Noname manuscript No.
(will be inserted by the editor)

On the evaluation of generative models in music

Li-Chia Yang · Alexander Lerch

Received: date / Accepted: date

Abstract The modeling of artificial, human-level creativity is becoming more and more achievable. In recent years, neural networks have been successfully applied to different tasks such as image and music generation, demonstrating their great potential in realizing computational creativity. The fuzzy definition of creativity combined with varying goals of the evaluated generative systems, however, make subjective evaluation seem to be the only viable methodology of choice. We review the evaluation of generative music systems and discuss the inherent challenges of their evaluation. Although subjective evaluation should always be the ultimate choice for the evaluation of creative results, researchers unfamiliar with rigorous subjective experiment design and without the necessary resources for the execution of a large-scale experiment face challenges in terms of reliability, validity, and replicability of the results. In numerous studies, this leads to the report of insignificant and possibly irrelevant results and the lack of comparability with similar and previous generative systems. Therefore, we propose a set of simple musically informed objective metrics enabling an objective and reproducible way of evaluating and comparing the output of music generative systems. We demonstrate the usefulness of the proposed metrics with several experiments on real-world data.

Keywords Objective Evaluation · Music Generation · Computational Creativity

Li-Chia Yang
Georgia Institute of Technology, Center for Music Technology
E-mail: richard40148@gatech.edu

Alexander Lerch
Georgia Institute of Technology, Center for Music Technology
E-mail: alexander.lerch@gatech.edu

1 Introduction

The desire to understand creativity has driven the development of computationally creative systems among a wide variety of tasks [5]. Just as deep learning has reshaped the whole field of artificial intelligence, it has reinvented generative modeling in recent years [63]. This thriving research area includes, for example, the creative generation or the style transfer of artwork such as paintings or music [15, 21].

Even with the research interest in generative systems, the assessment and evaluation of such systems has proven challenging. Formally, categorization of evaluation strategies can be derived from specifying the design ontology of the system. For instance, based on the Function-Behavior-Structure (FBS) ontology [18,62], we evaluate the actual behavior of a system compared to its expected behavior. The evaluation of creative systems can be categorized into function and structure evaluation, which relates directly to the so-called *summative* and *formative* approaches. While the former aims to assess whether the results of a system meet the stated goal of creativity, the latter focuses on monitoring how the instructional goals and objectives are being met [13,20,46]. Without a clear definition and consensus on the essence of (human) creativity, *summative* evaluation remains largely problematic [28].

As the ultimate judge of creative output is the human (listener or viewer), subjective evaluation is generally preferable in generative modeling. The challenges of designing and conducting an experiment leading to valid, reliable, and replicable results, however, are often underestimated. Controlling all relevant variables, eliminating bias, and recruiting a sufficient number of qualified subjects can easily blow the required resources out of reach for small-scale projects.The most common shortcomings

of subjective studies evaluating generative systems are closely related to both the available resources and the design of experimental methodology [28, 47].

Thus, a method for objective evaluation of generative systems is desirable.

The image generation community has benefited from the introduction of the idea of the *inception score* by Salimans et al. [47]. It uses a pattern recognition model to assess the generated sample. The general concept of the inception score is based on the assumption that a well-trained image classifier roughly has a human-like classification ability [47]. This idea has been adapted by multiple researchers to allow for an objective measure of various generative systems [26, 29, 39]. The idea of the inception score is convincing and the first results look promising; ultimately, however, the assumed correlation to human judgment still needs further scientific examination [19, 64].

The evaluation of generative music systems faces even harder challenges than that of image generation systems [9]. The sequential yet highly structured form, the ever-changing interaction between composition and performance, and the abstract nature of meaning and emotion in music [36, 61] make a semantic description of music exceedingly hard. The automatic analysis and categorization of music is, although having made great progress, not close to human-level performance [35]. This makes assessing music very difficult [3, 22, 41, 59] and partly explains why music assessment could not be automated by computational models so far.

Despite these high-level challenges, we will show below that state-of-the-art generative music systems struggle with creating musical content that follows basic technical rules and expectations. We argue that these technicalities have to be solved before addressing the questions of aesthetics of creative works with high-level structural and harmonic properties.

