SYNTHESIZING ENGAGING MUSIC USING DYNAMIC MODELS OF STATISTICAL SURPRISAL

Sandeep Kothinti[1]*, Benjamin Skerritt-Davis*[1]*, Aditya Nair*[2]*, Mounya Elhilali*[1]

[1]Johns Hopkins University, Baltimore, USA,
[2]University of Washington, Seattle, USA

ABSTRACT

Synthesis of music content generally leverages the underlying statistical structure of music to develop generative models, able to create new musical expressions within the same genre. In this work, we explore the statistical structure of a musical corpus and its effect on modulating the attention of listeners. The study specifically explores listeners' engagement to newly synthesized music and tests the hypothesis that maximizing statistical surprisal would result in increased auditory salience. The study employs a dynamical statistical model to estimate melodic line surprisal and develops an optimization procedure using parametrized codebooks to synthesize musical segments that maximize statistical surprisal. A behavioral experiment with a dichotic listening task is designed to probe salience of the synthesized melodies against original melodies by measuring listeners' engagement in a continuous-fashion. Results indicate that we can control the salience of sounds by manipulating the statistical surprisal, guided by the complexity of the temporal structure of the musical corpus. This work suggests that future work in automated music synthesis could leverage statistical models of music beyond musical aesthetics to also manipulate the degree of engagement.

Index Terms— Statistical surprisal, auditory attention, music synthesis, auditory salience, regularity extraction

1. INTRODUCTION

Music is often described as a series of moments of tension and release, where the music builds expectations over time that create tension when they are violated and release when they are met. Algorithmic composition and automated synthesis of music content attempt to leverage this underlying statistical structure by developing computational models to learn complex structures and rhythms that control the musical composition both at the local and global scales [1, 2]. Configurations based on generative systems such as adversarial models or recurrent networks have successfully yielded meaningful musical structures that were reasonably aesthetically pleasing to human listeners [3, 4]. Still, deep learning methods often obscure the statistical structure of music; hence making interpretation of the representations formed by these networks rather difficult.

Beyond musical aesthetics, the perceptual experience of listeners is colored by various aspects of the underlying statistics of the musical structure. In this work, we are interested in the relationship between the statistical constraints of a musical corpus and its effect on engagement of listeners in the musical experience. We specifically focus on the melodic pitch of monophonic pieces and examine how the statistical structure manipulates attentional focus of listeners. Evidence from brain responses indicate that perceived degree of expectedness of a musical piece are closely reflected in neural responses [5, 6]. Based on these results, we hypothesize that expectation violation makes certain moments of a melody more engaging; and by increasing these moments of expectation violations, we can produce more engaging music.

A probability measure for expectation violation is the negative log-likelihood of a given observation, conditioned on past observations. In information-theoretic terms, this quantity is referred to as *surprisal* and is equivalent to self-information. Surprisal gives a measure of how unexpected a given observation is under the assumed statistical model. In [6], surprisal was shown to be a good predictor of the behavioral judgment of unexpectedness of a particular note. Based on these results, we formulate our objective as maximizing surprisal to increase perceived unexpected moments in a melody.

Dynamic models such as Markov chains are often used to model the statistical structure of music. D-REX[7] and IDyOM[8] are two such examples of Markov models of music. While IDyOM uses variable-order Markov chains similar to n-gram language models, D-REX uses a multiple run-length based Markov model, to capture different time-scales of dynamics. These models provide a conditional distribution for each observation based on the past, which makes them convenient for sampling. In the present work, we use the D-REX model as it can provide robust estimates, even when prior training data is unavailable.

This study builds on surprisal estimates from the D-REX model to synthesize new melodies that maximize the violation of expectations of the underlying statistical melody, hence maximizing surprisal. We then verify our hypothesis that such newly synthesized melodies would engage listeners' attention more robustly compared to the original melodies. The study is specifically focused on listener engagement rather than musical aesthetics, and aspires to convey a clear relationship between statistical structure and estimates of surprisal based on musical statistics. Naturally, one of the questions to address in this exploration is the measure of musical engagement. Here, we explore a novel approach based on psychoacoustic testing using dichotic listening. We present subjects with concurrent melodies to each ear and measure which side engages listeners' attention without specifically alerting listeners to any aspect of the melodies. This paradigm is employed to test various key parameters in the musical statistical structure and estimates of surprisal and, ultimately, on attentional engagement of the listener.

