Toward Digital Musical Instrument Evaluation Using Crowd-Sourced Tagging Techniques

Michael Everman
University of Miami
Music Engineering Technology
Coral Gables, Florida, USA
m.everman@umiami.edu

Colby Leider
University of Miami
Music Engineering Technology
Coral Gables, Florida, USA
cleider@miami.edu

ABSTRACT
Few formal methods exist for evaluating digital musical instruments (DMIs). We propose a novel method of DMI evaluation using crowd-sourced tagging. Tagging is already used to classify websites and musical genres, which, like DMIs, do not lend themselves to simple categorization or parameterization.

Using the social tagging method, participating individuals assign descriptive labels, or tags, to a DMI. A DMI can then be evaluated by analyzing the tags associated with it. Metrics can be generated from the tags assigned to the instrument, and comparisons made to other instruments. This can give the designer valuable insight into the where the strengths of the DMI lie and where improvements may be needed.

Keywords
Evaluation, tagging, digital musical instrument

1. INTRODUCTION
A multitude of new digital musical instruments (DMIs) and controllers have been proposed in the literature and in the marketplace. Yet, for all the variety and innovation, very few are adopted by a circle wider than their creators and a few associates. Why, with so many inventive minds creating interesting devices, do so few of them see wider use?

To some extent, as noted by Wanderley and Orio [14], many of these were "designed to fit idiosyncratic needs of performers and composers," and as a result "have usually remained inextricably tied to their creators." Indeed, one of Cook's principles of DMI design in [4] is "make a piece, not an instrument or controller." Clearly, there is nothing wrong with designing a custom DMI to fit a specific creative niche. Judging by the number of companies that have been started to commercially market them, though, it is evident some of those creators hoped to see wider adoption of their inventions.

This leads to the conclusion that some sort of method evaluating designs is desirable. However, as noted by Barbosa, et al. [1], each year only a small number of papers presented at NIME consider any type of formal evaluation of new instruments. One reason may simply be that the DMI was designed to fulfill a particular performer's or composer's needs, and did so to their satisfaction. In such a case, a full evaluation was not worth the additional effort since the ultimate goal of the design was met. Another reason, however, may be that few formal evaluation methods have been devised. Though several authors (e.g., [1], [11], [14]) have explored the topic, the development of methods to evaluate and compare new devices remains a largely underserved concern. Most proposed methods focus on quantitative analysis using HCI methods. However, Johnston [7] points out "while ergonomics and efficiency are important, they are not the primary determinants of the quality of a musical interface." Stowell, Plumbley, and Bryan-Kinns [12] noted the need for qualitative analysis methods and proposed using Discourse Analysis (DA), which is borrowed from linguistics, psychology, and social sciences. However, they found that "DA of text is a relatively intensive and time-consuming method." Therefore, a qualitative analysis method that is easier to implement and analyze is still needed.

We propose a new method of evaluation utilizing social tagging techniques to supplement existing methods. Tagging has already been found to be useful for classifying things such as websites ([6]) and musical genres ([3]). Much like DMIs, these do not easily lend themselves to simple categorization or quantitative parameterization. Though tagging has not been applied to DMI evaluation to our knowledge, these papers suggest it may be a useful means of doing so.

Strictly speaking, a successful DMI is one that meets its musical design goals. As discussed above, these need not necessarily include wider adoption. However, the term *successful* is used throughout this paper to indicate an instrument or controller that has achieved or is likely to achieve that goal. Even if that is not a design goal, the evaluation method described is equally useful to instrument designers desiring feedback that will help them improve their designs.

2. SOCIAL TAGGING
Tagging was pioneered on the World Wide Web by the social bookmarking site Delicious (www.delicious.com) as a way of helping users organize and classify bookmarks [6]. On Delicious, users can assign descriptive words, or tags, to their bookmarks. Tagging has been subsequently adapted by numerous websites as a way for users to describe, track, search for, and rate web objects [5]. The terms *social tagging* and *collaborative tagging* are used to describe systems in which users are able see tags assigned by others and, in many cases, can tag items posted by others. Since some systems do allow tags to be made private and thus not visible to other users, the more general term *tagging* describes any system in which users can assign tags to objects, regardless of visibility to others.

