What's Broken in Music Informatics Research?
Three Uncomfortable Statements

Justin Salamon [1]

Abstract

The goal of this paper is to raise awareness to a number of issues impacting research in Music Informatics (MIR), with the hope that it will encourage us, as a community, to address them. It is not my goal to provide a comprehensive review of the literature related to these issues. Given the limited extent of this paper some of the statements I will make are purposely simplified, when in fact the issue is more nuanced. My hope is that the paper serves as a starting point for a conversation. A recording of the talk accompanying this paper is available on the paper's companion website[1].

1. A Very Quick Introduction

1.1. Music Informatics Research (MIR)

Music Informatics Research, also commonly referred to as Music Information Retrieval or MIR, is a research field concerned with the computational analysis, indexing, retrieval, recommendation, separation, transformation and generation of music (Müller, 2007). Like computer vision, much (but not all) of modern MIR research relies heavily on machine learning and, at the time of writing, deep learning.

1.2. MIR Through the Lens of Pitch Tracking

To make this discussion concrete, I will examine MIR research through the lens of a particular problem area, namely pitch tracking, i.e. estimating the fundamental frequency (f_0) of one or more voices/instruments in a music recording. Problems in this area include monophonic pitch tracking (Kim et al., 2018), melody extraction (Salamon et al., 2014) and multiple f_0 estimation (Bittner et al., 2017), all of which are problems I am interested in and have collaborated on.

2. The Elephant in The Room

When it comes to MIR, the elephant in the room is how to *define* the musical concepts we wish to model. What *is* a melody? What *is* musical mood? What *is* genre?

Taking melody as an example, numerous definitions have been proposed over the centuries with examples ranging from "the expression of motion in music" (Helmholtz) or "configuration of beauty" (Hanslick) to "pitched sounds arranged in musical time in accordance with given cultural conventions and constraints" (Ringer). More examples are provided in Chapter 1 of my thesis (Salamon, 2013). Since MIR requires a definition against which models can be evaluated, the research community has also proposed some definitions, including "a series of notes [which] is more distinctly heard than the rest" (Goto & Hayamizu, 1999), "the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music…" (Poliner et al., 2007), or the definition I provided for melody extraction in my thesis in an attempt to make it as unambiguous as possible: "fundamental frequency estimation of a single predominant pitched source in polyphonic music signals with a lead voice or instrument". Subsequently, Bittner et al. (2014) proposed to adopt multiple definitions of melody given that no single definition seemed adequate.

2.1. UNCOMFORTABLE STATEMENT 1

> In MIR, musical concepts are (often) defined by "whatever the annotations contain in the dataset I am using for my research".

More often than not, we use datasets in which musical concepts have been previously annotated by other researchers (or volunteers), consequently avoiding grappling with the question of how to define the musical concepts in the first place. This leads to "definition by annotation" where, in the worst case, we focus on fitting human responses without really knowing what we are actually modeling.

3. Evaluation

In 2012, we ran a comprehensive statistical analysis of melody extraction evaluation using as many datasets and

[1] Adobe Research, San Francisco, California, USA. Correspondence to: Justin Salamon <salamon@adobe.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] http://www.justinsalamon.com/news/whats-broken-in-music-informatics-research

algorithms as we could access (Salamon & Urbano, 2012). The conclusion was that the datasets were too small, artificial, or unrepresentative for the evaluation metrics (Salamon et al., 2014) to be statistically stable. In other words, there is little or no guarantee that the observed ranking between any two algorithms on one of these datasets would generalize to newer datasets. As a community, we continued using these datasets anyway. Bittner et al. (2014) addressed some of these issues by releasing MedleyDB, a significantly larger and more varied dataset. While the dataset was adopted in research, we continued using these small datasets in MIREX, the community's annual evaluation campaign (Downie, 2008). Bosch et al. (2016) released Orchset, another dataset aimed at addressing some of the aforementioned limitations. The dataset was adopted by MIREX, but the original set of problematic datasets remained in use too.

The problem does not only lie with the data, but also with the evaluation metrics themselves. In Figure 1 on this paper's companion website (cf. footnote 1) we can see the reference (ground truth) f_0 of a specific melody (a), and two algorithmic estimates in plots (b) and (c). All three can also be listened to on the website. While visually (and I would argue aurally too) the estimate in (b) seems preferable to the estimate in (c), according to the overall accuracy metric, the main metric used to evaluate melody extraction, both estimates obtain the same score (0.7 out of 1). To address this, Bosch et al. (2016) proposed new metrics that take pitch continuity into account. To the best of my knowledge, these metrics have not been broadly adopted by the community.

3.1. UNCOMFORTABLE STATEMENT 2

> Existing datasets for MIR are mostly too small, artificial or homogeneous, and many metrics in use have serious limitations, but we use them anyway!