This paper presents the methodology of the statistical model and surprisal maximization algorithm in Section 2. Section 3 presents the testing paradigm along with the model manipulations. Section 4 describes the main results of listener engagement and analysis of test

This work initiated from ideas explored at the neuromorphic Engineering workshop in Telluride, Colorado. The authors would like to thank Alain de Cheveigne and Malcolm Slaney for their input on initial ideas. The work was supported in part by NIH U01AG058532, R01HL133043, and 1F31DC017629, ONR N00014-19-1-2014, N00014-17-1-2736, and NSF 1734744.

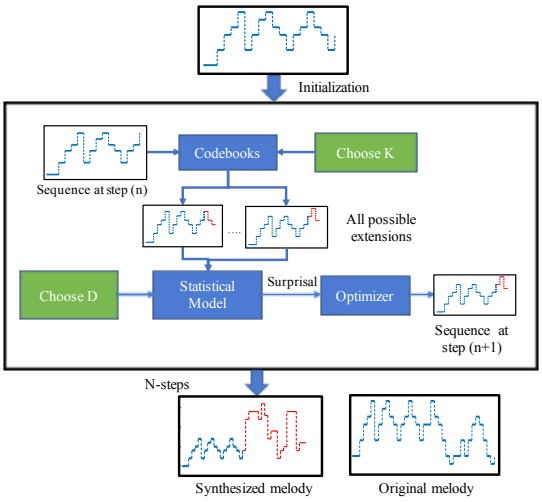

Fig. 1: Block Diagram explaining the synthesis using the statistical model and codebook based search algorithm

conditions; before presenting an overview of the main conclusions in section 5.

2. STATISTICAL MODEL OF MUSIC SURPRISAL

The present work probes a statistical model to understand how statistical structure affects salience of music. The methodology involves (i) inferring a statistical model of an existing musical corpus, (ii) using the model to guide estimates of expectations (or surprisal estimates of the melodic line), (iii) using an optimization algorithm to select components that maximize surprisal in newly synthesized pieces. These new melodies are tested to assess degree of engagement of listeners using a novel dichotic listening paradigm. The different elements of the methodology are outlined below.

2.1. Surprisal model

To model statistical structure of music, we use the Dynamic Regularity EXtraction (D-REX) model[1] presented in [7]. This model is based on a Bayesian inference framework designed to build sequential predictions from dynamic stochastic sequences containing unknown changes in the underlying statistical structure [9, 10].

The input to the model is a sequence of pitches $\{x_t\}$ assumed to be distributed according to a multivariate Gaussian with dimension D and unknown parameters $\theta = \{\mu, \Sigma\}$. The dimension D specifies the extent of temporal covariance collected by the model. For example, a model with $D = 1$ only collects marginal statistics (mean and variance), whereas a model with $D = 3$ additionally collects joint dependencies (i.e., covariances) between x_t, x_{t-1}, and x_{t-2}.

The model sequentially builds a predictive distribution for the next tone at time $t + 1$ given the sufficient statistics estimated from the context with length c: $\mathbb{P}(x_{t+1}|x_{t-c+1:t}) = \mathbb{P}(x_{t+1}|\hat{\theta}_c)$; where $\hat{\theta}_c = \{\hat{\mu}_c, \hat{\Sigma}_c\}$ are the sample mean and sample covariance estimated from the context $x_{t-c+1:t}$.

The model assumes the parameters θ change at unknown changepoint times, rendering all observations post-change statistically independent from those before the change. If the changepoints were known, the ideal context would exclude observations preceding the last changepoint. Because changepoints must be inferred from the inputs, the model maintains multiple hypotheses across different contexts and "integrates out" the unknown context to build a robust prediction: $\mathbb{P}(x_{t+1}|x_{1:t}) = \sum_c \mathbb{P}(x_{t+1}|c, \hat{\theta}_c)\, \mathbb{P}(c|x_{1:t})$. This weighted sum contains the predictions given each context c weighted by the belief in (or, equivalently, the posterior probability of) context c given the observed sequence. Upon observing the new input, the sufficient statistics for each context c, as well as the beliefs, are updated incrementally with x_{t+1} and used to predict subsequent inputs.

