, where for example: 'ability to change between playing styles' is commonly one of the challenges of executing a particular piece. Formally, what would constitute this playing style? and is it possible to infer from the score whether 2 consecutive measures would be executed using different styles? Is this dependent on rhythmic texture, part writing, or both? It could be difficult to answer this question from an absolute sense. For this feature, perhaps we can utilize the approach of [9] which they employed to capture

the rhythmic texture from polyphonic scores, as explained in section 2.2.2. Perhaps a playing style would be a combination of rhythmic pattern, dynamics and accents, fingers moving up/down, phrase crossing left hand and right hand (should be played smoothly)

8.1.3 Future Work

Within the line of R&P, it would be interesting to divide the songs into genres, and then understanding how the feature values vary across grades for specific genres. However, with only one curriculum from the R&P Syllabus, this was not possible because some genres exist almost exclusively in earlier grades (i.e synth-pop) and others in higher grades (i.e rock opera). This will be revisited when more curricula are available for research. Moreover, it is an observation that for music scores in higher grades especially, much more textual descriptions exist in the music score, so perhaps a text traversal of the annotations in the musicxml file to identify indications of style and contextualizing the difficulty accordingly would be useful.

Outside the line of R&P, an urgent of improvement is to implement the same analytic approach with the commentaries from the TCL classical piano songbooks and try to find a set of features more representative of classical, and in the process understand why the performance on classical was consistently worse than R&P.

Another line of future work which is related to the probabilistic difficulty approach. First of all, it would be interesting to try and create a piano-score model that encompasses rhythmic information as well, including information on coordination difficulties between the left and right hand. Perhaps the model can be built statistically based on actual performance data or student practice data. Creating annotated datasets with student errors of while playing songs from the TCL curricula can be very useful because, as explained in [1], performance errors could provide objective information to help us understand music score difficulty. Moreover, another idea is to try and understand the expressive range allowable by a piece, which could be approached by analyzing multiple performances of the same piece and observing the variability in performance actions (as termed by Giraldo and Ramirez in [7]). This

could be a way to objectively annotate score regions with high expressive variety.

Lastly, all if this work can be extended to different instruments.

8.2 Conclusions

This work was an attempt to contribute towards creating music education technology by tackling a research topic that has not received significant attention over the past years, which is music score difficulty characterization. The approaches in relevant literature fall under two categories: one which is based on feature extraction, and the other which is based on probabilistic difficulty, both reviewed in chapter 2. Both approaches were applied on music scores provided by Trinity College London (TCL), which belonged to 2 stylistically distinct curricula: Rock and Pop Keyboards, and Classical solo piano, as explained in chapter 3. In addition to applying each of these approaches on TCL data, which we do in chapters 4 and 6 for the feature extraction approach, and in chapter 7 for the probabilistic difficulty approach, of the main contributions of this work is the qualitative analysis in our attempts to evaluate either approach. Moreover, despite the proposed feature set for the feature extraction approach yielding mild improvements, the reports thorough analysis of the TCL educational material allowed us to adopt a more organized framework for viewing score difficulty parameters, which is shown in chapter 5.

List of Figures

1 Results of evaluating probabilistic difficulty. No-information, Gaussian, and Fingering models each refer to a particular pitch sequence model. The red arrows highlight the difficulty value in each model after which the error rate starts to increase. Source [1] 11

2 Basic pipeline of the steps involved in feature extraction 22

3 Results of both PCA (left) and T-SNE (right) dimensionality reduction on the R&P Keys curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs 28

4 Component Coefficients of the baseline features on the the R&P Keys curriculum for the first 3 Principal Components 29

5 Left: PCA on Baseline Set. Right: T-SNE on Baseline Set 30

6 Results of both PCA (left) and T-SNE (right) on the TCL Classical Piano curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs 30

7 Component Coefficients of the baseline features on the first 3 principal components for the Classical Curricula 31

8 Results of Spearman Correlation between the grade label and the baseline features for the R&P Curriculum 32

9 Component Coefficients of the baseline features on the first 3 principal components for the Classical Curricula 34

10 Snippets from the TCL R&P Songbooks through which we demonstrate why the distinct stroke rate feature is not an effective measure of LH and RH corrdination 35