4. Choosing Problems

The *Lean Startup* is a methodology for developing businesses and products, which aims to shorten product development cycles and rapidly discover if a proposed business model is viable (Blank & Dorf, 2012). The first stage of the process, Customer Discovery, is designed to help us determine the following: Are we solving a real problem? Who are the customers we are solving it for? Can we find customers who have this problem and would pay us to solve it? Is this the right solution that would actually solve the problem of the customers we found? While in MIR we often do not have direct customers, we do build systems targeted at end users, and we often motivate our research with applications we believe solve problems for these users.

According to the Lean Startup methodology, before writing any code at all you must get out of the lab, talk to 500 people (or more) who you think are potential users (while never mentioning your idea), and determine whether you have identified a real problem and whether your idea actually solves it. Once you identify a problem-solution pair, build the simplest version of your solution and iterate based on user feedback, and always be ready to discover it's not a real problem, or a good solution, and pivot to something else.

Conversely, in MIR we sometimes choose a problem either because we ourselves have this problem or because it motivates a research agenda or technical solution we are eager to explore, without ever talking to prospective end users or considering whether this is in fact a real (or big enough) problem. As a consequence, the impact of the research is sometimes limited. On the commercial end, many of the applications we use to motivate our research never actually make it to market.

4.1. UNCOMFORTABLE STATEMENT 3

> There is a disconnect between MIR research and potential users of MIR technologies.

Some of the biggest exceptions to this include music recommendation on streaming platforms and music fingerprinting, which have seen significant commercial success. Other exceptions doubtlessly exist.

5. Conclusion

Hopefully, these uncomfortable statements raise questions. For example:

1. Should we strive to define the musical concepts we want to model? If so, how? Should we turn to musicology for an answer? To music cognition? Should definitions be application-driven? User-driven?

2. How can academia best contribute in the absence of industry-scale data? Conversely, how can industry contribute back if it can't or won't share its data?

3. Should MIR research be more application driven?

It is beyond the scope of this paper to answer these questions, but I can offer some initial thoughts: I believe there is room for both application-driven and basic-science-driven MIR research, and I think how we define musical concepts should depend on which of the two we are conducting. Today's "blue sky" academic research could potentially transform into tomorrow's commercial application (the work by Goto & Muraoka (1994) on beat tracking back in the 90's, before it had any commercial application, is a case in point). But, if we wish to impact users today, we should factor them into the process of defining our research. In all cases, perhaps a good starting point is to be honest and clear about what kind of research we are conducting, and motivate it accordingly.

References

Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J. P. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *15th Int. Soc. for Music Info. Retrieval Conf.*, pp. 155–160, Taipei, Taiwan, Oct. 2014.

Bittner, R. M., McFee, B., Salamon, J., Li, P., and Bello, J. P. Deep salience representations for f_0 estimation in polyphonic music. In *18th Int. Soc. for Music Info. Retrieval Conf.*, Suzhou, China, Oct. 2017.

Blank, S. and Dorf, B. *The Startup Owner's Manual: The Step-by-Step Guide for Building a Great Company*. BookBaby, 2012.

Bosch, J. J., Marxer, R., and Gómez, E. Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. *Journal of New Music Research*, 45(2):101–117, 2016.

Downie, J. S. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29 (4):247–255, Jul. 2008.

Goto, M. and Hayamizu, S. A real-time music scene description system: Detecting melody and bass lines in audio signals. In *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pp. 31–40, 1999.

Goto, M. and Muraoka, Y. A beat tracking system for acoustic signals of music. In *Proceedings of the Second ACM International Conference on Multimedia*, pp. 365–372, San Francisco, California, USA, Oct. 1994.

Kim, J. W., Salamon, J., and Bello, J. P. Crepe: A convolutional representation for pitch estimation. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, Calgary, Canada, Apr. 2018.

Müller, M. *Information Retrieval for Music and Motion*. Springer, 2007.

Poliner, G. E., Ellis, D. P. W., Ehmann, A. F., Gómez, E., Streich, S., and Ong, B. Melody transcription from music audio: Approaches and evaluation. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(4):1247–1256, 2007.

Ringer, A. L. Melody. Grove Music Online, Oxford Music Online (last checked June 2019). URL http://www.oxfordmusiconline.com/subscriber/article/grove/music/18357.

Salamon, J. *Melody Extraction from Polyphonic Music Signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2013.

Salamon, J. and Urbano, J. Current challenges in the evaluation of predominant melody extraction algorithms. In *13th Int. Soc. for Music Info. Retrieval Conf.*, pp. 289–294, Porto, Portugal, Oct. 2012.

Salamon, J., Gómez, E., Ellis, D. P. W., and Richard, G. Melody extraction from polyphonic music signals: Approaches, applications and challenges. *IEEE Signal Processing Magazine*, 31(2):118–134, Mar. 2014.