ALGORITHMIC ABILITY TO PREDICT THE MUSICAL FUTURE: DATASETS AND EVALUATION

Berit Janssen[1] **Tom Collins**[2,3] **Iris Yuping Ren**[4]
[1] Digital Humanities Lab, Department of Humanities, Utrecht University, The Netherlands
[2] Music, Science and Technology Research Cluster, Department of Music, University of York, UK
[3] Music Artificial Intelligence Algorithms, Inc., Davis, CA, USA
[4] Department of Information and Computing Sciences, Utrecht University, The Netherlands
{berit.janssen@gmail.com, tomthecollins@gmail.com, y.ren@uu.nl}

ABSTRACT

Music prediction and generation have been of recurring interest in the field of music informatics: many models that emulate listeners' musical expectancies, or that produce novel musical content have been introduced over the past few decades. So far, these models have mostly been evaluated in isolation, following diverse evaluation strategies. Our paper provides an overview of the new MIREX task *Patterns for Prediction*. We introduce a dataset, which contains monophonic and polyphonic data, both in symbolic and audio representations. We suggest a standardized evaluation procedure to compare algorithmic musical predictions. We compare two neural network models to a baseline model and show that algorithmic approaches can correctly predict about a third of a monophonic segment, and around half of a polyphonic segment, with one of the neural network models achieving best results. However, other approaches to algorithmic music prediction are needed to achieve a more rounded picture of the potential of state-of-the-art methods of music prediction.

1. INTRODUCTION

Prediction of future events is fundamental to human and artificial intelligence, and has therefore been discussed as a core research interest bridging cognitive psychology and machine learning [5]. Music, with its complex event sequences extending over time, provides an excellent setting for the study of prediction.

In music cognition, human prediction of future events, or *expectation*, is studied from theoretic, behavioral, imaging, and modeling perspectives. Some theoretical work [2, 15] distinguishes between *veridical*, *schematic*, and *dynamic* expectations: veridical expectations occur due to familiarity with a musical piece, schematic expectations are elicited by familiarity with a genre, and dynamic expectations manifest in-the-moment predictions, when, consciously or subconsciously, a listener becomes attuned to a pattern in a novel piece. It has been claimed that the pleasure we derive from music resides in the tension between these three forms of expectation [15].

While the full complexity of expectation in music may still be hard to capture in computational models, the goal of this paper is to give an overview of how computational models may emulate human expectations through prediction of future musical events, and how we should evaluate such models.

Our main contributions are as follows: first, we review different approaches to modelling expectation in music. Second, we introduce a dataset on which such models can be trained and evaluated. Third, we propose two evaluation tasks and associated evaluation measures. [1] Fourth, we provide the results of a baseline and two more complex models on the tasks. Finally, we discuss findings and recommendations for future model development and evaluation.

2. RELATED WORK

2.1 Approaches to music prediction

2.1.1 Markov models

Markov models have been influential in music prediction: statistics on transitions between musical events may be used to generate predictions for unseen musical events. Musical events may be represented in various ways, such as pitch, duration, onset, metric weight, and so on. Therefore, it has been suggested to build distinct models for different combinations of music representations [11]. Markov models trained on music corpora may very well serve to model schematic expectation, such as a leading tone to be followed by an octave. Dynamic (i.e., work-specific) expectations may also be modelled through a Markovian approach through training on a musical piece itself, where the model is incrementally updated as the piece progresses. The question of how to combine models of various music representations, or how to combine models trained on corpora ("long-term models") and models trained incremen-

[1] The corresponding MIREX task description, datasets, and evaluation code can be found at https://tinyurl.com/y455cf97

tally on a piece ("short-term models") has been experimentally investigated [26], but on specific styles, which might not generalize to other musical genres.

Even though some models can theoretically extend over very long contexts, the question remains whether Markovian models, which are by nature "forgetful", will capture longer structure that may facilitate precise predictions, such as repetition of themes in a Classical piece of music, or the return of the theme in a jazz performance. [36]

2.1.2 Neural networks

Over the last two decades, interest in neural network models for music prediction has been increasing. The first attempts in this direction made use of recurrent neural networks (RNNs) [23], with an input, hidden and output layer, which predict future states of sequential input data. Variants of RNNs, such as long short-term memory models (LSTM), have also been applied to music [13]. Various extensions of such models have been presented since [4, 14].

Another class of neural networks, convolutional neural networks (CNN), is usually used for image data. A musical composition may also be thought of as an image rendered in time-pitch or time-amplitude space. Some authors therefore applied CNNs to piano roll representations of symbolic music [12], or to audio [34] for music prediction.