Therefore, we propose a *formative* evaluation strategy for systems generating symbolic music. The proposed method does not aim at assessing musical pieces in the context of human-level creativity nor does it attempt to model the aesthetic perception of music. It rather applies the concept of *multicriteria* evaluation [54] in order to provide metrics that assess basic technical properties of the generated music and help researchers identify issues and specific characteristics of both model and dataset. The usefulness of the presented method is demonstrated through a series of experiments, including dataset analysis, comparison of state-of-the-art music generation models, and assessment of generative music systems.

2 Related Work

As mentioned above, research on automatic music generation systems has suffered from the difficulty of designing evaluation methodologies [42]. The two challenges of measuring the success of a generative system are addressing the *summative* and the *formative* assessment of the system behavior. Subjective approaches to measuring the success of generative systems by means of listening experiments can often be categorized as *summative* assessment while objective evaluation strategies mostly fall into the category of *formative* assessment. Confusing these two challenges leads to unclear evaluation strategies. Although subjective evaluation is generally preferable for evaluating generative modeling, it might require significant resources. Objective methods, on the other hand, can be easily executed yet often lack musical relevance as they are often not based on musical rule systems or heuristics.

2.1 Subjective evaluation in music generation

Most assessments of generated symbolic music are based on inputs from human listeners. These evaluations either follow the concept of a musical Turing test [3] or use query metrics based on the modeled compositional theory [2].

The Turing test [55] follows an intuitive concept that evaluates whether a machine is able to exhibit behavior indistinguishable from humans. One strategy to adapt the Turing test to generative music systems is asking the subjects to identify the pieces they consider to be composed by a human as opposed to a computer [34]. This strategy has been used in several studies as listed in Table 1 [1, 21, 24, 25, 32, 49]. Over the past decades, shortcomings of the Turing test have been pointed out in various areas [2, 17, 44]. Many of these problems also apply to musical Turing tests. One of the fundamental issues, however, is that many studies confound the two questions on whether a piece is aesthetically pleasing and whether it is composed by a human.

The design of a listening experiment is complex due to the many variables ranging from the selection and rendition of audio examples, the listening environment, and the selection of subjects, to the phrasing of the questions. Without proper guidance (compare, e.g., [6]), we find that many contemporary studies struggle with presenting significant scientific evidence. Table 1 lists some of the variables for several major subjective evaluation studies in the context of music generation. It is worth noting that all of these evaluations are performed with a different problem configuration, i.e., different evaluation criteria are used, and both the questionnaires and

	[43]	[11]	[60]	[49]	[1]	[24]	[25]	[21]	[32]
Subject's background	T	N/A	T & UT	3L	N/A	N/A	N/A	3L	4L
Sample size	16	27	21	973/986	48	96	52	1272	759
Comp. w/ models	yes	yes	yes	yes	yes	no	yes	yes	no
Comp. w/ human-composed	yes	no	no	yes	yes	yes	yes	yes	yes
Comp. w/ random samples	no	no	no	no	no	no	no	no	no

Table 1: Experiment design for subjectively evaluating music generation research. The following abbreviations are use for the subjects' background: T (musically trained), UT (musically untrained), 3L (Three-point level of expertise), 4L (Four-point level of expertise).

the listening examples are proprietary (if not arbitrary) and hard to compare. Without addressing these issues properly, the reported results can only be understood as providing preliminary evaluation results and fail at representing a scientific benchmark. First, the majority of them ignore factors associated with the subjects themselves (e.g., their level of expertise), which influences further analysis and the reliability of the experimental result [6]. Second, most the studies rely —probably due to limited resources— on a relatively small sample size [11, 43, 60], which raises questions about the range of the confidence interval and the study's statistical significance (which are often not reported) [33]. Note that the common lack of reported statistical measures of confidence and significance in itself could be seen as an indicator of insufficient scientific rigor. Finally, some of the studies rely on the preference of one model over another [11, 60]. The drawback of such a test paradigm is the absence of a standard comparison or absolute reference. While it can be used to measure relative differences or improvements, it cannot provide any absolute measurement of quality.

Last but not least, these tests carry the risk of overestimating the subject's comprehension, as Ariza concludes after comparing several subjective evaluation methods (e.g, *Musical Turing Tests*, *Musical Directive Toy Tests* and *Musical Output Toy Tests*) [2].