11 Results of both PCA (left) and T-SNE (right) dimensionality reduction on the new feature set calculated from the R&P Keys curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs 52

12 52

13 Results of both PCA (left) and T-SNE (right) dimensionality reduction on the total feature set (baseline + new features) set calculated from the R&P Keys curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs 53

14 53

15 Results of both PCA (left) and T-SNE (right) dimensionality reduction on the new feature set calculated from the Classical Piano curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs 54

16 55

17 Results of both PCA (left) and T-SNE (right) dimensionality reduction on the new feature set calculated from the Classical Piano curriculum. Each point represents a song, and the colors reflect the TCL grade to which a song belongs 55

18 56

19 Feature Correlations with Grade for the new feature set for R&P Keys 56

20 Feature Correlations with Grade for the new feature set for Classical Piano 57

21 Probabilistic Score Difficulty for songs in the R&P curriculum based on the second order fingering model described in [2]. It is a combination of a box plot and a swarm plot. A black dot represents the overall difficulty score of one song 59

22 Example Measures in the song Gimme Some Lovin' of Grade 0 in the R&P 2018-2020 Curriculum 61

23 Example measures from the song 96 Tears of Grade 0 in the R&P 2018-2020 curriculum 62

24 Measures from the song knock on wood in grade 3 63

25 Probabilistic Score Difficulty for songs in the classical curriculum based on the second order fingering model described in [2]. It is a combination of a box plot and a swarm plot. A black dot represents the overall difficulty score of one song 64

26 Example measures from undervalued songs in the classical piano curriculum according to the probabilistic difficulty measure 65

27 Example measures from undervalued songs in the classical piano curriculum according to the probabilistic difficulty measure 65

28 More examples from the classical piano curriculum to demonstrate the effectiveness of probabilistic difficulty 65

29 Heatmap of the correlation matrix for the baseline feature set calculated on the R&P Syllabus 78

30 Heatmap of the correlation matrix for the baseline feature set calculated on the Classical Syllabus 79

31 Heatmap of the correlation matrix for the total feature set calculated on the R&P Syllabus 80

32 Heatmap of the correlation matrix for the baseline feature set calculated on the Classical Syllabus 81

33 82

List of Tables

1 Linear regression weights for each of the baseline features 33

2 Descriptions of the Range parameter across the different grades for the R&P Keys own choice parameters. Source [11] 40

3 Descriptions of the Part Writing parameter across the different grades for the R&P Keys own choice parameters. Source [11] 41

4 Descriptions of the Melodic Writing parameter across the different grades for the R&P Keys own choice parameters. Source [11] 41

5 Notable parameter values observed from R&P Grade 0, 1, and 2 songbook information and the music scores 44

6 Notable parameter values observed from R&P Grades 3, 4, and 5 songbook information and the music scores 45

7 Proposed features by parameter. Those suffixed with * are not implemented 48

8 Results of Correlation with grade for the new feature set on R&P . . 56

9 Distribution of Songs with Song Level Difficulty Scores Ranging from 21 to 25 in the TCL 2018-2020 R&P Curriculum 60

Bibliography

[1] Nakamura, E. & Yoshii, K. Statistical piano reduction controlling performance difficulty. *APSIPA Transactions on Signal and Information Processing* **7**, 1–12 (2018).

[2] Nakamura, E., Saito, Y. & Yoshii, K. Statistical learning and estimation of piano fingering. In *Information Sciences*, vol. 517, 68–85 (2020).

[3] Sébastien, V., Ralambondrainy, H., Sébastien, O. & Conruyt, N. Score analyzer: Automatically determining scores difficulty level for instrumental e-learning. In *ISMIR* (2012).

[4] Chiu, S. & Chen, M. A study on difficulty level recognition of piano sheet music. In *2012 IEEE International Symposium on Multimedia*, 17–23 (2012).