While these and other neural network architectures have resulted in generation and prediction of music, the output of the models in itself is often not easy to predict. One of the challenges for neural network models, as for Markovian models, is the degree to which they can capture large-scale structure in a musical piece, and recreate dynamic expectations that may arise within a piece in itself.

2.1.3 Pattern discovery

Yet another approach to predicting the musical future is to search for repeating patterns within the piece. This approach emulates dynamic expectations of a listener (patterns occur in earlier parts of a piece, leading to predictions based on later, partially complete occurrences of the same patterns [6, 26]), but less so schematic expectations.

Various algorithms have been proposed to discover repeated patterns within a piece [7, 18, 22, 28, 30, 35], which differ in the kinds of patterns they aim to discover, in the way music is represented, and in the algorithms used to find repetitions [16]. These algorithms have been tested on benchmarks of annotated patterns, while evaluation by prediction is suggested but yet-to-be implemented [21].

2.2 Evaluating music prediction

To ascertain how models compare to human expectations, various approaches have been used: some of them fall in the domain of music generation, while others fall in the domain of information theoretic measures.

2.2.1 Information-theoretic measures

In order to investigate musical predictions of a model with information theoretic measures, the model is trained on a corpus or a corpus subset, then exposed to a novel musical piece. For each musical event, the likelihood of that event according to the model is measured. Alternatively, the uncertainty of the model after each note in predicting the following melody note may be recorded. Ratings of likelihood or uncertainty may then be compared to human ratings from experimental research.

To compare likelihood as rated by a model to human ratings, priming experiments may serve as an evaluation ground: in such experiments, participants had to give an indication as to how well, given a melodic context, a note fitted their expectations [17, 31]. There were also experiments on uncertainty, in which participants were asked to indicate how uncertain they were of what might follow each note in a Bach chorale [19]. As phrase boundaries often coincide with points of greater uncertainty, human segmentations have also been occasionally used as a ground truth for evaluating model predictions [27].

2.2.2 Music generation

A very common way to evaluate predictions from a model is the demonstration of music generated by a given model (e.g., [11]). While this is informative, it is not self-evident how to judge the quality of such an output. Music practitioners of a given genre may be asked as judges [32], but aesthetic judgments alone may not reveal much about a model's shortcomings [1].

Another approach to evaluating music generation is through comparing generated music with human-composed melodies. Human compositions can then be used as the touchstone of how well a model captures structure and style [20]. We choose this approach by providing models with a short piece of music, or prime, and instruct the models to generate a continuation, which we evaluate against the true continuation. Moreover, we test how successfully a model distinguishes between the true continuation and an artificial, foil continuation.

3. A DATASET FOR MUSIC PREDICTIONS

3.1 Dataset construction

We provide small, medium, and large development datasets (100, $1,000$, and $10,000$ pieces, respectively). This caters to different approaches to designing models for the task, some of which are more data-intensive than others. Each dataset has audio/symbolic and monophonic/polyphonic variants.

Pieces were selected at random from the Lakh MIDI Dataset (LMD) [29] with the aim of creating primes lasting $\approx 35\,\mathrm{sec}$ according to tempo information. True continuations were selected – and foil continuations generated – such that onsets (note start times) occurred in a 10 quarter-note-beat window.

We also constructed a test dataset from another corpus of MIDI files, which is similar in nature to LMD. The test dataset also contains audio/symbolic and monophonic/polyphonic variants, and provides primes, true and foil continuations. In keeping with the MIREX guidelines,

