A First-Order Theory of Film Scores for Generation from Lightweight Specifications

Halley Young
Department of Computer Science
University of Pennsylvania
Pennsylvania, PA 19104, USA
halleyy@seas.upenn.edu

Abstract

This paper proposes a formal theory of the way film scores operate for the purpose of enabling semi-automatic generation. Among the contributions are a formalization of the entire generation process as a bit-vector-array satisfiability problem, an approach to music generation not taken in many previous papers. The paper also formalizes the idea of "thematic" and "stylistic" time-dependent variables and their inherited constraints in specification-driven generation. In order to make the result more coherent, the paper formalizes a regular-expression-like grammar of melodic contour. Synthesizing all of these contributions, the result is a program which can take a lightweight specification of the relevant information in each scene of a film, and produce a coherent and appropriate score to accompany it.

Introduction

The film industry turns out over $200 billion of films every year, and a substantial portion of that is spent creating appealing film scores ((?)). According to studies by Stuart Fischoff, himself both a film writer and scholar of media psychology, music scores to films account for much of our understanding of the emotional impact of film scenes as well as characterizations of different characters, locations, and events ((?)). However, there has not been substantial research on formalizing the way that film scores produce these effects. While there has been some research on producing film scores semi-automatically, these approaches are either statistical (and suffer the same limitations as most deep-learning based music, such as lack of memorable material or global structure), or don't include a background theory of film composition, and thus require extensive manual specification by the composer ((?) (?)). This study proposes a formal theory of film music, written in a decidable fragment of first-order logic. The theory allows for the generation of appropriate film music from lightweight annotations.

Related Literature

There is a wealth of literature available on generative music models. While machine-learning-based models such as music transformers have dominated attention recently (?), there has always been significant interest in logic- and constraint-based approaches, using languages such as Prolog (?) or Oz (?). However, to our knowledge this is the first paper to generate music using SMT solvers.

In parallel, there is a body of literature which is similar not in method, but in the aim to create a system which can respond to the user's desired states. This includes work which is specifically tailored to evoking a particular emotion (?) (?) (?), or work which allows users to control more technical details (?). In terms of usage contexts, these methods have been applied to creating music to accompanyvisual art (?) and videogames (?).

Algorithm Input and Output

We propose an algorithm for generating film scores from lightweight annotations. The input to this algorithm is a specification. A specification contains an arbitrary number of lines, each of which contains a list of variables, a specified duration time in seconds, and, optionally, a description of the scene (used only for documentation and not by the algorithm). Variables can either be stylistic variables (defined in the theory of film, as per the appendix), or thematic variables (only defined in the universe of the specific film). For instance, in the following specification of a short manufactured example, "Jose" (a character) is a thematic variable, while "flamenco" (a well-defined style of music found in Spain) is a stylistic variable, as is "suspense" and "happy":

```
Jose grew up in Spain. {Jose, Flamenco} 8

Then he moved to New Orleans. {Jose, Zydeco} 10

It was there that he met Sally. {Sally, Zydeco} 4

Sally was the most beautiful person
he'd ever met. {Sally, romantic} 6
```

```
They got married and moved back to Spain,
where they had a child.
{Jose, Sally, child, Flamenco} 6

But then an alien invasion came,
and infected the child.
{child, horror, alien, suspense} 8

In the end, Jose and Sally ended up having
to go to space to beg the alien king
to save their child.
{Jose, Sally, child, alien, suspense} 8

The alien king was touched by their plea,
and their child was saved.
{suspense, happy, alien, child} 6

They all lived happily ever after.
{sally, jose, child, happy} 6
```

The output of the algorithm is music which conforms to this specification, in that, for a sequence of variable lists $S_1 \dots S_n$ and durations $a_1 \dots a_n$, the variables in S_i are present in the music at timestamp $\sum_{j=0}^{i-1} a_j$ to $\sum_{j=0}^{i} a_j$. (A variable being "present" is defined below). Furthermore, the generated music is "musically coherent" (also defined below).

