PROBLEM-SOLVING STRATEGIES IN A MUSIC TRANSCRIPTION SYSTEM

Bernard Mont-Reynaud

CCRMA. Stanford University

ABSTRACT

Music transcription is a significant problem in machine perception. The system discussed in this paper, MANA. takes as input either the sound of a recorded performance, or data captured using a musical keyboard. It produces conventional musical notation as output. It has successfully handled pieces ranging from 18th-century piano music to improvisations on conga drums, in the Afro-Cuban style.

The paper describes the key ideas and techniques found in the temporal analysis component of MANA, the goal of which is to assign a rhythmic value to each note played.

Perception seems to result from the interplay of sensory evidence and pre-existing mental structures, data-driven agents pitted against model-driven agents in the formation of hypotheses that seem tolerable to our prejudices and not too distant from the data. To this first order view one must add a second order one, which allows the system to notice patterns among partially elaborated hypotheses, and to use these patterns to alter confidence ratings, according to "peer pressure" rules.

Concerning the use of multiple criteria for hypothesis evaluation and pruning, which necessary in the context of perception systems, it is argued that the use of partial orders over multi-dimensional spaces of natural criteria yield more robust methods that approaches based on combining criteria down to scalars.

I INTRODUCTION

An ongoing research project at CCRMA (Stanford University) is concerned with exploring an AI approach to the recognition of musical structures in sound and other forms of data. The effective Integration of numerical processing and symbolic processing towards the goal of machine perception is a central technical theme, while music transcription is the task goal.

Automatic transcription from sound is clearly more difficult than transcription from keyboard strokes. It requires, prior to the musical analysis which is needed in both cases, an initial stage of acoustic analysis. Further refinements may be obtained by feeding information gathered from the musical context back to acoustic levels of the analysis system.

It is perhaps less obvious that transcription from keyboard strokes alone presents a significant challenge. The problem is trivial if one either forbids expressive performance and rhythmic complexity, requiring strict metronomis accuracy in timing; or if one accepts musically absurd (but physically accurate) transcriptions, as may be obtained by setting the tempo and the metrical grid resolution to fixed values, and rounding durations to the nearest grid point. But one should keep in mind that, for example, an eighth note (1/2 of a beat) may be played shorter than a triplet eighth (1/3 of a beat) without confusing human listeners, provided the appropriate musical context. How to provide sufficient context mechanisms to reach correct decisions, often in spite of the physical evidence, without carrying context-driving to excess, may well be the most sensitive question in the design of robust perception systems.

In this paper, the discussion will be limited to the musical analysis component of MANA, and further restricted to the temporal aspects of the analysis. The reader is referred to [F0S82a-b, SCH84] for a discussion of acoustic analysis, and to [CIIA82, M0N84] for a more comprehensive account of the methods used in musical analysis. Section II of the paper provides an overview of the system, and presents some of the key ideas and techniques used in temporal analysis. Section III discusses some of the problem-solving strategies whose compounded effect is a carefully measured dose of context-driving. Section IV draws some conclusions and outlines the direction of future research, in which learning issues will play an increasing role.

II MUSICAL ANALYSIS

The performance goal initially set for the system was to exhibit a fair degree of musical wisdom while transcribing expressively played 18th-century music. The input was restricted to a single musical voice. A first redesign of the temporal analysis permitted opening up the range of musical styles without having to train specific styles into the program. Another redesign, which is is progress, is aimed at dealing with polyphonic sound. However, the system as discussed here only with one voice at a time

Further progress towards increased grasp of the musical structure. Including more robust methods for tempo tracking and meter determination, require moderately powerful learning techniques, of the "unsupervised" type. The idea Is to uncover the most significant musical patterns in a piece, and to use the temporal patterns of pattern occurences as hints to the tempo and meter. Near-misses also offer Important possibilities, notably for error detection, near-missed to initiate self-doubt.

A Rational Approximation Generation

Note values, metrical intervals and metrical positions are rational numbers, expressed In terms of a reference unit. Given (say) a note duration X and the duration R of the metrical reference unit, both in seconds, one wishes to

generate rational approximations of X/R as candidate note values for the given note, in terms of the given unit. The choice of rational approximations must take into account at least two criteria: closeness to the data (or *fit),* and simplicity of the fraction. The latter must be understood in terms of how simple or natural the resulting musical notation would be. The techniques of multi-criteria filtering (section III-B) come into play at this point, to limit the number of answers retained.