[5] Song, Y.-E. & Lee, Y. K. A method for measuring the difficulty of music scores. *Journal of The Korea Society of Computer and Information* **21**, 39–46 (2016).

[6] Parncutt, R., Sloboda, J. A., Clarke, E., Raekallio, M. & Desain, P. An ergonomic model of keyboard fingering for melodic fragments. *Music Perception: An Interdisciplinary Journal* **14**, 341–382 (1997).

[7] Giraldo, S. & Ramírez, R. A machine learning approach to ornamentation modeling and synthesis in jazz guitar. *Journal of Mathematics and Music* **10**, 107–126 (2016).

[8] Bantula, H., Giraldo, S. & Ramirez, R. Jazz ensemble expressive performance modeling. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 674–680 (2016).

[9] Levé, F. *et al.* Rhythm extraction from polyphonic symbolic music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (2011).

[10] Coca, A. E. & Zhao, L. Musical rhythmic pattern extraction using relevance of communities in networks. *Information Sciences* **329**, 819–848 (2016).

[11] London, T. C. *Keyboard Syllabus: Qualification Specifications for Graded exams from 2018* (Trinity College London, 2017).

[12] Cuthbert, M. S. & Ariza, C. Music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 637–642 (2010).

[13] F.R.S., K. P. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901).

[14] van der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

[15] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

Appendix A

Additional Figures

This section includes additional plots that are interesting but kept outside of their respective chapters so as not to disrupt the reading flow. These include: the full correlation plots for the baseline feature set then for the total feature set, for both classical and R&P, and running plots on feature subsets.

