THE FUTURE OF MUSIC IR: HOW DO YOU KNOW WHEN A PROBLEM IS SOLVED?

Eric Nichols
Indiana University
epnichol@indiana.edu

Donald Byrd
Indiana University
donbyrd@indiana.edu

1. INTRODUCTION

This workshop is intended to address the future of music IR research in the coming decades by exploring which topics will be important once the current important problems in the field are solved. However, we believe that most researchers as well as the public are unclear about what it even means for a music IR problem to be "solved". The best existing "solutions" to nearly all music-IR problems may give the desired result a high percentage of the time, but only for music in certain styles and genres—usually styles and genres of Western popular music, which research to date has focused on far more than any other kind. One ethnomusicologist who is particularly interested in music-informatics technology has commented that almost none of the software she has tried works for the type of music she's interested in (Polina Proutskova, personal communication, June 2009). How do we decide when the error rate is small enough and the range of music handled wide enough to call a problem solved?

Because the notion of a music IR "solution" is so problematic, we believe that the coming decades of work will turn out differently than many researchers expect. We suggest that confusion about the notion of solutions results in over-optimism for current results. Thus, to make reasoned prognostications about the future of music IR, we encourage researchers to think more critically about the state of the art.

2. WHAT IS A SOLUTION?

The field of mathematics, as opposed to music IR, could be expected to have clear-cut notions of problems and solutions; after all, mathematics is well-defined if anything is. In a famous 1900 presentation [1], the great German mathematician David Hilbert described 23 problems he deemed critical to future progress in mathematics. Overall, he was remarkably accurate. Nonetheless, some of his problems have turned out to be too vaguely stated to permit solution; for others, no consensus has been possible as to whether the problem is resolved or not. In a different field, computer graphics, Sutherland proposed a somewhat similar list of 10 problems [2]. However, some of these are considerably less well-defined and the issues are more difficult in determining if a problem has been solved. The problem of digital half-toning, for example, is to generate an image using different arrangements of dots, instead of solid color, to generate the appearance of another shade of color. Human visual perception is necessarily involved here, and the quality of a solution is dependent on human perception.

We believe music IR is much harder still. The reason is that, with music, almost everything is a question of human perception. Obviously, when we say a problem is solved, we mean it is solved *correctly*. The notion of "correct" mathematics certainly makes sense, and so does "correct" graphics; but what is "correct" music, or—perhaps more relevant—what is a "correct" statement about music? Music pedagogy has always had rules for correctness (e.g., in species counterpoint writing), but they have changed dramatically over the centuries, because composers have repeatedly ignored the current "rules". And when, in *any* art, there is a contest between creators and rules, the rules always lose—or, more precisely, they change to accommodate the aesthetics of the new art. Consider the mere existence of university courses like 17th-century counterpoint and 18th-century counterpoint, which teach different rules. In mathematics, correctness is always objective; in computer graphics, it is often objective, but often subjective. In music, correctness is almost always subjective.

3. FIVE DIFFICULT PROBLEMS

We now present several problems, ranging from some which most researchers probably consider solved, to problems for which few would make that claim. We explain why any existing solutions need to be improved, and suggest how work in the coming decades might make progress on all of these problems.

3.1 Monophonic Transcription

By monophonic transcription, we mean the process of converting an audio signal describing music with just one note at a time into a musical score, with no additional prior information such as knowledge of the tonality or

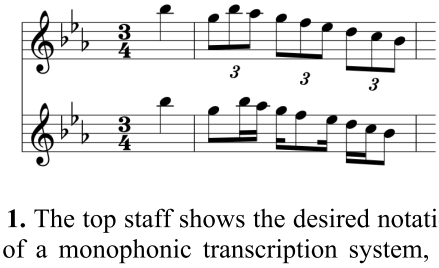

Figure 1. The top staff shows the desired notation for the output of a monophonic transcription system, while the bottom staff shows a hypothetical transcription result. Expressive and inexact performance result in a rhythmic pattern that is not isochronous as intended.

tempo. Many people consider this problem to be well-understood and long since acceptably solved; we disagree.