The use of tagging has further expanded into a wide variety of other areas, including music classification, music information retrieval, and music recommendation ([3], [8], [9], [13]). These last examples are interesting to the cause of DMI evaluation since, in a sense, the problems are similar: both music and DMIs are difficult to effectively analyze

NIME'13, May 27-30, 2013, KAIST, Daejeon, Korea.

quantitatively, with their appeal and success relying on many qualitative factors that often are not immediately obvious. Tags assigned by listeners are used to analyze and classify music, ascertain its essential features, and recommend other songs to users. These are similar to the goals of evaluating DMIs. Thus, it stands to reason that a method like social tagging may be useful in evaluating DMIs. Another significant advantage of tagging is that it is versatile, allowing any of the stakeholders described by O'Modhrain in [11] to evaluate an instrument.

2.1 Tagging vocabularies

Participants in social tagging use a vocabulary of words when assigning tags to objects. The choice of the vocabulary system is an important consideration in designing a tagging system.

In a *closed vocabulary*, the tags that may be assigned to objects are defined in advance. Participants can only use words from the provided set of tags, and cannot use any other words nor add any to the system. This vocabulary system is easier to implement and simplifies analysis since the set of possible tags is finite and known in advance. The disadvantage of a closed vocabulary is that it constrains the creativity of the participants, preventing them from using their intuition to devise tags that may have been initially overlooked or (and perhaps incorrectly) rejected. Similarly, it also allows for the possibility of unintentional bias being introduced into the system by its designers.

In an *open vocabulary*, the users may assign to an object any labels they consider descriptive of the object, thus avoiding all of the disadvantages of a closed vocabulary. The primary disadvantage of an open vocabulary is the increased difficulty in parsing and analyzing the tags. As described in [8], labels in open vocabulary systems are "noisy" and therefore harder to interpret.

2.2 Tagging digital musical instruments

Given the promise of social tagging as a means of evaluating DMIs, it needs to be adapted to the task. Using the social tagging method proposed in this paper, participating individuals (evaluators) assign descriptive labels, or tags, to a DMI. Tags generally are not mutually exclusive, and thus an object (in this case, a DMI) will almost certainly have multiple tags associated with it. A DMI can be evaluated by analyzing the tags assigned to it by users. Certain tags may be associated with the likely future success of the DMI, facilitate comparisons with other DMIs (perhaps even in cases where comparisons between the two would not be immediately intuitive), or give the designer valuable insight into where improvements may be needed.

2.2.1 Proposed vocabulary

For the purposes of evaluating DMIs, a *mixed* or *mostly-closed* vocabulary is proposed. The disadvantages of closed vocabularies are greatest for very large vocabularies and for diverse, unpredictable sets of objects. While there are a multitude of DMIs and controllers in a vast variety of forms to study, the relevant terms used to describe them are likely to be very similar since all musicians have largely similar concerns regardless of which instrument they happen to be playing. Terms like "intuitive", "expressive", and "familiar", as well as terms such as "difficult", "awkward" and "confusing", are concepts that musicians universally understand and consider when evaluating a new instrument. Therefore, the scope of terms to be used as tags is likely to be relatively small. Bias is unlikely to be an issue for the same reasons. Given a relatively small vocabulary, the need for participants to learn the vocabulary is not an onerous task. Therefore, the advantages of an open vocabulary are not significant in this case and do not justify the additional complexity and difficulty associated with an open vocabulary and the resulting noisy tags [9].

Considering that this is a new approach, however, it is conceded that the vocabulary proposed here is likely to require revision. Therefore, the interface for the tagging system will allow users to propose new tags. For purposes of simplifying the user interface, these tags will not be visible to other users in the study described in Section 4. Instead, they will be used as input for further refining the vocabulary for successive evaluations.

Table 1 shows the tags included in the tagging vocabulary. As a rule, tags are not hierarchical and each tag is assigned equal importance [5]. For purposes relating to the study proposed in Section 4, however, the tags are divided into two broad categories: classificatory and descriptive.