2.2 Objective evaluation in music generation

Given the advantages over subjective evaluation with respect to reproducibility and required resources, several recent studies have assessed their models objectively. We categorize the objective evaluation methods used by the recent studies on data-driven music generation into the following categories: (i) probabilistic measures without musical domain knowledge, (ii) task/model specific metrics, and (iii) metrics using general musical domain knowledge.

2.2.1 Probabilistic measures

The use of evaluation metrics based on probabilistic measures such as likelihood and density estimation has been successfully used in tasks such as image generation [54] and is increasingly used in music-related tasks as well [14, 52]. For example, Huang et al. propose a frame-wise evaluation computing the negative log-likelihood between the model output and the ground truth across frames [24]. Similarly, Johnson considers the note combinations over time steps of the training data as the ground truth and reports the summation of the generated sequence's log-likelihood across notes and time steps [27]. Since the recurrent model used in his study is trained with the goal of maximizing the log-likelihood of each training sequence, the measure is argued to be a meaningful quantitative measure of the performance. The used probabilistic measures provide objective information, yet Theis et al. observe that "A good performance with respect to one criterion does not necessarily imply a good performance with respect to another criterion" and provide examples of bad samples with very high likelihoods [54].

2.2.2 Model-specific metrics

As the approaches and models vary greatly between different generative systems, some of the evaluation metrics are correspondingly designed for a specific model or task. Bretan et al. proposed a metric for successfully predicting a music unit from a pool of units in a generative system by evaluating the rank of the target unit [8]. Mogren designed metrics informed by statistical measurements of polyphony, scale consistency, repetitions, and tone span to monitor the model's characteristics during its training [37]. Common to these evaluation approaches is the use of domain-specific, custom-designed metrics as opposed to standard metrics. Obviously, the authors realized the problems with using standard metrics (e.g., edit distance of melodies) as musically meaningless and implemented metrics inspired by domain knowledge. The variability and diversity of the proposed metrics, however, leads to comparability issues. The

design of non-standard metrics also poses additional dangers, such as evaluating only one aspect of the output, or evaluating with a metric that is part of the system design.

2.2.3 Metrics based on domain knowledge

To address the *multi-criteria* nature of generative systems and their evaluation [9], various humanly interpretable metrics have been proposed. More specifically, these metrics integrate musical domain knowledge and enable detailed evaluation with respect to specific music characteristics. Chuan et al. utilize metrics modeling the tonal tension and interval frequencies to compare how different feature representations can influence a model's performance [12]. Sturm et al. provide a statistical analysis of the musical events (occurrence of specific meters and modes, pitch class distributions, etc.), followed by a discussion with examples on the different application scenarios [52]. Similarly, Dong et al. apply statistic analysis including tonal distance, rhythmic patterns, and pitch classes to evaluate a multi-track music generator [14]. The advantages of metrics taking into account domain knowledge are not only in their interpretability, but also in their generalizability and validity — at least as long as the designed model aims to generate music under the established rules.

3 Method

Following the approach of using domain knowledge for designing human-interpretable evaluation metrics for generative music systems, we present a *formative* evaluation strategy based on a comprehensive set of simple yet musically meaningful features that can be easily applied to a wide variety of different symbolic music generation models.

The two targets of the proposed evaluation strategy are to provide (i) absolute metrics in order to give insights into properties and characteristics of a generated or collected set of data, and (ii) relative metrics in order to compare two sets of data, e.g., training and generated. The overall method is illustrated in Fig. 1 and described below.

In a first step, we gather two collections of samples as our input datasets. For the application of objective evaluation, one dataset contains generated samples, the other contains samples from the training (target) dataset. This approach can also be used for applications such as dataset analysis or the comparison of characteristics of two generative systems. We then extract a set of custom-designed features that are rooted in musical domain-knowledge yet easy to understand and interpret. These features encompass both pitch-based and rhythm-based features. After extracting these features for both datasets, we are able to compute both an *absolute measurement* (Fig. 1 top) and a *relative measurement*. The absolute measurement can provide useful insights to a system developer about the training dataset properties and generative system's characteristics.

The *relative measurement* (Fig. 1, bottom), on the other hand, allows to compare two distributions in various dimensions. It is computed by first applying pairwise exhaustive cross-validation to compute the distance of each sample to either the same dataset (intra-dataset) or to the other dataset (inter-dataset). The results are distance histograms per feature. Next, the Probability Distribution Function (PDF) of each feature histogram is estimated by kernel density estimation [50].