The program relies on the same generator for a variety of different tasks, but it varies parameters such as the set of acceptable numerators and denominators, and the error thresholds. The most obvious use is in the generation of hypotheses for individual note values. There are others, and they use different tunings.

Finally, the complexity measure for metric fractions is allowed to change over time. For example, fractions such as 1/3 and 2/3 may become *simpler* than 1/4 or 1/2 after sufficient statistical evidence of ternary meter has been gathered. This is an applications of *peer pressure* (cf. III-A).

B Approach to the rhythmic value problem

The goal is to express note values as rational fractions of a "whole note. It turns out that this goal must be approached in a round-about manner, which goes in 4 steps as follows: (a) choose a reference unit R; (b) express note values as fractions of a reference unit R; (c) analyze the patterns of note values in order to determine values for metrical divisions like beat, bar and (at least) the whole note; and (d) convert from R-values to whole-note values.

Problem (b) is the rhythmic value problem. Problem (c) is the meter problem, which is not discussed here. From a formal point of view, it appears that the value of R n step (a) is arbitrary. This is not so: the behavior of the rational approximator depends very much on the reference unit used. A first set of rules is used to determine R. Statistical clustering of note values plays a role there.

The next key idea is to localize the problem, to regions where the tempo is supposedly held constant. This factors out the global tempo fluctuation, which can be represented by a piecewise linear correspondence between musical time and physical time. The slope of each linear segment is a tempo value, and the list of successive segments is regarded as the *tempo line* of the piece. A second set of rules, based on simple rhythmic and melodic patterns is used to determine the *structural anchors.* These notes are singled out as likely candidates to occupy strong positions in the (as yet undetermined) metric grid. For instance, these notes might occur at downbeats, or at least at beat boundaries. The endpoints of the tempo line segments, on the physical time axis, are placed at these structural anchor points.

A third set of rules determines the metrical duration of each tempo line segment, in terms of R. This completes the construction of the tempo line. Once the tempo is known within each segment, it is possible to use the rational approximation generator to associate with each note a set of candidate rational values, in terms of R. Among the combination of approximations which add up to the desired metric length it is possible to select one that represents the best compromise between musical simplicity and closeness to the data, for this segment.

In terms of search paradigms, choosing the tempo line before assigning individual note values is a special case of solving a problem in a much smaller abstract *space.* Once an abstract solution S is obtained, one returns to solving the original problem under the constraints imposed by S.

III PROBLEM-SOLVING STRATEGIES

MANA operates primarily in a bottom-up fashion, but it does use top-down constraints. The important point is that very few a *priori* top-down constraints are used. In other words, the program tries to refrain from having too strong a notion of what music "must" be like. On the other hand, a *posteriori* constraints are heavily relied upon. Such context-driving operates in a top-down manner, but the context is acquired during previous bottom-up hypothesis generation. The idea is to promote homogeneity, or self-consistency. In other words, the program operates under the assumption that a piece defines its own "style" and then wants to see more of the same "style". This feedback mechanism, termed *peer pressure,* is further examined in section III-A.

Instead of backtracking, the system relies on a multiple-value technique, whereby a set of alternatives is operated on parallel. Techniques of *multi-criteria filtering* (cf III-B) are used to to prune the sets of alternatives whenever evaluation criteria are added or modified.

A Peer pressure

The system uses several levels of abstraction in its description of the data. It combines data-driven and context-driven methods of hypothesis generation and evaluation. Data-driven methods use features obtained at one level to generate hypotheses at the next level. Context-driven methods use information gathered *at* higher levels to re-evaluate and possibly re-generate lower-level hypotheses. A key problem is to arrange that feedback loops between the bottom-up and top-down modes either converge rapidly, either to a stable consensus, or to no effect at all. This problem is adressed by the *peer pressure* strategy, a method that allows a collection of hypotheses (obtained using statistics, clustering or pattern discovery methods) to promote hypotheses with similar contextual features.

Using rather broad terms here, let us call "data" some collection of hypotheses at a given level of description, and "partial model" some description of the the set of data. The partial model might be a set of patterns found in the data, or a statistical summary of some aspect of the data.