Consider rhythm in various kinds of music. We must know where the beats are for many purposes, e.g., in order to organize the music into readable notation. Even if we had a perfect method of beat detection, monophonic transcription would still be difficult because performers do not play metronomically. The length of each measure is also ambiguous, and it is easy to be wrong by a factor of two – should a piece be notated in 2/4 or 4/4? This kind of metric distinction is a subtle difference. More troubling is the determination of exact start and end times of notes. Even if we assume that a transcription will only use notes of a particular minimum relative duration (such as 32nd notes or eight-note triplets), the presence of rubato greatly increases the number of possible rhythmic transcriptions. See Figure 1 for an example of a common problem in monophonic transcription. The top staff is the beginning of a Schubert *Impromptu* with an entire page of triplets in the right hand. Fixing this problem would require a deeper understanding of the global structure of the music and the cognitive desire for repeated rhythmic patterns, in contrast to the local approach illustrated here. A solution would also require a good metric to determine the amount of mismatch with the desired solution.

3.2 Beat Detection

Detecting the time-points at which significant musical beats occur is something human listeners do remarkably well: at concerts, for example, crowds of thousands often clap or move their bodies in time with the music, and everyone seems to have a similar interpretation. An algorithmic approach, on the other hand, might work fairly well in certain genres, such as those with a strong drumbeat, but a large amount of musical context is still required to determine which beats are stronger, to establish a reasonable metric hierarchy.

This problem is much harder than it might seem. In many genres, beats are not necessarily regular, and music with no percussive element is common: think of unaccompanied folk songs, or choral music. Tempo can vary, and notes might be delayed or accelerated relative to the regular beat structure. The interplay between tempo, beats, rubato, expressive timing, etc. is highly complex, and it requires musical sophistication to tease apart the various factors which contribute to beat placement.

3.3 Optical Music Recognition

The OMR (optical *music* recognition) problem is the analogue of OCR (optical *character* recognition): given an image of a musical score, the goal is to generate a structured representation in a form such as MusicXML. OMR programs have been available commercially for at least 15 years; that fact alone suggests a solved problem. However, meaningful figures on accuracy of OMR programs are virtually impossible to come by, and there is little evidence that the state of the art has improved much since the first research report [3], in 1994. For several reasons this is not surprising. One is that the structure of music notation is so complex, it's very difficult to even say how accurate the OMR version of a given page of music is. With media like text, it is reasonable to assume that all symbols and all mistakes in identifying them are equally important. With music, that is not even remotely the case. It seems clear that note durations and pitches are the most important things, but after that, nothing is obvious. How important are redundant or cautionary accidentals? Fingerings? Mistaking note stems for barlines and vice-versa both occur; are they equally serious? A completely different reason is that music notation varies in difficulty so widely that most programs do very well on some pages—for example, chorales—but fail miserably on others; idiomatic piano music is particularly challenging. It is well-known in the document-recognition community that some types of documents pose far more difficult evaluation problems than others; music notation is certainly toward the difficult end of the spectrum [4].

3.4 Query by Humming

The problem of retrieving a piece based on a given audio query such as a hummed melody has received considerable attention. But query-by-humming research activity has greatly diminished in recent years as researchers have realized how difficult the problem is.

Few people are capable of singing their queries accurately. It requires much musical knowledge to infer what a person intends to sing given an inaccurate performance. For example, pitch intervals may be sung inaccurately so that notes do not stay within a particular tuning system; however the pitch contour may be correct [5]. Human experts may be able to focus on which features of the query are relevant, while it is hard for an automatic system to make such context-sensitive judgments.

In general, a good solution to the query-by-humming problem requires a sophisticated mapping – a musical analogy – to be made between features of the query and the retrieval target. For most pieces in the target database, an analogical mapping will not have a good fit, but for a few good results a relatively strong mapping should be established. This analogical mapping problem is difficult; it depends on picking out the critical features for each potential mapping. And, of course, the mapping depends on the genre of music in the database and the style of the query itself. We suspect that better solutions to this problem will come about via deeper study and application of music cognition.

3.5 Motif Detection

Music generally is based on repeated motifs. The problem of motif detection may seem like a simple pattern-matching exercise which could be solved by, say, exhaustively searching for repeated melodic contours [6]. However, consider the four-note motif which begins Beethoven's Fifth Symphony: three short notes of the same pitch and duration, then a long note a third lower (Figure 2a). As the piece progresses, the motif is presented in various altered forms, and detecting the various instances of the motif is surprisingly subtle: in the development, it occurs in a version consisting of only two short notes and one long, and all of the same pitch (Figure 2b). By itself, 2b would surely not be heard as an instance of 2a, but before

2b appears, many passages like 2c have occurred, developing the motif gradually so that, indeed, 2b *is* heard as an instance of 2a. We believe that a human listener recognizes motif instances as such via a process of musical analogy-making. As the piece progresses forward in time, a listener continuously compares what they hear with their memory of the previous music, and they recognize a motif instance when they can form an analogy between the current music and a previously-heard motif version.