Finally, we compute two metrics for the objective evaluation of generative systems from the training dataset's *intra-set* distance PDF (target distribution) and the *inter-set* distance PDF between the training and generated datasets: (i) the area of overlap and (ii) the Kullback-Leibler Divergence (KLD). The steps are introduced in detail in the following sections.

3.1 Input representation

Our proposed evaluation method reads input files in Musical Instrument Digital Interface (MIDI) format. MIDI is considered as one of the standard formats of symbolic domain representation of music [38]. Although a music generation system might have its own data representation and output format, the output is usually converted to MIDI format for distribution and auralization. The MIDI file format also provides useful musical metadata such as the time signature and the bar length through the resolution of the MIDI file.

For the current implementation of our method, the input samples are required to be monophonic melodies with a fixed number of measures.

3.2 Feature extraction

The features listed below are computed for both, the entire sequence, and for each measure in order to get some structural information.

3.2.1 Pitch-based features

1. *Pitch count* ***(PC)***: The number of different pitches within a sample. The output is a scalar for each sample.

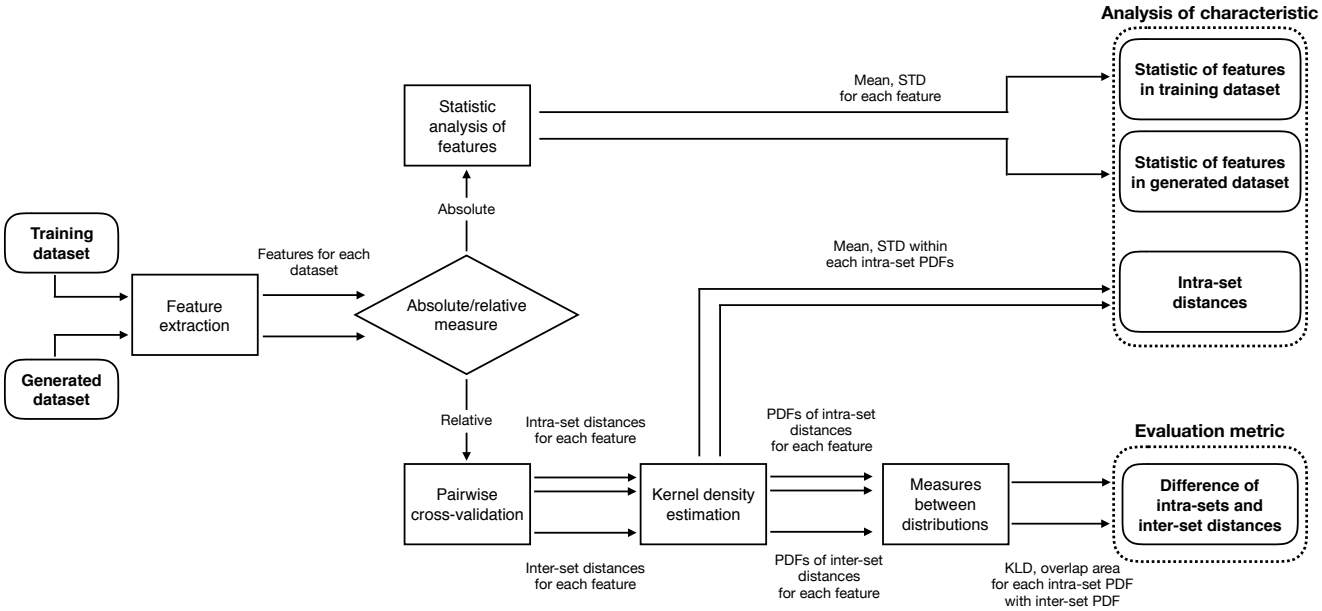


Fig. 1: General work flow of the proposed method

2. *Pitch class histogram* ***(PCH)***: The pitch class histogram is an octave-independent representation of the pitch content with a dimensionality of 12 for a chromatic scale [4, 40]. In our case, it represents the octave-independent chromatic quantization of the frequency continuum.
3. *Pitch class transition matrix* ***(PCTM)***: The transition of pitch classes contains useful information for tasks such as key detection [30, 53], chord recognition [31], or genre pattern recognition [10]. The two-dimensional pitch class transition matrix is a histogram-like representation computed by counting the pitch transitions for each (ordered) pair of notes. The resulting feature dimensionality is 12×12.
4. *Pitch range* ***(PR)***: The pitch range is calculated by subtraction of the highest and lowest used pitch in semitones. The output is a scalar for each sample.
5. *Average pitch interval* ***(PI)***: Average value of the interval between two consecutive pitches in semitones. The output is a scalar for each sample.