From the prediction, we derive a measure of surprisal, S_t, for each input x_t: $S_t = -\log(\mathbb{P}(x_t|x_{1:t-1}))$. Surprisal is a continuous measure of mismatch between the observed input and the prediction, thus more probable observations have low surprisal and less probable observations have high surprisal.

2.2. Surprisal optimization algorithm

The D-REX model (described in the previous section) provides surprisal for any given sequence of pitches. Next, we present an optimization-based synthesis used to produce music with maximum surprisal. The optimization uses a simplex search algorithm [11], which is a gradient-free multidimensional optimization technique.

The synthesized melody is built up sequentially with segments of length K. The choice of K is constrained by the model parameter D, which controls the structure of the statistical model of the melody. To guide the optimization design, we employ codebooks containing fragments of the musical corpus. A codebook C_k consists of pitch sequences of length k. Codebooks are used as building blocks of synthesis in order to confine the range of generated segments within the same expressive segments of the original corpus since no fully automated synthesis approach is used yet. The synthesis is shown in Figure 1 and described in detail below.

1. A sample melody is chosen from the music corpus and the first half of the melody is used for initialization of priors in the D-REX model.
2. For this iteration, parameter D is chosen randomly from a fixed range Γ_D. This choice of D controls the statistical structure used to compute surprisal.
3. The length of the codebook K is sampled from the same range Γ_D. A set of $K + 1$ candidate samples are picked randomly from the codebook C_K and appended to the melody to form $K + 1$ candidate melodies.
4. Surprisal is measured using the D-REX model for each of the candidate melodies.
5. A simplex search is performed to minimize the cost function of mean surprisal and quickly span the search space of melodies. The candidate melody from C_k that minimizes the L2-distance of pitches generated from the optimization is chosen as the output.
6. Now, the selected melody is used as the initialization for Step 2 and the optimization procedure is repeated until the length of the synthesized melody is the same as the original melody.

[1] code available at https://engineering.jhu.edu/lcap/software

3. EXPERIMENTAL SETUP

3.1. Musical corpus

Monophonic excerpts of music from Bach sonatas and partitas were used in this study. Melodies were synthesized with either violin or clarinet sounds sampled from the RWC Musical Instrument Database [12]. A total of 39 melodies were used with an average length of 57 notes presented isochronously at approximately 7Hz, leading to an average duration of about 8 seconds per melody. The stimuli were adapted from work by Di Liberto et al. [13], and used with permission of the authors.

3.2. Subjects

A total of 150 subjects were recruited using AWS Mechanical Turks via a web browser, designed using libraries from jsPsych[14]. Out of the 150 subjects, 93 were male, 56 female, and 1 non-binary. The average age was 33 years. Subjects were compensated after participating in the study. All procedures were approved by the Johns Hopkins Institutional Review Board (IRB).

3.3. Behavioral paradigm

To test engagement of the synthesized music, we designed a dichotic listening task, adapting a procedure previously used to study auditory salience [15]. Subjects were presented with the synthesized melody in one ear and the corresponding original melody in the other ear, therefore the presentation to the two ears only diverges in the second half of the melody. Subjects are asked to report which side they are focusing on at every moment, hence delivering a continuous measure of engagement rather than a single binary report at the end of the melody. Subjects can move the cursor to the side which they are attending to or keep the cursor in the middle to indicate attending to both ears or neither.

In dichotic listening tasks, it is often difficult to pay attention to one of the ears. To facilitate discrimination between the two melodies playing simultaneously, we use two instruments with distinct timbres. We chose violin to play one melody and a clarinet to play the other melody. To compensate for any preference towards one of the instruments, we conduct two trials for each melody combination such that each melody is presented with both violin and clarinet. The side playing the synthesized melody (left vs. right ear) is randomly chosen in each trial.