Table 1. Initial tagging vocabulary

Classificatory	Descriptive	
Augmented Instrument	Affordable	Heavy/Bulky
Instrument-like	Awkward	Innovative
Alternate Instrument	Complex	Intuitive
Note controller	Confusing	Limiting
Parameter controller	Cool	Musical
Processor	Difficult	Novel
Synthesizer	Easy	Portable
Installation	Ergonomic	Primitive
	Esoteric	Revolutionary
	Expensive	Simple
	Expressive	Strange
	Familiar	Theatrical

2.2.2 Classificatory tags

The classificatory tags are used to classify the DMIs by class and function. These are objective labels upon which there should be general agreement as to which applies to an instrument. They are not used directly in the analysis, but rather may be useful in determining whether a participant understood the basic nature of the DMI. This is particularly useful for the study described in Section 4.

The first three correspond to the categories of DMIs described in [10]. They are *augmented instruments* (acoustic instruments to which additional sensors or controls have been added), *instrument-like* or *instrument-inspired* (DMIs which are based on acoustic instruments), and *alternate instruments* (instruments which have no apparent origin in acoustic instruments).

DMIs can perform one or more basic functions: *note control* (instructing another device to trigger an event), *synthesis*, *processing*, and *parameter control*. Note that these are often not mutually exclusive. For example, a typical keyboard synthesizer is both a note controller and a synthesizer; in addition, most can also serve as a parameter controller, and some can also act as a processor.

2.2.3 Descriptive tags

These tags form the core of the tagging system. Participants use these to tag the DMI with the words they believe best describe the instrument. These are the tags that will be used for evaluation. Participants are not given strict definitions of the words in the context of DMIs. The interpretation of the meaning of the words and their relationship to each DMI is up

to them. This keeps the tags simple and reduces the amount of instruction or training required, and makes the system more flexible at the expense of being less precise. The terms used are predominantly qualitative in nature, and are believed to have generally accepted meanings in this context.

3. TAG ANALYSIS

After a sufficient number of evaluations have been collected for each instrument, the next task is to glean useful insight from them. As a general matter, there are two ideal outputs from the methodology: (1) a *score* (or set of scores) that can be used for quantitative analysis and comparisons, and (2) the identification of a set of *properties* (i.e., tags) shared by successful designs. This section discusses two proposed methods of achieving those outcomes. Other methods, including correlation with a "ground truth" score such as the one generated in the historical analysis in Section 4, and machine learning algorithms such as a perceptron, will also be used, as shown in Figure 1.

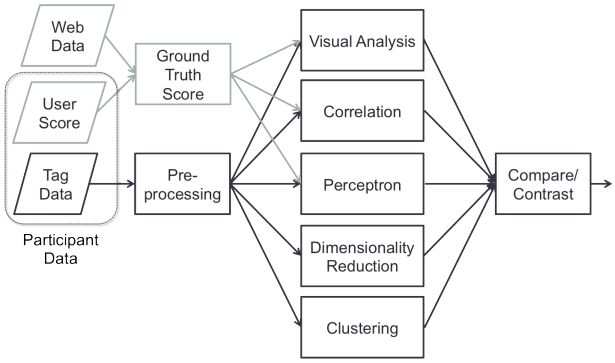

Figure 1: Tag Analysis Diagram

3.1 Dimension reduction

It is obvious from inspecting the list of proposed tags that there is likely to be a large number of data points for each instrument. Furthermore, it is expected that this number will increase as more tags are proposed by participants and added to the vocabulary. Looking more closely at the vocabulary, one may suspect that there is likely to be correlation between some tags. Therefore, dimension reduction methods such as principal components analysis (PCA) and self-organizing feature maps (SOFM) may be useful in the analysis by reducing the dataset, eliminating redundancies, and elucidating relationships between tags.

A secondary but perhaps more useful outcome of using dimensionality reduction methods is that by reducing the dimensions, it may be possible to discover a small number of dimensions on which DMIs can be scored. These scores can in turn be used for quantitatively comparing DMIs.

3.2 Cluster analysis

It seems intuitive that there likely are certain traits (and therefore certain tags) shared by successful DMI designs. Use of clustering algorithms such as K-means can be used to group tags together, which may further elucidate relationships between tags. It may also identify commonalities shared by successful instruments – for example, a set (or sets) of positive traits linked with success (and perhaps as well some negative ones that may be indicators of the opposite), especially when used in conjunction with the dimension reduction analysis.