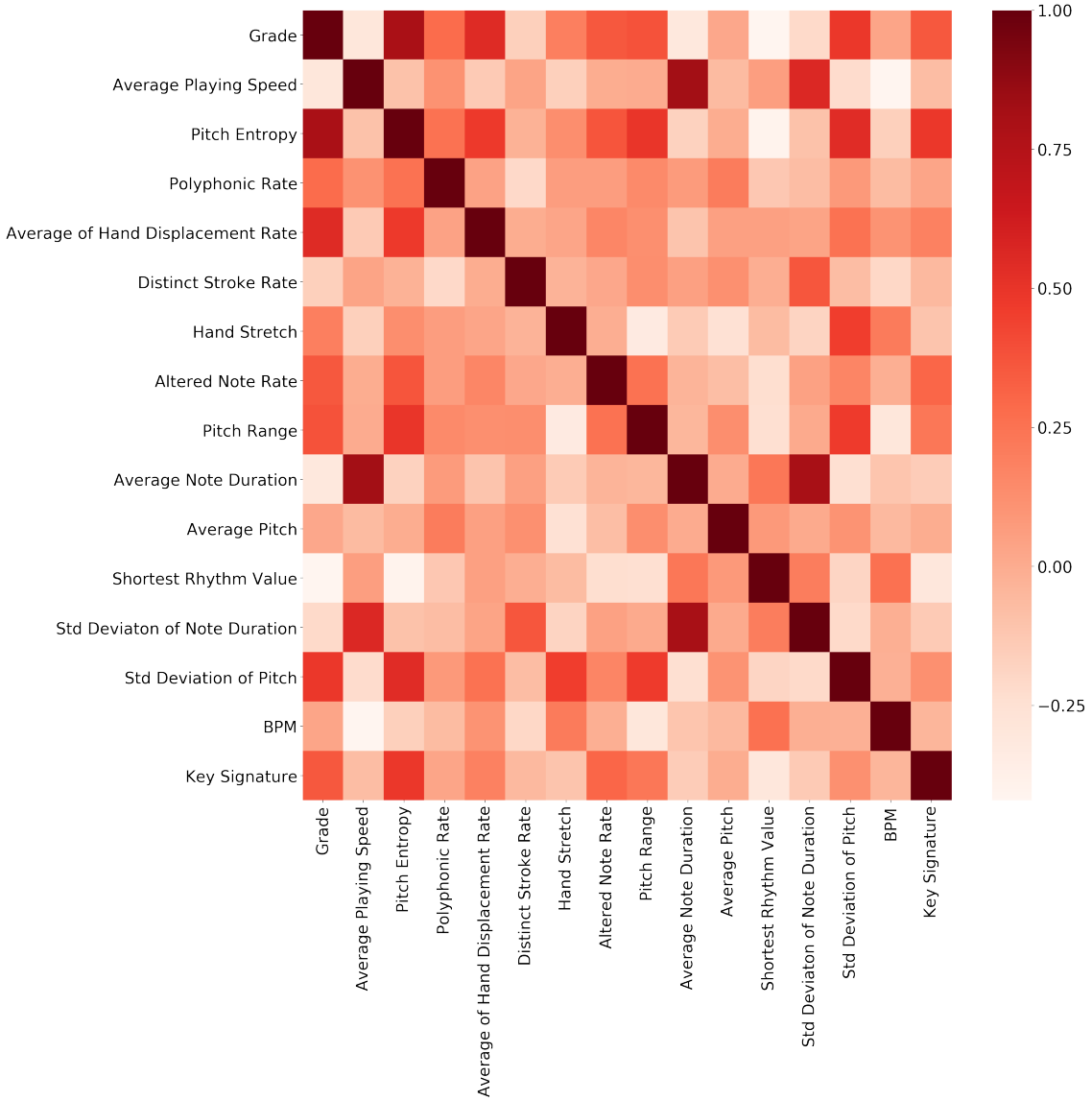


Figure 29: Heatmap of the correlation matrix for the baseline feature set calculated on the R&P Syllabus