3.4. Test parameters

One of the open questions is the direct impact of the underlying statistical structure of the corpus on the optimization procedure for synthesizing new melodies. Specifically, we are interested in the influence of the extent of temporal structure captured by covariances over different time windows. The D-REX model controls such structure via the model parameter D. In this work, we contrasted two degrees of complexity in the underlying model, a lower-order model with $D \in \{1, 2, 3\}$ versus a higher-order model with $D \in \{8, 9, 10\}$. The former (referred to as *short*) assumes a more local dependency between notes as they evolve in the melodic line, while the later (referred to as *long*) considers more complex phrasing and temporal relationships and could arise from presence of patterns like arpeggio cycles. For the short condition, both variables D and K are sampled from the range $\{1,2,3\}$. This corresponds to synthesizing in shorter steps and smaller D, reducing the time-scales captured by the D-REX model. In the long condition, we sample D and K from $\{8, 9, 10\}$, incorporating more temporal complexity in the statistical model.

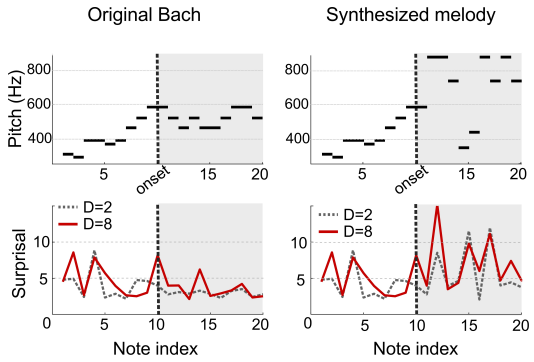

Fig. 2: Example Bach melody (left panel) and synthesized version (right panel) showing changes in melodic pitch. Bottom panels show estimate of statistical surprisal for each melody using the a short model (D=2) and a long model (D=8).

To ensure any effects observed are due to the statistical model driving the synthesis, we performed a control experiment where we chose random notes from the codebook with sizes $K \in \{1, 2, 3\}$. In this *control* case, the procedure for the short condition is replicated but instead of maximizing surprisal, a random segment from the codebook is selected. This control case serves as a baseline to examine whether the mere fact of shuffling the melodic line of an existing Bach melody would in fact attract listeners' attention beyond our stated hypothesis that presumes that this attentional engagement is contingent on maximizing surprisal.

4. RESULTS

Figure 2 shows an example of an original Bach melody (left) along with a continuous estimate of surprisal for each note using the D-REX model with an underlying multivariate Gaussian model with short temporal structure (D=2, dashed line) versus an analysis of the same melody using a longer temporal structure (D=8, solid line). The right panel shows the corresponding synthesized melody using the left melody for initialization. Surprisal estimates (bottom panels) highlight the differences between the underlying statistics that employ different granularity of sufficient statistics. A given note can vary widely in terms of how well it fits expectations of a short-term vs. a long-term model indicating that a choice of such statistics needs to be carefully considered for a specific musical corpus to capture the musical phrasing and complexity of melody dynamics. The right panel also shows how the synthesis procedure (shown after half-way point) not only changes the local estimates of surprisal of individual notes; but also results in widely different estimates for different statistical models.

Next, we examine the effectiveness of the dichotic procedure in capturing the engagement of listeners with these newly synthesized melodies. The analysis procedure averages the response of each subject for each trial, yielding a curve that is scaled between [0,1], where 0 indicates attending to the original Bach melody, 1 indicates attending to the synthesized stimuli, and 0.5 indicates subjects have no preference between the original and synthesized melodies. The synthesized stimulus deviates from the original melody at the halfway point; thus, we refer to responses before this point when melodies presented to both ears are the same as "pre-onset" and responses af-

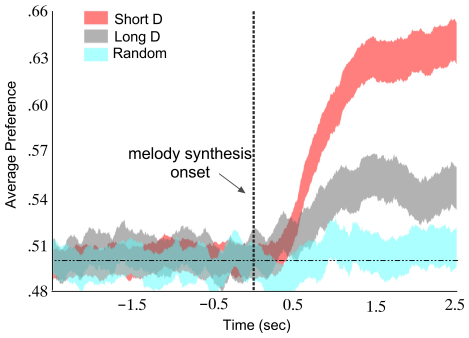

Fig. 3: Response profile of listeners showing preference to newly synthesized melodies (towards 1 on the y-axis) as compared to no preference (0.5 on the y-axis). Plots show 95% confidence intervals pooled across subjects and melodies for the short and long synthesis conditions, as well the control condition.