4. EVALUATING THE EVALUATION

To determine if the evaluation method proposed in this paper is valuable, we conducted a study in which a 'ground truth' score for each instrument was calculated. The tagging results were then compared to the ground truth scores. A total of 35 subjects participated in the study. The majority (25 out of 35) of the subjects were between the ages of 18 and 21, five were aged 22 to 29, three were aged 30 to 39, and two were over age 40. All were musicians. A total of 21 instruments were evaluated.

4.1 Evaluation System

Ideally, participants would have performed the tagging after performing with the DMI or witnessing a live performance. Then, the long-term success of the instrument could be compared to the prediction based on the tagging data. This would provide a solid 'ground truth' against which the method could be judged. However, logistical and financial issues made direct, hands-on evaluations with a significant number of instruments impossible. Moreover, even if such trials could have been arranged for new instruments, it would be some time before the success of the instruments, and therefore that of the method itself, could be evaluated. Therefore, a historical study was conducted. In this study, participants observed pre-recorded video performances using various DMIs. After viewing each performance, the participants were asked to assign tags to the DMI based on their observations during the performance (see Figure 2). While this removes the participants from directly interacting with the instrument, it offers one significant benefit: the results of their tagging can be compared against the historical performance of the DMI.

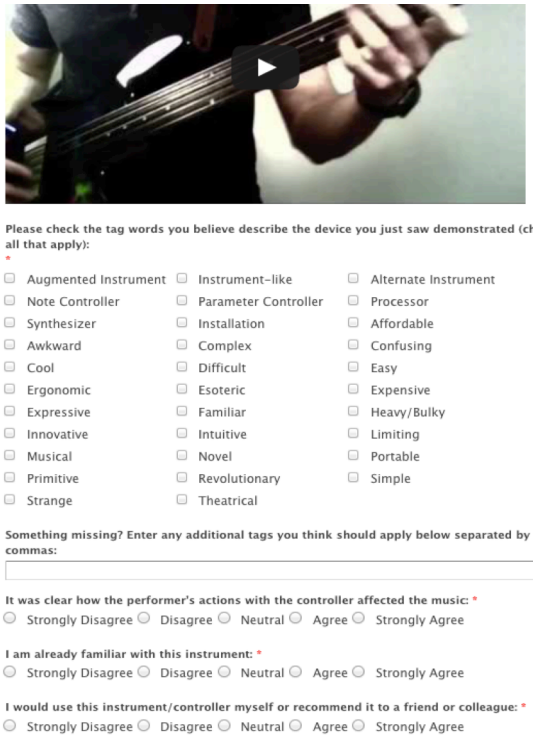

Figure 2. Example tagging questionnaire

By comparing the results of the tag analysis to the historical data, an attempt to evaluate the method itself can be made. This historical data took several forms. For commercially available products, the historical analysis could be as straightforward as examining sales volumes. However, most DMI manufacturers do not publicize sales volume information for individual products. The data must also be indirectly gathered for devices proposed in academia since the number of devices produced is often very small. Therefore, the data was drawn from documented performances, tags on social websites, and web search results. A score was generated and contrasted with "scores" generated from the data analysis of the instrument

tagging evaluation. Users were also asked to rate on a Likert scale the statement: *I would use this instrument/controller myself or recommend it to a friend or colleague.* This provided an additional reference against which to compare the tagging results.

4.2 Preliminary Results

A first-order, visual analysis of the tag data shows that there is correlation between certain tags and the instruments ground truth score. Preliminary results with PCA show that instruments that received a high ground truth score in the historical analysis (see Section 4.1) received higher scores in the first principal component (see Section 3.1). Instruments that received a low ground truth score received lower values in the first principal component. Instruments that received high ground truth scores also cluster together (see Section 3.2), while instruments that had low ground truth scores also cluster together. These suggest that 'successful' instruments share some common traits that can be determined from the tag data. Preliminary data also show a high correlation between certain tags and the ranking of the instruments by the participants.

One concern with this approach is that participants are not able to directly use or observe the use of the instruments since their understanding of the instruments is limited to the information provided in the videos. Users were asked to self-report their understanding of the instrument. It was intended that the classificatory tags would allow for an additional, objective determination of the participants' understanding of the instruments. However, this approach has been found to be problematic since the classificatory tags are not as unambiguous and objective as initially believed. In particular, it was found that the line between a DMI being instrument-like and an alternate instrument is a blurry one in some cases.