Building on SMT solvers - the Set-Theoretic Universe of Film Score Theory

Generating the basic building blocks of a film score involves determining a set of values of "mid-level" musical variables at every moment in time. Some of these properties are completely independent (rhythmic density and harmonic progression), while others have mutual constraints (the number of rhythmic values must be the same as the number of pitch values). Some properties depend on the duration of time in which they are being used (for example, it's not realistic to have a full Andalusian progression in 2 seconds or less).

Musical Types and Values

The mid-level universe consists of a set $T = t_0 \dots t_n$ of types of mid-level variable, such as "harmonic rhythm" (the rate at which the chords change), "rhythmic density" (the average duration of a single note in the melody), or "has unpitched percussion". These types are associated with their range of values, which can be boolean (e.g. whether or not there is an ostinato), a bounded integer (e.g. the degree of tension), a finite set (e.g. the list of possible chord progressions), or a list of bounded integers of bounded size (e.g. the melodic contour). In principle this can be extended to bounded floating point numbers, but for simplicity bitvectors were used. Boolean variables in the SMT logic are modeled simply as Boolean SMT variables, bounded integers are modeled as bit-vectors, sets are modeled as 1-hot bit-vector variables, and lists of integers are modeled as a tuple of a fixed-size arrays of bit-vectors of some maximal size k_{max}, and a value $k_{actual} < k_{max}$ such that all entries with indices $> k_{actual}$ are ignored.

Time-Span Sets and Thematic Sets

Consider a specification with m scenes and n total thematic variables in all of the scenes (n can easily be obtained by counting the number of unique elements of each S_i that are not designated as "stylistic" by the theory). Under these assumptions, there will be a set $M_0 \dots M_m$ of sets of mid-level properties corresponding to each scene. These sets will be total in the sense that for every musical type t in T, one possible value of t will be in M_i. There will also be a set $N_0 \dots N_n$ of mid-level properties corresponding to each theme. However, these sets will not be total - some sets may, for instance, include one element of the set of possible chord progressions but no possible value of rhythmic density, while another may contain a possible valuation of rhythmic density but not of chord progressions. This is because, as film music scholar Andrew Powell acknowledges, a leitmotif (the musical elements that together make up a "theme", or a memorable gestalt which can appear in various versions but still be recognizable), can be one of a variety of musical markers which "which serve[s] to distinguish a character, idea, or symbol", rather than one or several necessary and sufficient musical characteristics ((**?**)).

Axioms of the Formal Theory

Axioms of the formal theory relate the specification of a film score to constraints on its attributes. These constraints include formal definitions of the high-level stylistic elements that are described in annotations, rules regarding the existence of a well-defined leitmotif when appropriate, some basic rules which establish musical coherence, and ontological claims which relate to the logically necessary relationship between various variables.

Stylistic

Stylistic axioms tie the stylistic definitions assumed by annotators to mid-level variables. Unfortunately, there is a shortage of academic papers on the specific musical attributes associated with more broad words encompassing ideas such as genre or emotion. Where possible, I took definitions from academic sources, including ((**?**) (**?**)). However, in some instance it was necessary to simply survey the non-academic resources available ((**?**)).

Below are a few of the definitions used:

1. **"Flamenco"** style implies at least three of: the use a flamenco percussive pattern, the use of castanets and clapping as percussive instruments, the use of

guitar, the use of phrygian mode, and the use of an Andalusian chord progression.

2. **“Horror”** implies at least three of: existence of a repeating ostinato, use of dissonance, use of a chromatic chord progression or chord transformation, use of low register.
3. **“Jazz”** implies: the use of a stereotypically jazz percussive pattern, use of dominant seventh chords, and use of a high level of syncopation.
4. **“Happy”** implies: the use of a major scale, and either the use of a fast tempo or the use of a high register.