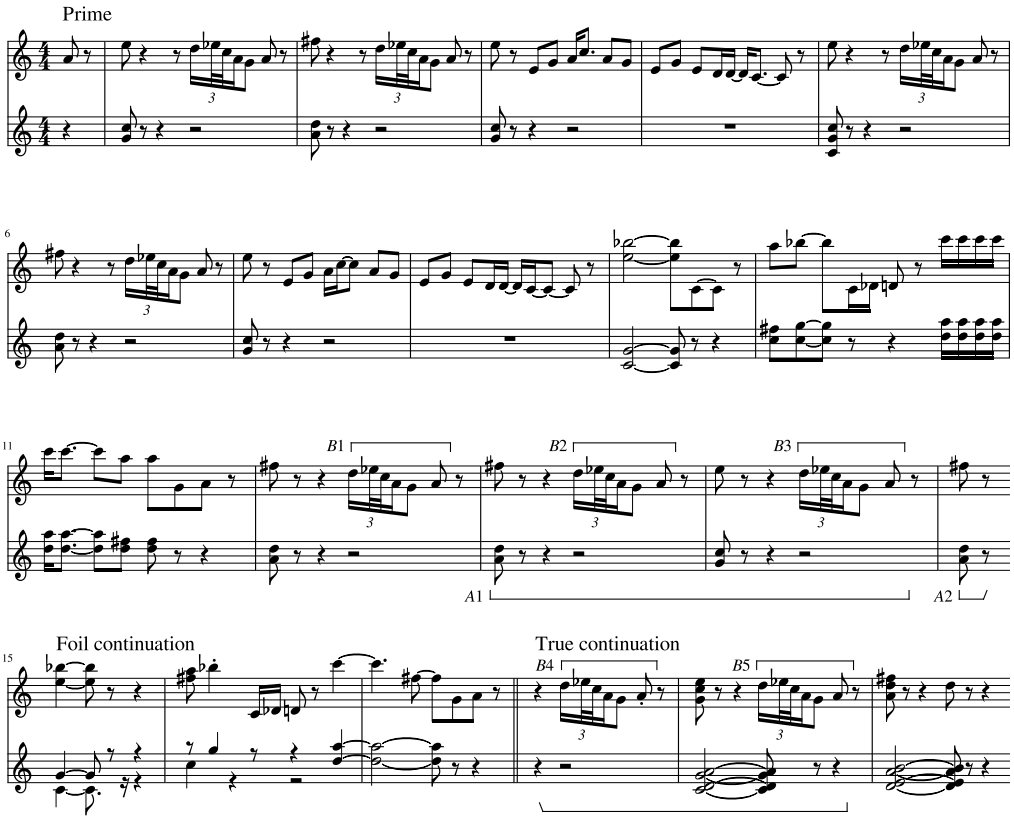

Figure 1. Item from the large, polyphonic variant of the development dataset. Musical provenance unknown.

we will not reveal details about the provenance of the test dataset. What we can say is that the source for the test dataset contains approximately $30,000$ files, and that both this source and LMD are gathered from sites that represent musical interests and tastes, broadly construed. The number of items in the test source tagged as "pop", for instance, is $9,165$. Other items have similar tags, however, such as "latin pop". To our knowledge, no such analysis of genres exists for the LMD, so remarks about the overlap of genres and content between our training and test datasets are necessarily speculative.

MIDI files were selected at random, imported using midi-convert, [2] quantized, and then the following criteria were applied when generating monophonic primes and continuations:

1. A prime had to contain at least 20 notes;
2. The maximum inter-onset interval in a prime could not exceed 8 quarter-note beats;
3. A continuation had to contain at least 10 notes;
4. The channel from which material was selected had to be suitably monophonic prior to *skylining* (see below), meaning at least 80% of minimal segments [25] had to be single notes or rests.

Skylining means to select the highest-sounding notes at each onset and return only those notes, perhaps with modified durations, so that the output is truly monophonic. The rationale for only skylining material that was already 80% or more monophonic is that skylining inherently polyphonic material often results in odd-sounding or implied-polyphonic output. For polyphonic dataset generation, criteria (1)-(3) were the same, but a replacement for (4) was needed because parsed MIDI files sometimes contain erroneously long notes. In place of criterion (4), polyphonic dataset generation involved clipping any notes longer than 8 quarter-note beats.

If a prime or continuation did not satisfy one or more of the above criteria, generation proceeded to the next randomly selected piece (rather than, say, selecting a different excerpt from the same piece). Approximately $1/6$ random selections passed the criteria, meaning we had to process $6N$ pieces to produce a dataset of size N.

Our baseline for generating foil continuations is the Racchman-Jun2015 model [9] described in section 4. Since previous work has emphasized the need to progress from Markovian approaches to modeling music [36], this seemed to be the most appropriate baseline. Examples of a prime, true and foil continuations are shown in Figure 1. This excerpt came from an LMD song called "Dirtyluv". We were unable to identify further title or artist information – an issue when working with this source.

Returning to the discussion of pattern discovery for prediction, the example is annotated with pattern occurrences $A_1, A_2, B_1, B_2, \ldots, B_5$ to indicate how such an approach would be fruitful in this case. The annotating lines are placed above and below the stave for clarity, but encompass notes from both staves that begin in the indicated time spans. The prime ends by repeating the first 3 notes of A_1.