5. FUTURE WORK

The evaluation system proposed here is a proof-of-concept prototype, implemented as part of a larger ongoing project in DMI evaluation. Considerable room for enhancing and testing the system remains. First and foremost, the system should be deployed in some systematic manner that allows participants to tag a DMI after performing with it or while observing a performance using it, in sufficient quantity to provide a large data set. Ideally, the system would be deployed at a major conference or trade show featuring DMIs, such as NIME or NAMM, allowing attendees and performers to evaluate the instruments they encounter. The results could then be returned to the designers for consideration. In the long term, the results could be compared with actual outcomes to determine how the tagging system compares with other evaluation methods.

Second, the vocabulary needs further refinement. The choice of a mostly-closed vocabulary was made to make its implementation more manageable during the early stages of development. The system can be considered collaborative in the sense that the vocabulary is allowed to expand between successive generations of the system, allowing future participants an improved, more specific vocabulary. A more mature system could not only allow users to suggest additional tags but allow those to be immediately visible to other users. The results of such an open-vocabulary system could then be compared to the more restricted vocabulary proposed here to determine which offers a better compromise between complexity and accuracy.

6. CONCLUSION

We proposed a new method of evaluating digital musical instruments using concepts from social tagging developed for the World Wide Web. The method has many advantages: the gathering of data is quick and simple, it allows for qualitative analysis, and any stakeholder may use the system to evaluate a DMI. Ongoing work will study the effectiveness of the method, and future work will enhance and improve it.

7. REFERENCES

[1] Barbosa, J., Calegario, F., Teichrieb V., Ramalho, G., and Mcglynn, P. Considering audience's view towards an evaluation methodology for digital musical instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME '12)* (Ann Arbor, MI, USA, May 21-23, 2012), 403–408.

[2] Bellotti, V., Back, M., Edwards, W. K., Grinter, R. E., Henderson, A., and Lopes, C. Making sense of sensing systems: five questions for designers and researchers. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '02)* (Minneapolis, MN, USA, April 20-25, 2002). ACM Press, New York, NY, 2002.

[3] Chen, L., Wright, P., and Nejdl, W. Improving music genre classification using collaborative tagging data categories and subject descriptors. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, (Barcelona, Spain, February 9-11, 2009). ACM Press, New York, NY, 2009, 84-93.

[4] Cook, P. Principles for designing computer music controllers. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME '01)* (Seattle, WA, USA, April 1-2, 2001), 3–6.

[5] Golder, S. A. and Huberman, B. A. The structure of collaborative tagging systems. In *Journal of Information Science*, 32, 2, 2006, 198–208.

[6] Gupta, M., Li, R., Yin, Z., & Han, J. An overview of social tagging and applications. *Social Network Data Analytics* (2011), 447-497.

[7] Johnston, A. Beyond evaluation: linking practice and theory in new musical interface design. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME '11)* (Oslo, Norway, May 30 – June 1, 2011), 280-283.

[8] Law, E., Settles, B., and Mitchell, T. Learning to tag from open vocabulary labels. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2010)* (Barcelona, Spain, September 2010) 211-226.

[9] Law, E., Settles, B., and Mitchell, T. Learning to tag using noisy labels. *European Conference on Machine Learning*, 2010.

[10] Miranda, E., Wanderley, M. *New Digital Instruments: Control and Interaction Beyond the Keyboard*, A-R Editions, Inc., Middleton, WI, USA, 2006.

[11] O'Modhrain, S. A framework for the evaluation of digital musical instruments. *Computer Music Journal*, 35, 1, 2011, 28–42.

[12] Stowell, D., Plumbley, M. D., and Bryan-Kinns, N. Discourse analysis evaluation method for expressive musical interfaces. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME '08)* (Genova, Italy, June 5-7, 2008), 81–86.

[13] Turnbull, D. Indexing music with tags. In *Music Data Mining*, 2011, pp. 276–302.

[14] Wanderley, M. and Orio, N. Evaluation of input devices for musical expression: borrowing tools from HCI. *Computer Music Journal*, 26, 3, 2002, 62–76