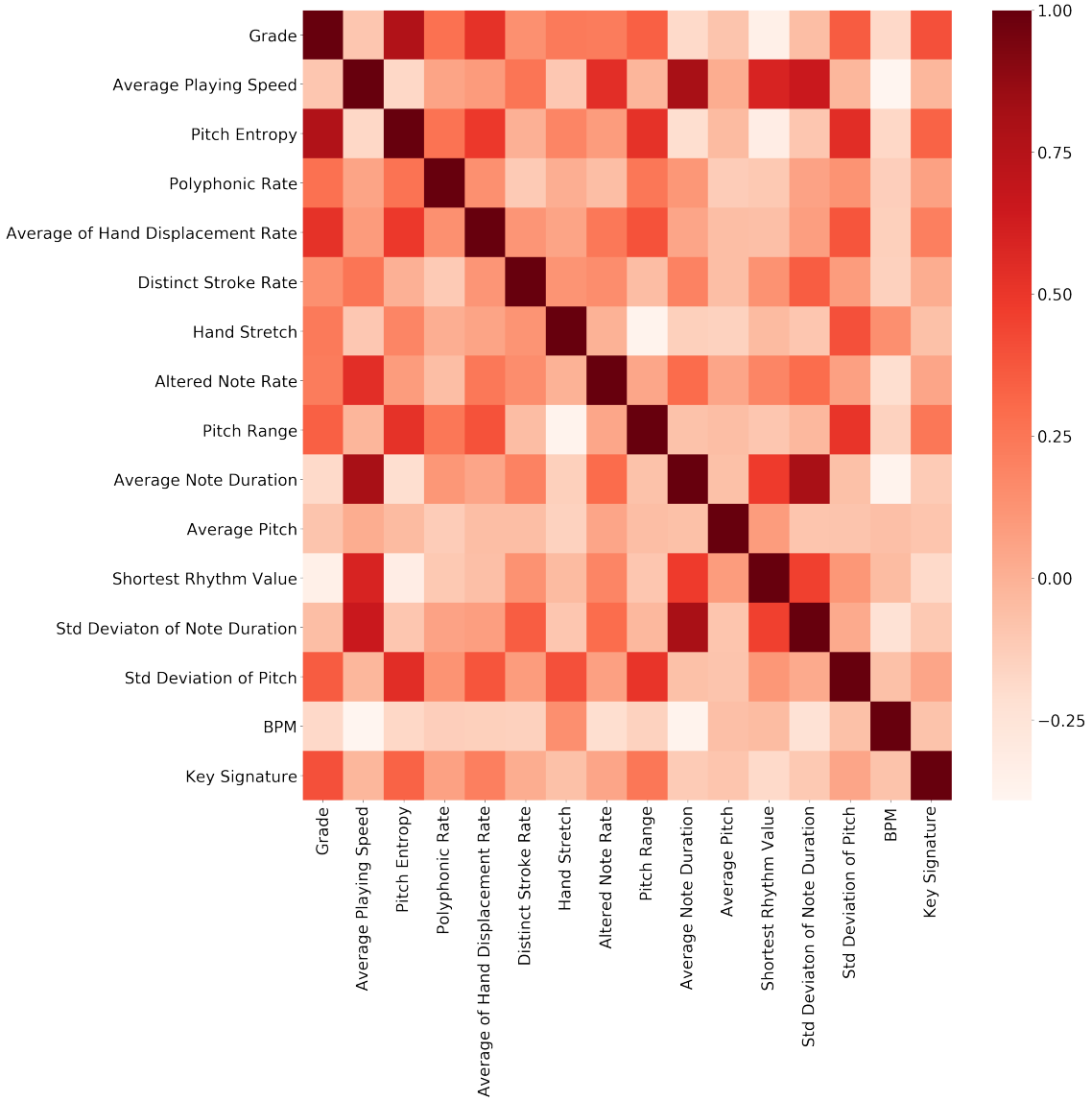


Figure 30: Heatmap of the correlation matrix for the baseline feature set calculated on the Classical Syllabus