[2] `https://www.npmjs.com/package/@pioug/MidiConvert`

Therefore, one reasonable prediction for the continuation is that it will proceed to restate the remainder of A_1. Comparing such a prediction to the true continuation, we see that it would be quite successful – some extra notes in the left hand in measure 16 are the only difference between A_1 and A_2. In an analogous fashion, the regularity of occurrences of B_1, B_2, B_3 could be used to make a prediction about B_4 and B_5 appearing in the true continuation.

3.2 Dataset characteristics

Figure 2 contains three violin plots showing basic distributional characteristics of the symbolic, monophonic, medium-size development ("dev") dataset and the symbolic, monophonic test dataset. Inspection of these plots suggests that the development and test datasets share similar characteristics. A slightly more marked peak can be seen around inter-onset 0.5 in the primes and true continuations of the test compared to the development datasets (Figure 2A), and the test dataset has a slightly lower mean MIDI note number than that of the development dataset (Figure 2C). While the test dataset is separate from the development dataset (LMD) and it would be inappropriate to report the extent to which they overlap in terms of content, evidently their distributional characteristics are similar. It is worth noting that there is healthy representation of "bass lines" in monophonic – but not polyphonic – variants of the datasets (see the modal concentration around MIDI note 35 in Figure 2C), as a result of the selection criteria outlined above.

4. COMPARED MODELS FOR MUSIC PREDICTION

We compare the output of foil continuations by the first-order Markov model (see section 3.1), in the following referred to as *baseline* model, with two recurrent neural network models, *BachProb* [10] and *Seq2SeqP4P* [24].

BachProb is a deep-gated, recurrent neural network with three consecutive layers. Notes are represented as triplets of pitch, duration and inter-onset interval with respect to the previous note. Durations and inter-onset intervals are rounded to durations commonly found in musical scores. *BachProb* is trained on the development dataset using truncated back propagation, with separate models for the monophonic and the polyphonic parts of the dataset.

Seq2SeqP4P is a long short-term memory network with two layers. Music was represented as the MIDI commands *note-on*, *note-off*, which define when a given pitch starts or ends, and time shifts between those commands, quantized to 12 subdivisions per beat. Such a sequence of MIDI commands and time shifts was used as the input to training the model on the development dataset. By virtue of design, Seq2SeqP4P was trained only on the monophonic part of the dataset.

The baseline model consists of a first-order Markovian generator nested in other processes intended to ensure the output has long-term repetitive and phrasal structure [9]. The state space consists of beat of the mea-

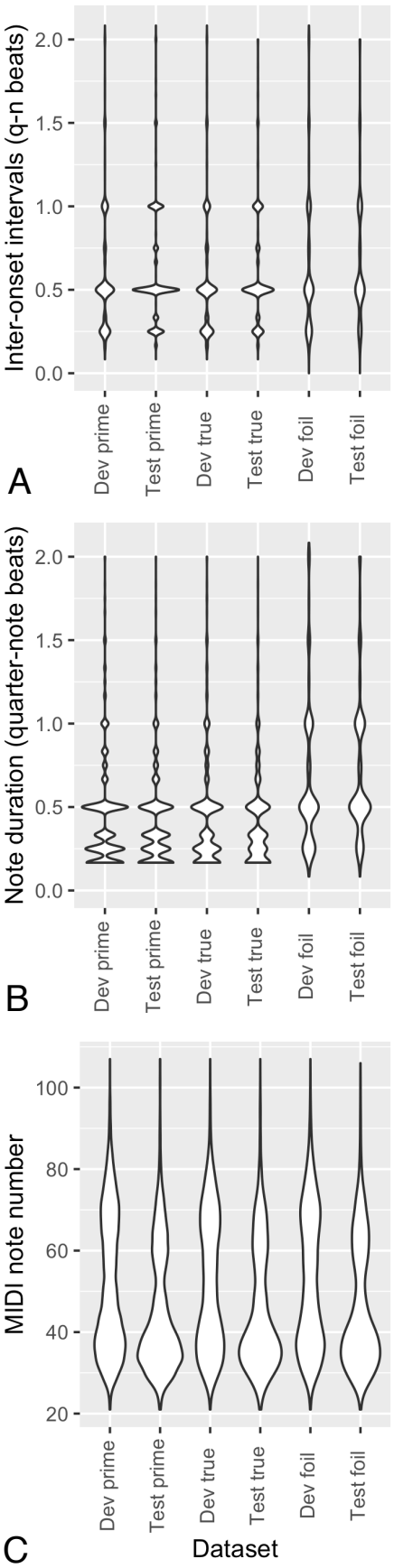

Figure 2. Characteristics of the symbolic, monophonic, medium-size dataset: (A) Inter-onset interval distributions of development and test datasets; (B) Duration distributions; (C) MIDI note number distributions.

sure on which notes occur, and MIDI note numbers relative to tonal center. The Markov generator alone, called Racchman-Jun2015, is a useful benchmark, because any longer structures that emerge here do so by chance.

