Meiji University

Graduate School of Advanced Mathematical Sciences

Academic Year 2019

Doctoral Dissertation
(Abstract)

Automatic Music Completion through Statistical and Musicological Modeling

統計学的および音楽理論的モデルに基づく音楽自動補完の研究

submitted by
Christoph Matthias Wilk

Frontier Media Science Program

1 Research Goal and Contributions

This dissertation proposes automatic music completion as a new class of music information problems, which are characterized as tasks of automatically generating complete music pieces from any incomplete fragments of music. These fragments can belong to multiple levels of musical abstraction, such as notes, harmonies, and musical keys. Therefore, one can interpret automatic music completion as a generalization of several conventional music information problems, including automatic melody generation and harmonization (generating harmonies from melodies).

The goal is to turn any musical idea of a user into music pieces, allowing users to quickly explore new ideas, as well as enabling musically inexperienced users to create their own music. Thus, automatic music completion is motivated as a fundamental principle for music composition assistance with a focus on user input. In contrast to previous research, this dissertation follows the new direction of allowing user input to be as free as possible, which implies the challenge to handle input of any size (a few notes, melodic fragments, almost complete melodies, no input, or multiple note candidates to choose from), as well as the goal to provide as many modes of input as possible (notes, multiple melodies, harmonies, rhythm, musical keys, abstract tuning parameters, etc.).

This principle is applicable to a wide variety of music, and the dissertation presents mathematical models and algorithms for the automatic completion of four-part chorales. The composition of such chorales is a fundamental compositional discipline of Western music, because such music pieces combine principles of harmony theory and voicing (the arrangement of chord notes in multiple voices), both of which are central to the composition of many styles and genres of Western music. The interdependence of harmony and voicing, as well as a significant number of compositional rules, make chorale composition a complex and challenging problem. The automatic generation of chorales is also particularly suited for research, because the large amount of music theory that exists for chorale composition makes it easier to evaluate generated music (writing four-part chorales is a common task for students of music composition, and can be evaluated by teachers quite objectively). The automatic music completion systems implemented for evaluation allow

users to freely constrain the melodies of the four chorale voices (soprano, alto, tenor and bass) as well as the underlying harmony progression.

Overall, the main contributions of this dissertation include (1) the formulation of automatic music completion as a music information problem, (2) multiple mathematical models and algorithms to solve this problem for four-part chorales, (3) proposing parameterized modeling of harmony as a new approach to automatic music generation, and (4) presenting several evaluation methods as well as their application to the models and algorithms of this dissertation.

2 Musicological Problem Analysis and Mathematical Formulation

The theory and rules for composing four-part chorales can be separated into harmonic structure and voicing constraints, which are, however, not completely independent from each other. Especially in classical music, but also in many other genres of Western music, harmony progressions follow a certain order of chords which fulfill specific harmonic functions related to tension and resolution. A mathematical model of harmony should be able to reproduce this order, but to increase the freedom of user input, it should also be able to handle input that contradicts these common patterns. On the other hand, the theory for voicing consists of a considerable number of rules and constraints for placing chord notes with respect to each other. For example, the pitches of consecutive notes in a melody should not be too far apart from each other, and the same is true for the pitches of the notes of two neighboring voices (e.g., soprano and alto). There are also more complex rules that state that specific intervals (distance between two notes) between two voices should not occur in succession. A mathematical model should avoid violating these rules, while being flexible enough to handle user input that entails rule violations.

An automatic music completion system that follows these rules of music theory guarantees a certain degree of musicality and reliability, because a user can expect it to generate music according to widely accepted principles. However, since the design principle of free user input dictates that a user should be able to intentionally violate these rules, they are treated as loose constraints. This means that the used mathematical models

should assign high probabilities to note constellations that follow music theory and low probabilities to those which do not. The problem of automatic music completion can then be formulated as the optimization problem of finding a set of notes and harmonies that contains the user input as well as the generated missing parts, and follows music theory as best as possible according to the corresponding mathematical models. The notes and harmonies for the missing parts have to be chosen from a large pool of possible candidates, which makes this optimization problem a computational challenge.

3 Mathematical Models and Algorithms

The musical models of this dissertation were designed in a modular fashion. A complete model of four-part chorales consists of a harmony model (hidden structure) and a voicing model (observed notes). The basis for this modularity is the assumption that harmony and voicing, while interdependent, are individual aspects of music composition that can be considered separately. The dissertation presents two harmony models as well as two voicing models.

The first harmony model is based on learnable n-gram probabilities, a concept of natural language processing, adapted to harmony based on the similarity between harmonic structure and language grammar. The model uses 3-gram probabilities, but more important than the length of the n-grams is the information content of each harmony, i.e., how a single harmony for n-gram training is defined. If harmonies were only defined as sets of specific pitches, such as C:maj (C E G) or F:maj (F A C), the harmony model would be unable to differentiate whether the chord sequence C:maj → F:maj occurs in the key (harmonic context) of F major or in the key of C major. However, these two cases would sound significantly different to a listener, the former being a harmonic resolution or conclusion, while the latter usually opens up a harmony progression. Therefore, harmonic information is encoded with respect to underlying keys, also considering the possibility of key modulation (the change of keys).