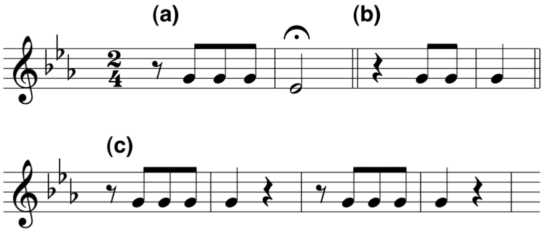

Figure 2. Three versions of a Beethoven motif.

Thus, whether a series of notes, chords, or other musical material is heard as another instance of a motif is very dependent on the precise musical context.

4. MUSIC COGNITION AND MUSIC IR SOLUTIONS

To make significant progress in the problems above, we suggest that researchers must think carefully about what it means to solve these problems. Any solution should carefully define under which contexts it provides a good answer. Furthermore, musical concepts are intimately connected with the human perceptive and cognitive process. The limits of human short-term and long-term memory, perception abilities, and attention result in a listening process that is idiosyncratic and results in unique representations that are radically different from computer representations of musical data.

One might think that our focus on human cognition is unnecessary: after all, computers routinely perform better than humans on tasks such as mathematics or playing chess by using different methods than humans. But we cannot remove human cognition from the music IR equation in the same way as in chess, because music is defined by human perception, not sharp-edged rules. MP3 and the other audio data-compression methods that have revolutionized the music world in recent years were evaluated by human listeners, not algorithms [7].

An algorithm for motif detection, for example, could try to exhaustively find all the instances of a given sequence of notes, subject to a series of transformation operators, but many such transformations are not cognitively relevant. Motifs only come into existence for a listener when they are recognized as such in the mind, so the very definition of a motif and algorithms to search for them must include understanding of human cognition. The same is true for the related audio query by humming problem. OMR is likewise dependant on human visual perception to make sense of notated musical symbols. Beat detection for listeners at a concert involves not only a computation of accented notes, tempo, and rubato, but also likely involves physiological issues such as the embodied experience of nodding the head or tapping the foot along to the music. Finally, the transcription problem also is really the problem of recording what the typical human hears when listening. Despite minor rhythmic variations and inaccuracies, the residual memory of the heard music is a critical piece of the transcription puzzle. Otherwise, transcription would be nothing more than a digital recording process. Instead, it is also sophisticated and intimately connected to the human listening experience.

In summary, music IR involves sophisticated, deep problems for which progress will require appropriately sophisticated and deep thought. How much progress have we made so far? Jeremy Pickens (personal communication, 2009) wrote that only one problem in music IR has really been solved, namely the “Shazam” problem of recognizing a given recording in a noisy audio signal, but that it’s not clear that it *is* a music-IR problem. We agree.

We recommend that as researchers consider the future of music IR, they keep in mind two guiding principles: 1) be realistic about the extent to which a problem has been solved, and 2) respect the importance of human cognition in musical experience. The ultimate test of a solution involves human evaluation, so the development of successful music IR solutions will incorporate careful contemplation of the role of the human listening process.

5. REFERENCES

[1] Hilbert, D. *Bulletin of the American Mathematical Society,* 8, 1902, pp. 437-479.

[2] I.E. Sutherland, "Ten Unsolved Problems in Computer Graphics," *Datamation*, Vol. 12, No. 5, May 1966, pp. 22-27.

[3] Selfridge-Field, E., Carter, N., et al (1994). “Optical Recognition: A Survey of Current Work; An Interactive System; Recognition Problems; The Issue of Practicality”. In Hewlett, W., & Selfridge-Field, E. (Eds.), *Computing in Musicology,* vol. 9, pp. 107–166.

[4] Droettboom, Michael, & Fujinaga, Ichiro (2004). “Micro-level groundtruthing environment for OMR”. In *Proceedings of ISMIR 2004,* Barcelona, Spain, pp. 497–500.

[5] Meek, C. & Birmingham, W. (2002). “Johnny Can’t Sing: A Comprehensive Error Model for Sung Music Queries”. In *Proceedings of ISMIR 2002*, Paris, France, pp. 124–132.

[6] Weyde, T. & Datzko, C. (2005). “Efficient Melody Retrieval with Motif Contour Classes”. In *Proceedings of ISMIR 2005,* London, UK, pp. 686–689.

[7] Pohlmann, Ken (2005). *Principles of Digital Audio,* 5th Edition. New York: McGraw-Hill, pp. 407-413.