5. EVALUATION

We evaluate music prediction in two ways:

- Explicit task. Models are provided with a prime, from which they generate continuations. The output of the models is then judged according to how many events of the true continuation they correctly predicted (metric definitions below).
- Implicit task. Models are provided with correct and foil continuations after a prime, from which they have to select the correct continuation.

5.1 Explicit task

For the explicit task, the evaluation proceeds as follows: within a time interval of ten quarter notes, we step through the time line by small time increments. We choose a time increment of $t = 0.5$ quarter notes, i.e., an eighth note.

We represent each note in the true and algorithmic continuation as a point in a two-dimensional space of onset and pitch, giving the point-set $\mathbf{P}$ for the true continuation, and $\mathbf{Q}$ for the algorithmic continuation. We calculate differences between all points p_i in $\mathbf{P}$ and q_j in $\mathbf{Q}$, which represent the translation vectors $\mathbf{T}$ to transform a given algorithmically generated note into a note from the true continuation [8, 33].

We then search for the largest set match achievable through translation with any vector, leading us to the number of correctly predicted notes cp:

$$\text{cp}(\mathbf{P}, \mathbf{Q}) = \max_{\mathbf{T}} |\{q_j | q_j \in \mathbf{Q} \wedge q_j + \mathbf{T} \in \mathbf{P}\}| \quad (1)$$

We define recall as the number of correctly predicted notes, divided by the cardinality of the true continuation point set $\mathbf{P}$. Since there exists at least one point in $\mathbf{Q}$ which can be translated by any vector to a point in $\mathbf{P}$, we subtract 1 from numerator and denominator to scale to $[0, 1]$.

$$\text{Rec} = (\text{cp}(\mathbf{P}, \mathbf{Q}) - 1)/(|\mathbf{P}| - 1) \quad (2)$$

Precision is the number of correctly predicted notes, divided by the cardinality of the point set of the algorithmic continuation $\mathbf{Q}$, scaled in the same way:

$$\text{Prec} = (\text{cp}(\mathbf{P}, \mathbf{Q}) - 1)/(|\mathbf{Q}| - 1) \quad (3)$$

The F_1-score is the harmonic mean of precision and recall. As the measures we propose are not defined for cases in which either the true or the algorithmic continuation contain fewer than two events, we start evaluation from onset 2.0, i.e, two quarter notes after the end of the prime, which ensures long enough sequences.

5.2 Implicit task

In the implicit task, a prediction model has to judge which of two continuations after a given prime is the true continuation. The foil is generated by the baseline model (see Section 4).

To evaluate the implicit task, we measure the success rate, i.e., the number of cases in which the model correctly picks the true continuations, divided by the total amount of decisions undertaken by the model.

6. RESULTS

Our evaluation and results focus on the symbolic variants of our test dataset. For the monophonic part of the dataset, we compare all three models, whereas for the polyphonic part, we only compare *BachProb* against the baseline, as *Seq2SeqP4P* has not been trained on polyphonic data yet.

6.1 Explicit task

For the monophonic dataset, the various models predict around a third of the events correctly, with *BachProb* outperforming the baseline and *Seq2SeqP4P* slightly, especially at the start of the predicted continuation (see Figure 3B).

Seq2SeqP4P predicts more of the notes in the true continuation correctly than the baseline, but at the cost of precision (Figure 3A), which means that its F_1 score is also lower overall than the baseline (Figure 3C).

For the polyphonic dataset, *BachProb* achieves a much higher recall than the baseline Markov model (Figure 3E). In precision, it performs close to the baseline, which results in very similar F1-scores, too (Figure 3D, F).

In general, the recall, precision, and F_1 score of the models decrease as the onset of the generated continuation increases, even though the baseline model has fairly stable performance over the evaluated time interval for the monophonic dataset, and *Seq2SeqP4P* increases in performance at the start of the continuation.

6.2 Implicit task

BachProb achieves a success rate of 0.85 for the monophonic continuations, i.e., 85% of the true continuations were identified correctly. For the polyphonic continuations, *BachProb* scores a success rate of 0.90. According to the binomial distribution, a success rate of 0.54 or higher constitutes above-chance performance on this task. At present, *Seq2SeqP4P* has not been implemented for the implicit task, so there are no results for it at this stage.