Note that in practice, these definitions do not ensure the desired feeling, and indeed in the examples it can be difficult to discern exactly what style is being evoked. However, they can be thought of as probably necessary conditions, such that $P(style(m) = x|m \models \Phi)$, where x is a style and Φ are its related constraints, is much higher than $P(style(m) = x|m \not\models \Phi)$. More research is necessary to determine other variables which would increase $P(style(m) = x)$ for various styles.

Ontological

Ontological constraints include constraints which are inherent to the meaning of the different mid-level variables. For instance, in order for a scene to have “a violin playing the accompaniment”, it is necessary both for the scene to be accompanied and for the scene to contain a violin; in order for a scene to have a “Andalusian cadence” (a specific pattern defined under the rules of 12-tone tuning and undefined for tunings where $n \neq 12$), it is necessary for the tuning to be 12-tone. To be precise, ontological constraints occur when there exist two variable assignments, $p\hat{h}i$ and $\hat{\theta}$, such that there is no possible music satisfying the condition $p\hat{h}i \wedge \not{t}\hat{h}eta$. In an end-to-end system, where the entire music generation could be described as a single SMT instance, than if $\hat{\phi}$ were constrained to be true than the system necessarily would return a result such that $\hat{\theta}$; however, due to tractability issues discussed elsewhere, it was necessary to decouple the generation of “mid-level” and “low-level” variables. Therefore, the system has no a-priori knowledge that the variable describing “Andalusian cadence” is dependent on the variable describing “12-tone.”

Leitmotivic/Thematic

The second type of constraint concerns making sure that leitmotifs are associated with the correct theme. In a major simplification, we assume that the specification can be cleanly separated into stylistic and thematic variables, so, while “the alien” might only occur in sci-fi scenes, it is not itself assumed a-priori to have different defining properties than “the cowboy” (although the cowboy may only occur in scenes which are designated as “Western”, and thus will also in effect be associated with this genre). The thematic variables impose additional constraints on the mid-level variables associated with each time span, as *a musical element can only represent a theme if it is present in every scene where the theme is included in the specification, and not present in any scene where the theme is not included in the specification.* In addition, themes must be unique, and must be noticeable. In a simplification, this is expressed as the following constraint on the thematic sets $N_0 \dots N_n$: The sets must be completely disjoint, and each must contain at least two elements. Thus, for any two given themes a and b, either the value of type t associated with a is different than the one associated with b, or a includes a value of type t while b does not.

Axioms of Contour - Creating Coherent Motivic Material

The axioms regarding melodic contour require their own section, as they are slightly more complex. As discussed above, rhythmic and melodic contours are lists of bounded size of bounded integers. A tuple of a rhythmic contour and a melodic contour form a “motivic contour,” and this tuple is one of the variables which can be assigned to a theme. Furthermore, an ordered list of motivic contours are assigned to each scene, with the number of elements roughly correlating to the duration of each scene. Thus, a scene can contain individual motivic contours corresponding to multiple themes if the scene duration is large enough.

A rhythmic or pitch contour is a list of numbers $C = c_0 \dots c_k$ such that, if $x \in C$, then $\forall 0 \geq j < x, j \in C$. The rhythmic (or pitch) contour is associated with the following constraint on the rhythmic values $r_0 \dots r_k$: $\forall i < k$, $\forall j < k$, if $c_i < c_j$ then $r_i < r_j$, if $c_i > c_j$ then $r_i > r_j$, and if $c_i = c_j$ then $r_i = r_j$. Thus, the contour restricts the relative size of the different durations without restricting absolute sizes or even size ratios. This is a standard definition among modern music theorists (**?**).

Why Motivic Contours?