Peer pressure, which may be viewed as a noise reduotion technique, operates in two steps. The first step extracts a partial model from the data. The model must be "safe" before proceeding: under poor conditions, peer pressure may amplify noiset The second step modifies the data to create a better agreement with the partial model. Each data point is re-examined in terms of this agreement. If the agreement is weak, the data point becomes a candidate for modification, that is, deletion, replacement by one or more other data points, or simple adjustment. Strong patterns always get stronger by the use of this technique, which decreases the perceived disorder in the interpretation of the data.

B Multi-criteria filtering

In an attempt to achieve robustness over a wide range of examples and styles. MANA relies on a variety of independent methods for hypothesis generation, and also on a multiplicity of evaluation criteria. Rather than using backtracking, MANA usually carries a small number of

hypotheses in parallel. Excessive pruning of a set of hypotheses, most likely to occur during early stages of analysis, may cause a system to overlook a good solution whose value is not yet obvious at that stage of the game. Too little pruning, on the other hand, not only slows things down, but leaves noisy points which may affect the operation of peer pressure. Thus, it is important to maintain a balanced degree of indecision.

In a multiple-criteria situation, a popular approach is to use a weighting scheme to produce a scalar from all the criteria. MANA considers this method to be one of last resort, as it all too often leads to erroneous decisions. Trying to improve the approach by making the weighting scheme dependent on context leaves the problem unchanged for initial decisions, which must be made before sufficient context is established. It also forces one to develop ways of changing the weights as a function of context, a rather hazardous enterprise.

It seems better to deal with the original multiplicity of criteria directly, especially during pruning *stages.* The intuition behind the scheme we use is quite simple: If hypothesis A is no worse than hypothesis B in any of the criteria, and better in at least one criterion, then (and only then) B should be pruned. This idea immediately generalizes to an arbitrary partial orderings in the space of criteria. We say that an hypothesis is *dominated* if it is larger than some other hypothesis, with respect to the chosen partial ordering. *Undominated* hypotheses are those that correspond to minimal elements in the space of criteria. They are the only ones retained past a pruning stage.

Since the retained hypotheses are minimal in the partial ordering, they trade one criterion for another: one hypothesis might be close to the observed data while the other is simpler but farther away, and a third is intermediate in both respects. The scheme is uniform, but leaves much flexibility in the choice of the partial ordering. In our application, this technique has been found extremely effective.

IV CONCLUSIONS

Elements of a methodology for achieving robustness in perception tasks have been gathered, along with some ideas more specific of the transcription domain. In terms of performance, the automatic transcription system described has also produced some rather interesting results. Naturally, there is much more work to be done, both to improve robustness in single-voice examples, and to deal with polyphonic data.

Experiments underway suggest that the response to these challenges will be based on (a) extending the use of adaptive feedback strategies into the acoustic levels of the system, so that signal processing and reasoning work hand-in-hand, and (b) developing the techniques for pattern discovery and selection to the point where second-order patterns become a reliable source of information.

In fact, in the process of this research, it has become more and more apparent that the musical domain provides an ideal setting for in-depth studies in machine perception and machine learning. Statistical pattern recognition and unsupervised inductive learning [MIC83] both have Important roles to play in a flexible musical understanding system.

REFERENCES

[CHA82] Chafe, C, B. Mont-Reynaud and L. Rush. Toward an Intelligent Editor of Digital Audio: Musical Construct Recognition, Computer Music Journal 6:1 (1982), 30-41.

[FOS82a] Foster. S.. J. Rockmore and W. Schloss. Toward an Intelligent Editor of Digital Audio: Signal Processing Methods, Computer Music Journal 6:1 (1082), 42-61.

[F0S82b] Foster, S. and A. Joseph Rockmore. Signal Processing for the Analysis of Musical Sound, ICASSP Proceedings. Paris. May 3-6, 1082.

[MIC83] Michalsky. R.. J. Carbonell. and T. Mitchell, eds. *Machine Learning.* Tioga Publishing Co., Palo Alto, 1883.

[M0N84] Bernard Mont-Reynaud, et. al. *Intelligent Systems tor the Analysis of Digitized Acoustic Signals, Final Report.* Technical Report STAN-M-16, Department of Music, Stanford University. Stanford, California (1984)

[SCH84] Schloss, W. On the Automatic Transcription of Percussive Music. PhD thesis, Department of Speech and Hearing. Stanford University, Stanford. California (in preparation)

• This work was supported by the National Science Foundation under Contracts NSF MCS-8012476 and DCR-8214360.