7. DISCUSSION

How events in the recent or more distant past may be appraised – consciously or otherwise – so as to be better adapted for what lies ahead is a phenomenon that has intrigued researchers from diverse disciplines such as cognitive psychology, philosophy, computer science, and music. In this paper, we focussed on music as a vehicle for studying the ability of computational models to predict continuations of given primes, and described datasets, evaluation procedures, and results to this end.

BachProb, utilizing a gated recurring neural network, outperforms the other two models. *Seq2SeqP4P*, based

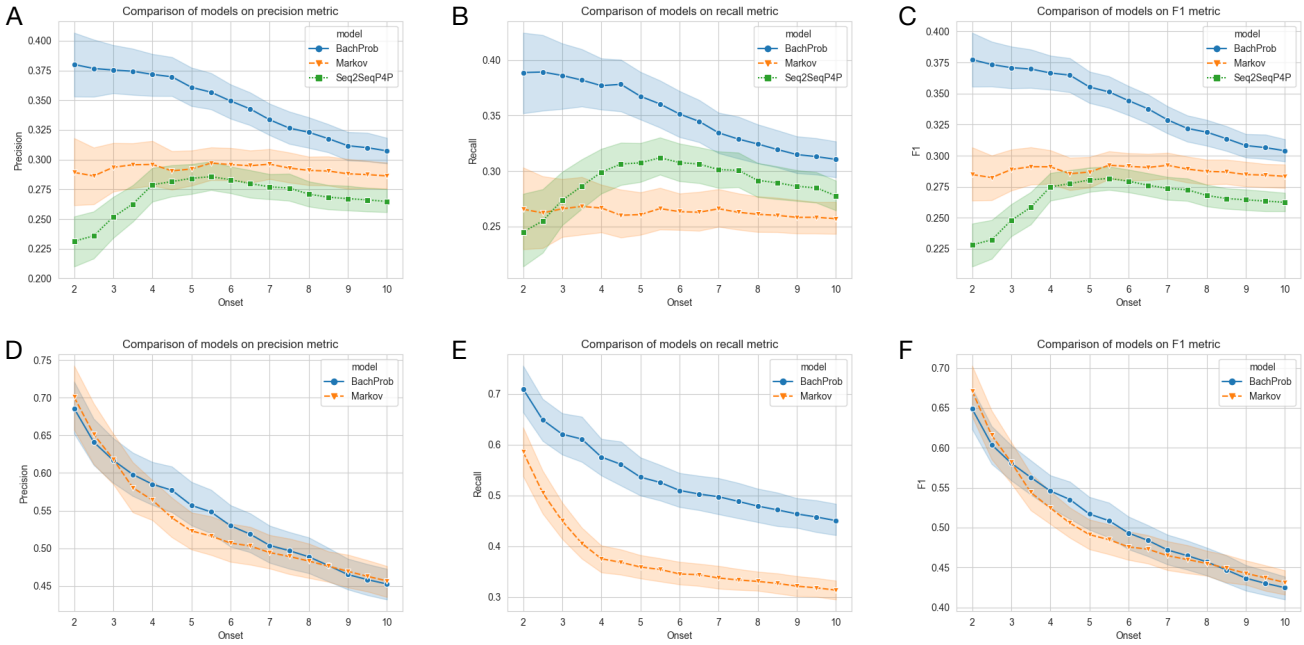

Figure 3. (A) Precision, (B) recall, and (C) F_1-measure of the three monophonic prediction models; (D) precision, (E) recall, and (F) F_1-measure of the two polyphonic prediction models. Evaluation measures are not defined for very short sequences, so evaluation starts at onset 2.0. Shading around the curves indicates one standard error from the mean.

on a long-short term memory model, predicts less musical material correctly. Arguably, music representation may be a larger factor in this than network architecture: as the model reportedly produces repeated pitches in some cases, and tends to always assign the same durations between the note-on and note-off events [24], the sequences of decoupled pitch and duration related information may not be suitable for the model to learn musical structure. Repeated short durations in the output may also explain the high recall and low precision of the model.

The overall higher recall, precision and F1-score of *BachProb* and the baseline on the polyphonic dataset, as compared to the monophonic results, is surprising. A possible explanation may be that repetitive chords, for instance in a piano or guitar part, are frequently present in the dataset and may be relatively easy to predict. The comparatively lower precision of *BachProb* suggests that while many notes from the true continuation are generated, there are also many spuriously generated notes.