Most accounts of melodic ideas involve specific pitches and rhythms. Arnold Schoenberg was perhaps the first to explicitly promote the motivic contour as a core component of a musical idea; however, uses of constant contoural structures over changing pitches have been an element of Western music at least since Bach ((**?**) (**?**)). For this algorithmic approach, it is useful to cleanly divide between contoural structure and the specific pitch and rhythmic elements so that both can be constrained and assigned values independently. For instance, a $[0, 2, 1]$ pitch contour can be associated with any type of scale or chord, thus increasing the size of the possibility space by an order of magnitude. It could be fulfilled by a pitch sequence [C4, G4, E4] (a major triad, or one particular kind of harmony), or [C4, G4, D4] (a sus4 triad, or a harmony with a different emotional valence).

Grammatical Contours

I introduce a language for describing valid grammatical contours. This language is based on prior work by cognitive and computational musicologists. It can be used to enumerate a sequence of contoural values which is more likely to sound "musical" than a contoural sequence generated by randomly choosing numbers on a given interval in $\mathcal{Z}$. Readers can subjectively compare tunes generated by the two methods by going to `https://www.seas.upenn.edu/~halleyy/random-and-verified-melodies`.

Note that this grammar assumes that the user wants the theme to be coherent and uphold the sort of contoural constraints seen in pre-20th century music. This is not always the case for avante-garde music, nor for film music in general. An extension of this work would eliminate the contoural grammar in very specific scenarios where doing so would create the appropriate effect.

Musicological Antecedents of the Contoural Grammar

The development of the contoural grammar draws on work by authors including Larson, Narmour, Ockelford, and Meredith, all of whom developed melodic theories ((**?**) (**?**) (**?**) (**?**)). Central to all of their work (whether under the label of "inertia" in Larson's physics-based theory or as "compressibility" in Meredith's computational theory) is the idea of a necessary degree of repetition and controlled variation, including several stereotyped methods of variation. Larson also introduces other operators such as "gravity", or the idea that after a leap a pitch should tend to fall down, which will be incorporated into the contoural grammar.

The Rhythmic and Melodic Contoural Languages as Interpreted Subsets of Regular Grammars

The contoural languages take the following form: a list of values (with an optional repetition exponent), references, and transformations on references, followed by a list of reference valuations. The reference valuations are lists of values, with an optional repetition exponent.

Examples of Famous Works in Regex form

The main theme of Mozart's 25th piano concerto is one of the most beloved melodies in classical music. Below is the pitch contour of the theme:

$$[0,0,0,1,1,2,2,3,1,3,5, \tag{1}$$

$$4,4,3,3,2,2,2,2,3,3,4,4, \tag{2}$$

$$5,1,3,5,3,3,2,2,1] \tag{3}$$

This can be read as the interpretation of the following regex:

$$(012(ui1)(m0)(uu1)2(i1)) \tag{4}$$

$$((0,0,0),(1,1,2,2,3),(1,3,5)) \tag{5}$$

In English, this can be interpreted as "the first pattern (0,0,0) followed by the second pattern (1,1,2,2,3) followed by the third pattern (1,3,5) followed by the second pattern transposed to start at the current value and inverted (4,4,3,3,2) followed by the first pattern transposed up two levels followed by the third pattern followed by the inverted second pattern".

Similarly, the rhythmic contour of the main theme of Smetana's Moldau has the following pattern:

$$[0,1,0,1,0,1, \tag{6}$$

$$2,2,2,3,1,2, \tag{7}$$

$$1,0,1,0,1,0,3,1,3] \tag{8}$$

This can be read as the interpretation of the following regex:

$$(0(2)^3 1(r0)(a^3 1)) \tag{9}$$

$$((0,1)^3,(3,1,2)) \tag{10}$$

In English, this can be interpreted as "the first pattern (itself a tri-fold repetition of a simple pattern), followed by a tri-fold repetition of the value 2, followed by the second pattern, followed by a retrograde of the first pattern, followed by the second pattern with the third value augmented."