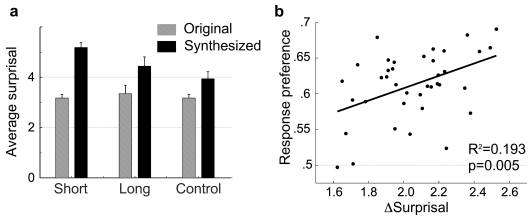

Fig. 4: **(a)** Estimate of average statistical surprisal for each melody using the short and long-term models. **(b)** Relationship between change in surprisal (after synthesis) for the short model and listeners' engagement as reflected in their average preference of ear of entry.

ter this point as "post-onset". Our analysis performs statistical tests by averaging responses 2 seconds before the onset and contrasting them with average responses 2 seconds after a delay of 0.5 seconds after the onset. The delay of 0.5 seconds is chosen to account for the latency in the subjects' response.

First, we focus on the "pre-onset" phase to validate the experimental paradigm. Before any induced change in melodic line, we check whether the average preference of listeners hovers around 0.5, when the two stimuli presented to each ear differ only in instrument. Pooling trials by instrument shows a significant preference towards the ear playing the violin for the test conditions(unpaired t-test, p=0.01(short), 5e-4(long)). This preference was not significant for control(p=0.16) which could be attributed to differences in subject inclinations. This result is consistent with reported differences between timbres of the violin and clarinet suggesting stronger salience of violin over the clarinet [16]. However, our paradigm counter-balanced melodies played by each instrument to either ear, and pooling across all trials regardless of instrument and ear of presentation confirms that the average pre-onset response is not significantly different from 0.5 (p=0.99(short), 0.49(long), 0.69(control)). Next, we examine subjects' response to synthesized melodies. Figure 3 shows the average response profile pooled across subjects for the original and synthesized melodies using the short and long temporal structures. A t-test on the post-onset preference averaged across stimuli shows that the preference for synthesized melodies is significantly higher than 0.5 for both short (p=4e-7) and long (p=1e-4) cases indicating that on higher average preference towards synthesized melodies. Moreover, it is clear from the plot that subjects were more engaged by the melodies synthesized using the local structure (i.e. short or low D values), as opposed to melodies driven by expectations over longer-term correlations (i.e. long or high D values). A comparison of the two cases shows a statistically significant difference between the two curves (p=2e-8).

Also interesting to note is that the control experiment did not show any indication of engagement of listeners and revealed preferences that continued to hover around 0.5. The control paradigm consisted of newly synthesized melodies that were not specifically maximizing statistical surprisal but were randomly concatenated segments from the codebook. Post-onset average preference when averaged across stimuli, was not statistically different from 0.5 (p=0.53). Thus, listeners did not have a higher preference towards synthesized melodies that were randomly shuffled, suggesting the increase in average preference post-onset in the short and long conditions was driven by maximizing statistical surprisal.

We compare the change in average surprisal for original and synthesized melodies. Figure 4a shows the average melodic surprisal for Bach melodies quantified using the short model(D=2) for the test and control experiments; and long model(D=8) for long case. Overall, the original melodies have comparable average surprisals under both models. After the optimization algorithm, it appears that the short model is able to generate melodies with higher average surprisal than the long and control algorithms.

To quantify how surprisal affects listener preference, we performed a correlation analysis on the post-onset average preference and the difference in average surprisal between the synthesized and original melodies. Figure 4b shows the details of this analysis and reveals a significant correlation(R^2=0.193, p=0.005) for the short condition. For the long(R^2=0.01, p=0.48) and control(R^2=0.08, p=0.08) cases, there was no significant correlation.