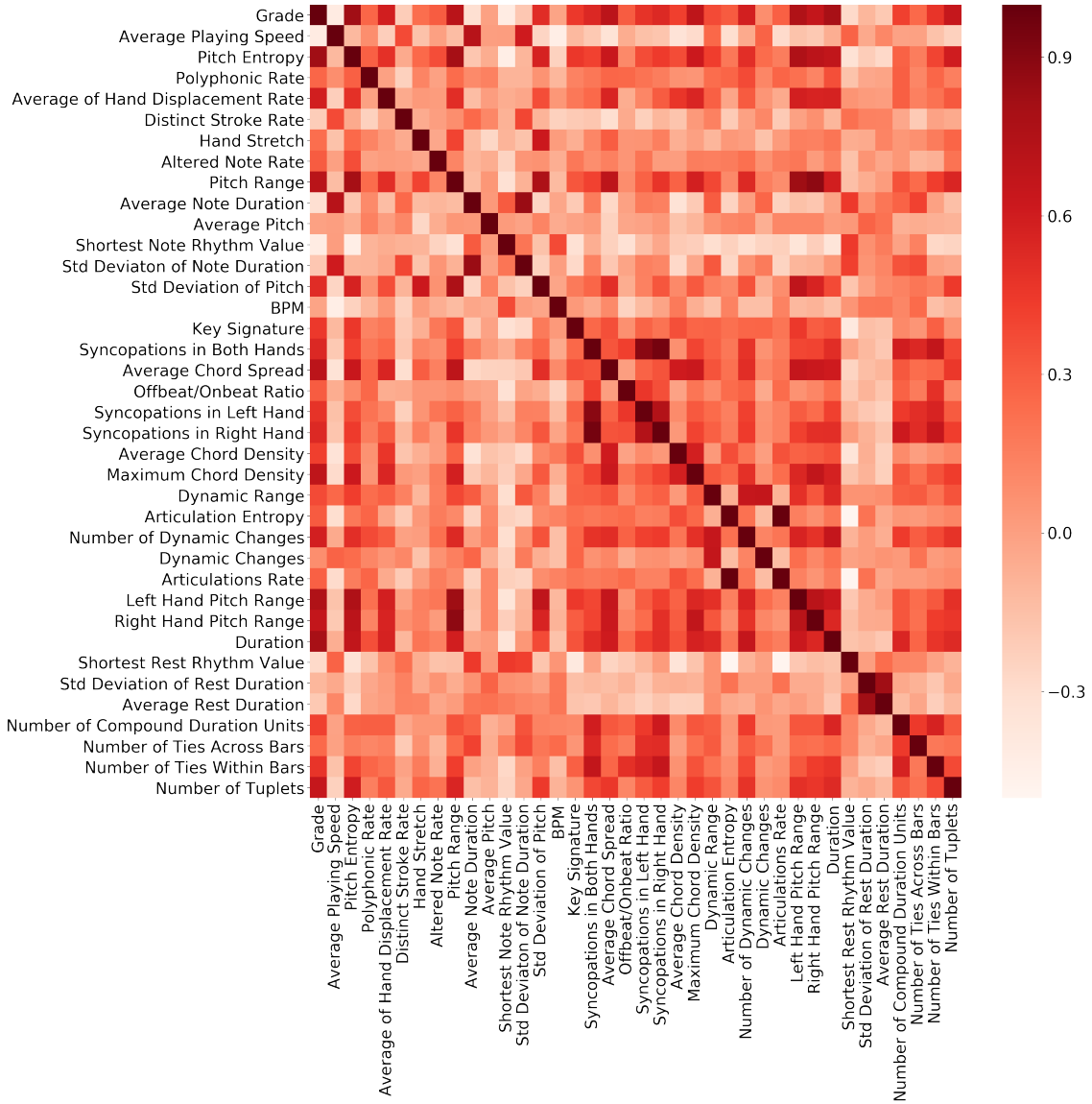


Figure 31: Heatmap of the correlation matrix for the total feature set calculated on the R&P Syllabus

Figure 32: Heatmap of the correlation matrix for the baseline feature set calculated on the Classical Syllabus

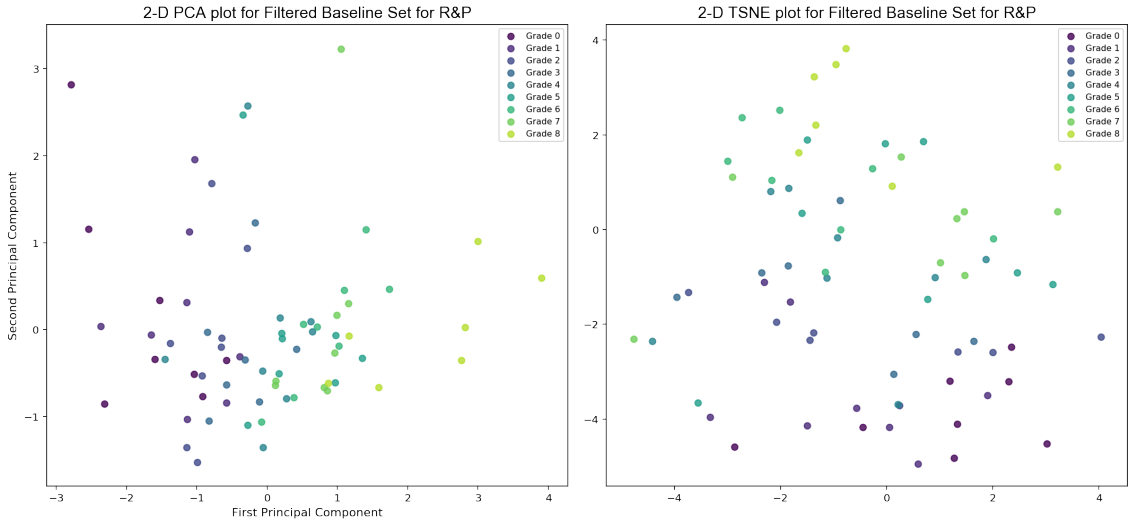

2-D PCA plot for Filtered Baseline Set for R&P
2-D TSNE plot for Filtered Baseline Set for R&P
Grade 0
Grade 1
Grade 2
Grade 3
Grade 4
Grade 5
Grade 6
Grade 7
Grade 8
Second Principal Component
First Principal Component

Figure 33