Necessary Constraints on Melodic Contours - Axioms of "Coherence"

According to several authorities cited above, coherent music necessarily must involve a substantial (but not an excess) of repetition, and specifically varied repetition. It is thus necessary to constrain the valuations of the regex. The following constraints were imposed:

1. If the melody is of sufficient length (> 6 seconds), each of the patterns has to either be used in its original form at least once and then used in some other form twice, or used twice in its original form.
2. If the melody is very short (¡4 seconds), only one pattern can be used, and if it is somewhat short (¡6 seconds), only two patterns can be used.

In addition, as per Larson's description of the consequences of "gravity" and "inertia", there was a constraint on what can follow a leap (contoural values x_i and x_{i+1} such that $abs(x_{i+1} - x_i) > 4$, namely that a leap has to be followed by a “step” - $abs(x_{i+1} - x_i) < 2$ - in the opposite direction.

Note that the only tested melodies were at most 15 seconds long (about 8 bars, which is typical of an antecedent-consequent style Classical theme), which significantly reduced the possible complexity of the melodies. More research is necessary in order to achieve coherence across larger time spans.

From mid-level materials to notes

To generate notes from the values (including contoural values and mid-level variables) output by the SMT solver in the first pass, further satisfiability and constrained optimization problems were constructed. First, generating a rhythm from the rhythmic contour was framed as constrained optimization: It was necessary for the contoural constraints to be recognized for all i and j, ($length(x_i) > length(x_j)$ if and only if the i^{th} contoural value was greater than the j^{th}), the total length of the rhythmic pattern was constrained to be the specified length of the scene, and an optimization was sought that maximized the sense of meter. After the rhythm was generated, melody was generated with constraints maintaining contoural values and definitions of chord progressions and scales (each pitch modulo 12 has to be either in the respective chord, in the respective scale and in between two notes less than 3 semitones apart, or in between two notes less than two semitones apart). Finally, the accompaniments were chosen to match the instrumentation, dissonance level, thickness, spacing, etc. of the other mid-level variables.

Restriction sequences and tractability

The method of determining first the mid-level variables, then rhythmic values, then pitch values, and finally accompaniment and timbral features in a series of disjoint SMT instances suggests an interesting avenue of research. In the first implementation, features were to be generated all at once in a single constrained optimization instance. However, the search space was apparently far too large for the optimization to terminate. Even the difference between separating duration and pitch generation vs. determining them together was decisive in determining feasibility (separate runs proved tractable while joint generation was not). One could understand the ordering of variables to be synthesized as a sequence of operations, each of which further restrict the search space over all possible melodies.

Empirical Results

Generation from random scores

A suite of 30 random film specifications were generated by assigning fixed probability distributions over seeing a given style over each scene’s timespan, joint probabilities over theme variables, and a fixed distribution of number of scenes. According to this analysis, **20.0%** of randomly generated film scores were satisfiable. In contrast, all of the three handcrafted synthetic film specifications and three handcrafted specifications for existing films were satisfiable. This discrepancy suggests that the distribution of styles and thematic material in real films is non-uniform.

Generation from hand-crafted film and specification

Three stories of lengths 27-62 seconds were handcrafted for the sake of this research. They were intended to be realistic, but also erred on the side of having a large amount of thematic and stylistic variety. It took an average of **189.6 seconds** for the process of generation. Each of these film specifications had a satisfying generation. The reader is free to evaluate the results at `https://www.seas.upenn.edu/~halleyy/synthetic-film-score-generation`. In particular, it is worth noting the drastic difference in quality between the example where a composer manually wrote the piece but constrained herself to use the generated mid-level variables (example 1), and where the end-to-end system was used. This suggests that the mid-level generation may be more robust than the middle-to-low-level system.

Generation from lightweight annotation of existing film

Three short film clips of length 28-62 seconds were chosen, and specifications were written for each. In practice, it took less than five minutes to create each specification, suggesting that specification creation itself is not a limiting factor. Each of these film specifications had a satisfying generation. The reader is free to evaluate these results at `https://www.seas.upenn.edu/~halleyy/real-film-score-generation`.