For the implicit task, it is remarkable that *BachProb* distinguishes between true and foil continuations highly above chance. While the explicit task shows that the Markovian continuation reproduces a modest percentage of notes in the true continuations, the implicit task shows that *BachProb* learned details of the musical structure which could not be emulated by the Markovian foil.

We hope to evaluate more models for music prediction in the future, which might give us more insights into what constitutes successful prediction. As such, our proposal of a dataset and evaluation measures opens up the ground to discussion of how comparison of music prediction models may be improved.

First, we need to consider improved, or additional evaluation measures for the explicit task: our current approach to evaluating the explicit task entails that algorithmic continuations will be evaluated as correct continuations even if they are shifted in onset or pitch. The proposed measures may also penalize deviations from a true continuation that might be almost imperceptible to a human listener, such as an added chord note, or the reordering of chord tones.

Second, the evaluation of the implicit task also needs to be reconsidered: it depends heavily on the quality of the foil continuation. Perhaps the Markov baseline generates material which is too easily distinguishable from the true continuation. Moreover, we measure success rate, which has the advantage of easy interpretation, but does not take into account a model's confidence in its distinction between the true and foil continuation. Alternative foil continuations, or more fine-grained measures of the models' distinctions, would certainly give additional insights on model performance.

Third, additional tasks and datasets may be needed. We envision bringing together outcomes of music prediction models with evidence on human expectations in music. The continued systematic comparison of various models for music prediction can teach us much about the successes and shortcomings of prediction models in relation to each other, as well as about the influence of music representation and model parameters. Studies which measure human responses on their levels of surprise when hearing the continuation of a musical prime [19], or studies which ask humans to improvise a continuation [3] may inform improved tasks and evaluation strategies, and underpin models to predict the musical future.

8. ACKNOWLEDGMENTS

We thank Anja Volk and James Owers for sharing their insights and comments throughout our research.

9. REFERENCES

[1] Christopher Ariza. The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems. *Computer Music Journal*, 33(2):48–70, 2009.

[2] Jamshed J. Bharucha. Tonality and Expectation. In Rita Aiello and John A. Sloboda, editors, *Musical Perceptions*, pages 213–239. Oxford University Press, 1994.

[3] James C. Carlsen, Pierre I. Divenyi, and Jack A. Taylor. A Preliminary Study of Perceptual Expectancy in Melodic Configurations. *Bulletin of the Council for Research in Music Education*, pages 4–12, 1970.

[4] Srikanth Cherla, Tillman Weyde, Artur Garcez, and Marcus Pearce. A Distributed Model For Multiple-Viewpoint Melodic Prediction. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 15–20, 2013.

[5] Andy Clark. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *The Behavioral and Brain Sciences*, 36:181–204, 2013.

[6] John G. Cleary and Ian H. Witten. Data Compression Using Adaptive Coding and Partial String Matching. *IEEE Transactions on Communcations*, COM-32(4):396–402, 1984.

[7] Tom Collins, Andreas Arzt, Sebastian Flossmann, and Gerhard Widmer. SIARCT-CFP: Improving Precision and the Discovery of Inexact Musical Patterns in Point-Set Representations. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 549–554, 2013.

[8] Tom Collins, Sebastian Böck, Florian Krebs, and Gerhard Widmer. Bridging the Audio-Symbolic Gap: The Discovery of Repeated Note Content Directly From Polyphonic Music Audio. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, pages 1–12, 2014.

[9] Tom Collins and Robin Laney. Computer-generated Stylistic Compositions with Long-term Repetitive and Phrasal Structure. *Journal of Creative Music Systems*, 1(2), 2017.

[10] Florian Colombo. Generating and Discriminating Symbolic Music Continuations with Bach-Prob. `https://www.music-ir.org/mirex/wiki/2018:Patterns_for_Prediction_Results`, 2018. Accessed: 2019-04-09.

[11] Darrell Conklin and Ian H. Witten. Multiple Viewpoint Systems for Music Prediction. *Journal of New Music Research*, 24(1):51–73, 1995.

[12] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. MuseGAN: Demonstration of a Convolutional GAN Based Model for Generating Multi-Track Piano-Rolls. In *Late Breaking/Demos, 18th International Society for Music Information Retrieval Conference*, 2017.