3.2.2 Rhythm-based features

1. *Note count* ***(NC)***: The number of used notes. As opposed to the pitch count, the note count does not contain pitch information but is a rhythm-related feature. The output is a scalar for each sample.
2. *Average inter-onset-interval* ***(IOI)***: To calculate the inter-onset-interval in the symbolic music domain, we find the time between two consecutive notes. The output is a scalar in seconds for each sample.
3. *Note length histogram* ***(NLH)***: To extract the note length histogram, we first define a set of allowable beat length classes *[full, half, quarter, 8th, 16th, dot half, dot quarter, dot 8th, dot 16th, half note triplet, quarter note triplet, 8th note triplet]*. The rest option, when activated, will double the vector size to represent the same lengths for rests. The classification of each event is performed by dividing the basic unit into the length of $(barlength)/96$, and each note length is quantized to the closest length category. The output vector has a length of either 12 or 24, respectively.
4. *Note length transition matrix* ***(NLTM)***: Similar to the pitch class transition matrix, the note length transition matrix provides useful information for rhythm description [57]. The output feature dimension is 12×12 or 24×24, respectively.

Obtaining these domain-knowledge based features give us a generally interpretable representation of the data. The features, however, have different dimensionality and normalization, complicating their direct use. Therefore, additional processing is applied to all these features.

3.3 Absolute measurement

During the model design phase of a generative system, it can be of interest to investigate absolute metrics from the output of different system iterations or of datasets as opposed to a relative evaluation. A typical example is the comparison of the generated results from two

generative systems: although the model properties cannot be determined precisely for a data-driven approach, the observation of the generated samples can justify or invalidate system design choices. (e.g., Sect. 4.2).

To acquire the analysis, the mean and standard deviation[1] of each feature of the data are computed.

3.4 Relative measurement

In order to enable the comparison of different sets of data, the relative measure generalizes the result among features with various dimensions; the features are summarized to (i) the *intra-set* distances and (ii) the difference of *intra-set* and *inter-set* distances.

3.4.1 Pairwise cross validation

To compare the distance of the features within and between sets of data, a pairwise exhaustive cross-validation [16] is performed for each feature. In each cross-validation step, the Euclidean distance of one sample to each of the other samples is computed. If the cross-validation is computed within one set of data, we will refer to it as *intra-set* distances. If each sample of one set is compared with all samples of the other set, we call it the *inter-set* distances. The output of this process is a histogram of distances for each feature.

3.4.2 Kernel density estimation

In order to smooth the histogram results for a more generalizable representation, kernel density estimation [50] is applied to convert the histograms into PDFs. A Gaussian kernel and Scott's rule of thumb of bandwidth selection [48, 56] is used for all features in *inter-set* and *intra-set* distances.

Note that the feature dimension plays a role impacting the robustness of density estimation. Silverman provides examples for the relation of sample size and dimensionality for the density estimation and the corresponding mean square error [50].

For the estimated PDFs, simple statistical measures such as mean and standard deviation (STD) can be extracted and directly convey properties of the input datasets. For instance, the mean value in the *intra-set* distances corresponds to the diversity of the samples within a dataset, and the mean value of the *inter-set* distances is a measure of the average similarity of the two input datasets in this feature dimension. On the other hand, the STD value serves as an indication of the reliability of mean value.

[1] The deviation here refers to an element-wise standard deviation, which retains the dimension of each feature.