Future Work

One of the appealing features of this approach is the lightweight nature of the specifications - it took less than 5 minutes for the author to write up the annotations for a real-world 46-second scene, which according to several internet forums would be viewed by a professional composer as a task deserving of $375-750 ((**?**)). However, the opportunity to expand the specification could decrease the gap in expressivity between music generated automatically and music written by professional composers. For instance, in this approach there is a clear and simple distinction between stylistic and thematic variables. However, in a more expressive language, it would be possible to explicitly associate certain properties with characters as well as how those properties change over the course of the film. In addition, several film theorists have suggested that leitmotifs can be changed in a very deliberate manner through the span of the movie so as to suggest character development, a very important expressive possibility that is completely absent in this work.

Long-term structure is notoriously hard in film music, and in music in general. This approach incorporates long-term structure in that there are recurring leitmotifs and in that each individual scene is scored using a principled musicological approach, but the sense of continuity between scenes is still significantly less than one would typically find in most music. An improvement on this approach would be to include constraints on the distances between subsequent scenes, although the nature of these constraints are not obvious.

Due to the nature of SMT solvers, the valuations of each mid-level variable are not independent across executions of this algorithm (even on completely different scripts). This is a deficit because, as discussed above, composers ideally would like their material to sound relatively unique. Furthermore, the algorithm is not stochastic, as the output is determined by the heuristics used by the SMT solver. Thus, it is difficult to obtain a diverse list of possible outputs from a single specification. The most obvious solution to this is to use a Uniform-SAT module to maximize the independence between executions. However, at this time the number of SAT clauses is too large to apply Uniform-SAT. Optimizations which either reduce or modularize the number of SAT clauses could make this approach feasible, which would drastically increase the appeal of this algorithm.

Areas for Collaboration

As mentioned above, there is not sufficient academic literature on what makes something sound "underwater" or "eerie." Collaborating with music experts could prove useful in developing more precise and accurate definitions, as could partnering with HCI experts who work on learning conditional user preferences.

In particular, collaborating with film composers could provide much needed feedback on the approach as well as benchmarks to compare to and specific advice on areas for improvement within the algorithm.

Collaborations with experts in SMT solving could prove as rewarding as the interactions with film composers. This is because the approach is fundamentally limited by what is tractable to compute, and significant sacrifices were made in the name of efficiency (namely deciding rhythmic contour/harmonic progression, rhythm, and pitch as three separate steps). If it were tractable to produce end-to-end systems, we would avoid issues such as pitches and rhythms being unable to fit a given harmonic progression well.

Conclusion

In conclusion, this paper proposes a logical theory of film scoring, as well as a theory for creating and verifying coherent melodic contours. Empirical studies suggest that film scores do have significant structure and that this method may be promising. User studies in the future could enhance the impact of this research.

List of Stylistic Terms

1. Suspense (defined by a high tension level, resulting from some combination of having an ostinato, tremolos, dynamic contrast, chromaticism, rising contour, and dissonance
2. Relaxed (defined by having a low tension level)
3. Zydeco (defined by having three of the following: flamenco percussion, guitar, Andalusian cadence, and phrygian mode)
4. Americana (defined by using 3 of 4 of major scale, harmonica, I-IV-V progression, and washboard drum pattern, as well as lack of synth-based sounds)
5. Sci-fi (defined by having two of three of synth-based sounds, ostinatos, and modes of limited transposition)
6. Jazz (defined by having dominant seventh chords, a jazz-kit-based rhythm, and electric guitar or other stereotypically jazz instruments)
7. Romance (defined by 3 of 4 of major key, moderate rhythm, high pitch, string instruments)
8. Happy (defined by 3 of 4 of major key, fast rhythm, high pitch, consonance)
9. Sad (defined by 3 of 4 of minor key, slow rhythm, low pitch, dissonance)
10. Underwater (defined by use of marimba or whole-tone scale)