The second harmony model proposes a completely new approach to generating music, based on tuning parameters derived from music theory. These parameters quantify music-

theoretic properties of harmonies, and allow users to individually specify what kind of harmony progression they want to be generated. The parameter values can be set individually for each beat or bar in the music piece to be generated, meaning that the character of the harmony progression can dynamically change throughout the piece. The three parameters introduced in the dissertation are related to active tones, cadences and key modulation. The term active tone is used to describe a note that does not belong to the current harmony or demands resolution towards another note for another reason (e.g., leading tones, seventh notes or altered notes in altered chords). All these types of notes have in common that they introduce harmonic tension into the music, and therefore, the corresponding parameter allows users to specify how much active tone induced tension an algorithm should generate at each position in the music piece. The second parameter allows users to influence the generation of cadences, which are harmonic resolutions of varying strength, enabling users to specify where and how strongly certain musical phrases should be concluded. Lastly, the third parameter allows to tune where and how abruptly the underlying key of a harmony progression should change to another key. This parameterized approach to harmony generation enables users to define their own individual styles of harmony progressions, without relying on learned probabilities that would result in imitating the genre or style of the training data.

The major challenge when designing voicing models is to handle the high combinatorial complexity of four-part voicings. Assuming typical voice ranges of singing voices, each of these four voices can potentially perform 25 different notes. Furthermore, proper composition of four-part voicing requires not only information about the notes in the current voicing, but at least also information about the notes of the previous voicing in order to write good melodies. There are 25^8 possible pairs of two voicings, which inevitably leads to problems with data sparsity in the training data.

The first voicing model handles the high combinatorial complexity of four voices by assuming probabilistic independencies based on human understanding of music composition. It combines statistical learning from data with some heuristics to account for more

complex music-theoretic principles. In particular, there are some compositional rules that are very specific and also quite strict, i.e., violation should be avoided if possible. The most popular of these rules is the avoidance of parallel fifths and octaves, which are note constellations where the notes of two voices are a fifth or octave (distances between notes) apart and their melodies move by the exact same amount, causing the two voices to end up in the same interval. This rule is implemented heuristically by suppressing the probability of such note constellations by a constant factor. However, there are also less clear voicing rules, e.g., how to balance the spacing between notes in the voicing, or how to write smooth melodies (the melody intervals should be small, but it is not a major problem if they are sometimes larger). To model these rules, the probabilistic model is split into structure (notes within a voicing) and transition (intervals between notes in consecutive voicings) probabilities, which are learned from data, and factorized such that one can apply probability smoothing to account for the remaining data sparsity.

The second voicing model reduces the need for heuristics and is independent from the number of voices, i.e., even if trained on a data set containing music pieces for four voices, it could be applied to generate music pieces with three or five voices. It is based on several trainable factors that each capture a part of the musical context a composer would consider. The dissertation introduces three different types of factors: One type only captures the immediate connection between two consecutive notes in a melody. This factor is computed for each voice melody, i.e., in the final model there are as many factors of this type as there are voices. Another factor captures the larger melodic context, but less precisely, namely it only distinguished between small and large movements upwards and downwards instead of exact intervals, but thanks to the reduced number of possibilities this reduction entails, it is possible to consider multiple consecutive notes in a melody with fewer sparsity problems. The last type of factor captures the relative motion between the melodies of two voices, i.e., how the distance between their notes change during the melodic development. This enables the model to, for example, learn the avoidance of parallel fifths and octaves from data.

Given a complete music model consisting of one of the harmony models and one of

the voicing models, one needs an optimization algorithm to compute a complete music piece based on the chosen model. The dissertation presents three search algorithms as well as multiple performance improvements for these algorithms. The first algorithm is uniform cost search, which guarantees optimality of the solution, but is generally to slow to be applied to the highly combinatorially complex problem of four-part music generation. The second algorithm relies on beam search to reduce the runtime while possibly sacrificing quality, which can be tuned by the user setting the beam width. Lastly, the dissertation introduces a nested beam search algorithm, that keeps track of harmonies and voicings individually (keeping a certain number of harmonies, and a certain number of voicings for each harmony in the search beam) in order to increase the diversity of the beam content for handling unexpected user input.

The presented performance improvements for these algorithms generally restrict the search space while avoiding to make the problem infeasible or remove possibly optimal solutions. Some restrictions are based on heuristic constraints derived from music theory, while others dynamically restrict the search space based on user input.

4 Experimental Evaluation

While the evaluation of generated music is inherently difficult due to the subjectivity of taste, the dissertations explores multiple methods for this task. Generated music is evaluated with respect to quantitative metrics derived from music theory, put to the test in subjective evaluation experiments where users can use the system to turn their own musical ideas into music pieces, as well as subjected to the analysis of a professional composer.

The quantitative experiments showed that the automatic music completion system is good at following music theory, surpassing results obtained from a state-of-the-art deep learning model in that regard. The feedback from both the subjective evaluation experiment as well as the professional composer was positive. The composer furthermore provided comments on specific details of the system that could be improved, such as voice leading during key modulations or more sophisticated use of nonharmonic tones.