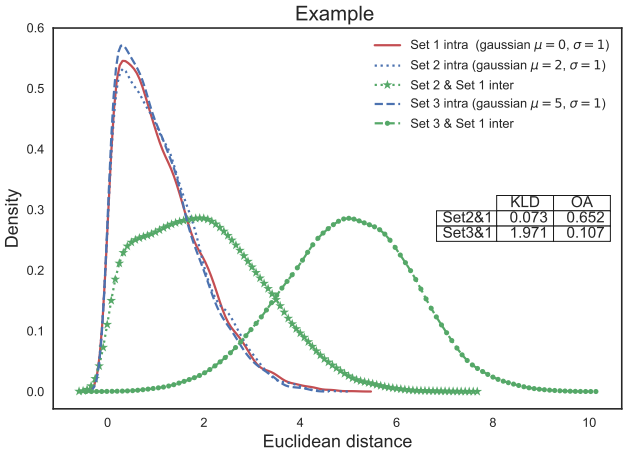


Fig. 2: Example of the proposed evaluation metric: measuring difference of *intra-set* and *inter-set* distances by Kullback-Leibler divergence (KLD) and Overlapped area (OA).

3.4.3 Kullback-Leibler divergence and Overlapped area

In addition to the statistical measures representing *intra-set* distances or *inter-set* distances, similarity measures between distributions are also of interest in the application of evaluating music generative systems. Two metrics are computed, the Kullback-Leibler Divergence (KLD) and Overlapping Area (OA) of two PDFs. We propose to compute the distance between the target dataset's *intra-set* PDF and the *inter-set* PDF.

Although the KLD is the most common measure of how two PDFs diverge from each other, it is unbounded and asymmetric, i.e., $D_{KL}(A||B) \not\equiv D_{KL}(B||A)$); for this reason we further calculate the OA to provide a bounded measure in the range $\in [0, 1]$.

The above similarity measures can indicate the behavior of the evaluated system, as it compares the similarity of two input datasets to each other and within themselves. An artificial example is illustrated in Fig. 2, where we calculate the *intra-set* and *inter-set* distances among three sets of randomly sampled entries from Gaussian distributions with same variance but different mean value (Set 1: $\mu = 0, \sigma = 1$; Set 2: $\mu = 2, \sigma = 1$; Set3:$\mu = 5, \sigma = 1$). Three datasets all have identical *intra-set* distances, but distinct *inter-set* distances. By applying the proposed metric, the smaller KLD and larger OA between Set 2 & Set 1 *inter-set* distances and Set 1 *intra-set* distances shows that Set 2 is more similar to Set 1.

	Folk				Jazz			
	Intra-set		Absolute measure		Intra-set		Absolute measure	
	mean	STD	mean	STD	mean	STD	mean	STD
PC	2.242	1.658	9.300	1.962	3.101	2.355	8.570	2.740
PC/bar	4.178	1.340	-	-	5.635	1.982	-	-
NC	11.583	8.281	47.020	10.018	11.386	8.510	26.270	10.001
NC/bar	5.415	2.446	-	-	7.615	2.905	-	-
PCH	0.339	0.107	-	-	0.480	0.150	-	-
PCH/bar	1.746	0.264	-	-	2.702	0.389	-	-
PCTM	0.307	0.057	-	-	0.417	0.107	-	-
PR	3.773	2.830	15.900	3.318	4.013	3.202	12.150	3.612
PI	0.557	0.439	2.694	0.499	0.989	0.818	2.590	0.903
IOI	0.031	0.027	0.277	0.029	0.838	3.754	0.922	2.706
NLH	0.769	0.504	-	-	0.607	0.229	-	-
NLTM	0.729	0.429	-	-	0.557	0.162	-	-

Table 2: Experimental result of dataset evaluation (see Sect. 4.1)

4 Use-case demonstration and discussion

Three experiments are conducted to demonstrate the value of the proposed analysis of musical characteristics:

1. *Exp. 1 — Dataset evaluation*: the analysis of datasets is one of the fundamental processes of a data-driven experiment. In this experiment, we evaluate (the differences between) two datasets from different music genres, and how this result could inform the developer of a generative system.
2. *Exp. 2 — System comparison*: as mentioned above (see Sect. 2.1), the comparison between two generative systems is a common approach in subjective evaluation experiments. In this experiment, we evaluate two music generation systems and compare the results with the *summative* answers from a subjective evaluation of these systems.
3. *Exp. 3 — Performance evaluation*: a typical problem after prototyping a generative system is the parametrization of the system. This experiment is an example for the typical usage of the objective evaluation method. We discuss how parameters can influence the result of a generative system by comparing the generated samples with the training dataset.