5. CONCLUSION

In this paper, we presented a novel methodology to synthesize engaging monophonic melodic music by taking into account the statistical structure of music and well-known perceptual phenomena of expectation violation into consideration. By means of well designed psycho-acoustic experiment, we tested the synthesized melody for increased engagement compared to an existing melody. Experimental results show that the proposed methodology can be used to synthesize engaging music. It should be noted that the current study did not attempt to control for musical aesthetics but was solely concerned with the degree of salience of newly synthesized melodies. Control experiments confirmed that random selection of new musical segments is not sufficient to attract attention of listeners to random melodic lines. Instead, only melodies that were constrained by surprisal within an underlying statistical model were considered engaging by listeners. Furthermore, the Bach corpus chosen for this study appears to favor a rather local temporal structure resulting in stronger listener responses for synthesized melodies controlled by the short statistical model. There is an also the possibility that such favorable outcome with a short model may reflect the dynamics of the optimization algorithm in terms of timescale of updating samples from the codebook. Overall, this work sets an initial exploration of measures of musical engagement and its relationship with underlying statistics of a musical corpus. These aspects of musical expression can then be incorporated with future work aiming to provide a richer artistic expression for computer generated music.

6. REFERENCES

[1] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet, "Deep learning techniques for music generation-a survey," *arXiv preprint arXiv:1709.01620*, 2017.

[2] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts, "Gansynth: Adversarial neural audio synthesis," in *International Conference on Learning Representations*, 2019.

[3] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent, "Modeling temporal de- pendencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *Proceedings of the 29th International Conference on Machine Learning*. JMLR. org, 2012.

[4] Allen Huang and Raymond Wu, "Deep learning for music," *arXiv preprint arXiv:1606.04930*, 2016.

[5] Stefan Koelsch, Tomas Gunter, Angela D Friederici, and Erich Schröger, "Brain indices of music processing:"nonmusicians" are musical," *Journal of Cognitive Neuroscience*, vol. 12, no. 3, pp. 520–541, 2000.

[6] Marcus T. Pearce, María Herrojo Ruiz, Selina Kapasi, Geraint A. Wiggins, and Joydeep Bhattacharya, "Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation," *NeuroImage*, vol. 50, no. 1, pp. 302 – 313, 2010.

[7] Benjamin Skerritt-Davis and Mounya Elhilali, "Detecting change in stochastic sound sequences," *PLOS Computational Biology*, vol. 14, no. 5, pp. e1006162, 5 2018.

[8] M.T. Pearce, "The construction and evaluation of statistical models of melodic structure in music perception and composition," December 2005.

[9] Ryan Prescott Adams and David J. C. MacKay, "Bayesian Online Changepoint Detection," Tech. Rep., University of Cambridge, Cambridge, UK, 2007.

[10] Matthew R Nassar, Robert C Wilson, Benjamin Heasly, and Joshua I Gold, "An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment," *Journal of Neuroscience*, vol. 30, no. 37, pp. 12366–12378, 2010.

[11] John A Nelder and Roger Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[12] M Goto, H Hashiguchi, T Nishimura, and R Oka, "RWC music database: Music genre database and musical instrument sound database," *Proceedings of International Symposium on Music Information Retrieval*, pp. 229–230, 2003.

[13] Giovanni M. Di Liberto, Claire Pelofi, Roberta Bianco, Prachi Patel, Ashesh D. Mehta, Jose L. Herrero, Alain de Cheveigné, Shihab Shamma, and Nima Mesgarani, "Cortical encoding of melodic expectations in human temporal cortex," *bioRxiv*, 2019.

[14] Joshua R De Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," *Behavior research methods*, vol. 47, no. 1, pp. 1–12, 2015.

[15] Nicholas Huang and Mounya Elhilali, "Auditory salience using natural soundscapes," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2163, 3 2017.

[16] S McAdams, S Winsberg, S Donnadieu, G De Soete, and J Krimphoff, "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes," *Psychological Research*, vol. 58, no. 3, pp. 177–192, 1995.