[13] Douglas Eck and Juergen Schmidhuber. Finding Temporal Structure in Music: Blues Improvisation with LSTM Recurrent Networks. In *Proceedings of the 12th IEEE workshop on neural networks for signal processing*, pages 747–756. IEEE, 2002.

[14] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music Transformer, 2018.

[15] David Huron. *Sweet Anticipation. Music and the Psychology of Expectation*. MIT Press, Cambridge, Massachusetts, 2007.

[16] Berit Janssen, W. Bas de Haas, Anja Volk, and Peter van Kranenburg. Finding Repeated Patterns in Music: State of Knowledge, Challenges, Perspectives. In M. Aramaki, editor, *10th International Symposium, CMMR 2013, Revised Selected Papers*, number 8905, pages 277–297. 2014.

[17] Carol L. Krumhansl and Edward J. Kessler. Tracing the Dynamic Changes in Perceived Tonal Organization in a Spatial Representation of Musical Keys. *Psychological Review*, 89(4):334–368, 1982.

[18] Olivier Lartillot. In-depth Motivic Analysis Based on Multiparametric Closed Pattern and Cyclic Sequence Mining. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 361–366, 2014.

[19] Leonard C. Manzara, Ian H. Witten, and Mark James. On the Entropy of Music: An Experiment with Bach Chorale Melodies. *Leonardo Music Journal*, 2(1):81–88, 1992.

[20] Gabriele Medeot, Srikanth Cherla, Katerina Kosta, Matt McVicar, Samer Abdallah, Marco Selvi, Ed Newton-Rex, and Kevin Webster. StructureNet: Inducing Structure in Generated Melodies. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 725–731, 2018.

[21] David Meredith. COSIATEC and SIATECompress: Pattern Discovery by Geometric Compression. `https://www.music-ir.org/mirex/wiki/2013:Discovery_of_Repeated_Themes_\%26_Sections_Results`, 2013. Accessed: 2018-02-09.

[22] David Meredith, Kjell Lemström, and Geraint A. Wiggins. Algorithms for Discovering Repeated Patterns in

Multidimensional Representations of Polyphonic Music. *Journal of New Music Research*, 31(4):321–345, 2002.

[23] Michael C. Mozer. Neural Network Music Composition by Prediction: Exploring the Benefits of Psychoacoustic Constraints and Multi-Scale Processing. *Connection Science*, 6(2-3):247–280, 1994.

[24] Eric Nichols. Seq2SeqP4P: A Sequence-to-Sequence model for Monophonic Music Continuation. `https://www.music-ir.org/mirex/wiki/2018:Patterns_for_Prediction_Results`, 2018. Accessed: 2019-04-09.

[25] Bryan Pardo and William P. Birmingham. Algorithms for chordal analysis. *Computer Music Journal*, pages 27–49, 2002.

[26] Marcus Pearce and Geraint A. Wiggins. Improved Methods for Statistical Modelling of Monophonic Music. *Journal of New Music Research*, 33(4):367–385, dec 2004.

[27] Marcus T. Pearce, Daniel Müllensiefen, and Geraint A. Wiggins. The Role of Expectation and Probabilistic Learning in Auditory Boundary Perception: A Model Comparison. *Perception*, 39(10):1367–1391, 2010.

[28] Matevž Pesek, Aleš Leonardis, and Matija Marolt. A Compositional Hierarchical Model for Music Information Retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 131–136, 2014.

[29] Colin Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-Midi Alignment and Matching*. PhD thesis, Columbia University, 2016.

[30] Iris Yuping Ren. Closed Patterns in Folk Music and Other Genres. In *Proceedings of the 6th International Workshop on Folk Music Analysis*, 2016.

[31] E. Glenn Schellenberg. Expectancy in Melody: Tests of the Implication-Realization Model. *Cognition*, 58:75–125, 1996.

[32] Bob L. Sturm and Oded Ben-Tal. Taking the Models back to Music Practice: Evaluating Generative Transcription Models Built Using Deep Learning. *Journal of Creative Music Systems*, 2, 2017.

[33] Esko Ukkonen, Kjell Lemström, and Veli Mäkinen. Geometric Algorithms for Transposition Invariant Content-Based Music Retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 193–199, 2003.

[34] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop*, pages 125–125, 2016.

[35] Gissel Velarde, Tillman Weyde, and David Meredith. An Approach to Melodic Segmentation and Classification Based on Filtering with the Haar-Wavelet. *Journal of New Music Research*, 42(4):325–345, 2013.

[36] Gerhard Widmer. Getting Closer to the Essence of Music: The Con Espressione Manifesto. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):19, 2017.