4.1 Experiment 1: Dataset evaluation

Musical style is defined by a set musical characteristics. Due to the complexity of musical content, observing style and properties of a music dataset can be a major challenge. This experiment aims to demonstrate how the proposed approach allows to characterize data from two different music genres and provide insights into genre-specific properties.

4.1.1 Input datasets

The chosen two genres are folk and jazz music. The folk music dataset is the Irish Tunes collected from the Henrik Norbeck's ABC Tunes website [23]. The jazz music dataset comprises jazz lead sheets from both the Wikifonia database [51] and publicly available jazz solo transcriptions collected by Mason et al. [8].

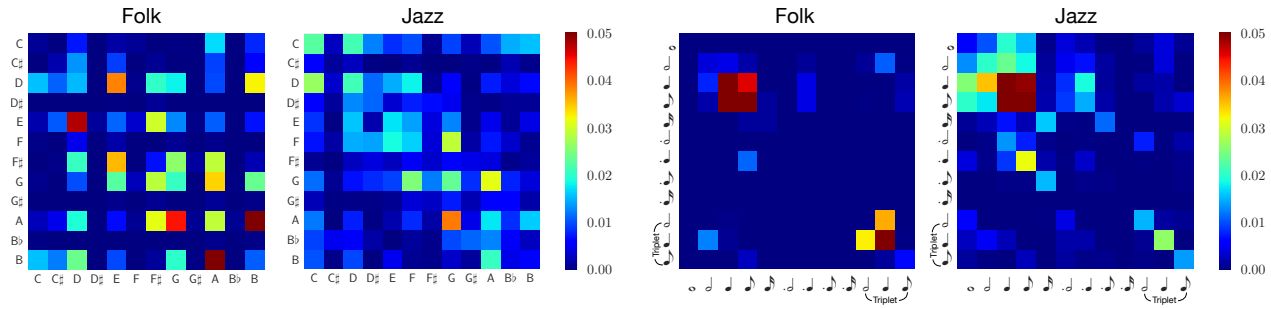
The folk and jazz music datasets contain 2351 and 392 entries, respectively. A pilot experiment determining the necessary amount of samples was carried out. The experiment was then executed with 100 randomly selected songs from each dataset. Of these songs, only the first 8 bars are considered.

4.1.2 Analysis and discussion

Table 2 lists the results for both the *intra-set* distances and the absolute measurements for features with one dimension. We can make the following observations. First, the higher mean of the *intra-set* self-distance for nearly all features in the jazz genre as compared to folk indicates that samples in the jazz genre generally have a higher diversity, a result that matches expectation as folk is often based on simple patterns [45] while jazz generally allows more freedom in its musical composition [7]. Second, we observe considerable differences for the absolute measures of features such as note count and average inter-onset-interval.

Figure 3(a) illustrates the average pitch class transition matrices (PCTM). The folk dataset is more restricted in the usage of certain pitches (i.e., D♯, F, G♯, B♭.) and shows a comparably sparse matrix compared to jazz, where both pitches and pitch transitions tend to have more variety.

We can also observe that the folk music dataset shows a larger mean for features such as note length histogram (NLH), and note length transition matrix



(a) average PCTM (b) average NLTM

Fig. 3: Example of absolute measurement: (a) average pitch class transition matrix (PCTM) and (b) average note length transition matrix (NLTM) of Jazz and Folk music dataset (see Sect. 4.1)

(NLTM). However, by illustrating the average NLTMs in Fig. 3(b), we notice that folk dataset again shows a sparse matrix as compared to the jazz dataset. This implies that the jazz dataset has a higher variety of note length transitions within a song while having a lower diversity of note length transition across the dataset.

In data-driven approaches to music generation, the output of the generative system should directly relate to the characteristics of the training dataset. The presented absolute measures allow for a musically intuitive way of highlighting various dimensions of such characteristics. This can help with the critical step of designing a generalizable dataset, possibly from various sources, for training a generative system.

4.2 Experiment 2: System comparison

The second experiment compares MidiNet [60], a generative adversarial network (GAN) for symbolic domain music generation, with the melody lookback recurrent neural network (Lookback RNN) of the Magenta project [58]. As discussed in the previous Sect. 3.4, the proposed objective evaluation can assist studying different model structures and behaviors when the training datasets for both models are available. In some cases, however, the training datasets are inaccessible as is the case for Magenta. Given this issue, we consider this scenario for the proposed method to compare the characteristics of different models. We again exploit the *intra-set* distances and the absolute measurement utlized in the previous experiment. Furthermore, we attempt to relate reported subjective evaluation results to the identified characteristics.

4.2.1 Input datasets

We implement and train the so-called MidiNet "Model 2" [60], below referred to as MidiNet 2, by using 526 MIDI tabs with 8 bars parsed from the TheoryTab.[2]

The MidiNet model and the public accessible pre-trained model of Magenta's Lookback RNN generate 100 samples each. Each sample contains a melody with 8 bars. The first bar is provided by the user while the remaining 7 bars are generated by the models.

4.2.2 Analysis and discussion

The results of Exp. 2 are shown in Table 3. It can be observed that the two model outputs are distinctly different in several dimensions such as pitch count, pitch interval and pitch range; this is shown by the fact that the mean values of the *inter-set* distances are larger than the mean values of both *intra-set* distances. Furthermore, the absolute measurements NC and PR indicate that MidiNet 2 tends to use more notes and has a higher average pitch range than Magenta's lookback RNN.

The fact that the outputs of these two systems have been used previously in a subjective study [60, Sect. 5] allows us to compare the subjective results with these objective results. The listening test resulted in a comparable rating for the questions *How real* and *How pleasing* the model outputs are; for the question *How interesting*, however, MidiNet acquired a slightly higher rating. This interestingness result might be related to the characteristics of higher pitch range, pitch count, and note count that we find in the absolute measures.

Magenta's RNN, on the other hand, shows a higher mean among the *intra-set* distances in these features; this somewhat contradicts the result of the subjective test. Therefore, we investigate this issue further by looking into the STD value, as a higher STD might hint at a

[2] https://www.hooktheory.com/theorytab

	Magenta				MidiNet				Inter-set	
	Intra-set		Absolute measure		Intra-set		Absolute measure			
	mean	STD	mean	STD	mean	STD	mean	STD	mean	STD
PC	2.897	2.400	7.820	2.647	2.214	1.708	11.300	1.967	4.097	2.490
PC/bar	4.766	1.594	-	-	4.866	1.324	-	-	4.885	1.446
NC	10.228	9.534	27.310	9.837	6.086	4.596	30.740	5.366	8.940	7.576
NC/bar	6.870	2.903	-	-	7.511	1.855	-	-	7.359	2.220
PCH	0.490	0.156	-	-	0.385	0.127	-	-	0.440	0.142
PCH/bar	2.575	0.371	-	-	2.591	0.283	-	-	2.584	0.326
PCTM	0.441	0.099	-	-	0.300	0.049	-	-	0.386	0.079
PR	4.796	3.975	12.650	4.383	3.013	2.631	19.600	2.814	7.681	4.052
PI	1.209	1.274	2.940	1.236	1.105	0.812	5.559	0.965	2.773	1.275
IOI	0.257	0.241	0.653	0.248	0.108	0.095	0.531	0.101	0.205	0.212
NLH	0.538	0.223	-	-	0.237	0.085	-	-	0.420	0.180
NLTM	0.491	0.187	-	-	0.271	0.059	-	-	0.399	0.152

Table 3: Experimental result for characteristic comparison of generation models (see Sect. 4.2)

lower reliability of the mean value. No clear conclusions can be drawn as the limited sample size in the listening test does not allow for more detailed analysis.

Finally, Fig. 4 showcases another visualization of data characteristics. The PDF of the *intra-set* distances among features (PCH, PCTM, NLH, and NLTM) is shown in a violin plot, an intuitive visualization of PDFs. The plot echoes the previous argument, where a significant higher skewness indicates a less diversified *intra-set* behavior and a higher STD indicates a lower reliability of the similarity measure.

4.3 Experiment 3: Performance evaluation

The final experiment demonstrates the use case of evaluating a generative system. We compare two parametrizations of MidiNet, "Model 1" and "Model 2" [60]. Both models have identical architecture and share the same training data. The difference between the models is that one model does not use feature matching regularizers (MidiNet 1) while the other model does (MidiNet 2). *Feature matching* is a technique for stabilizing the GANs by urging the model follow patterns within the training data more closely [47].

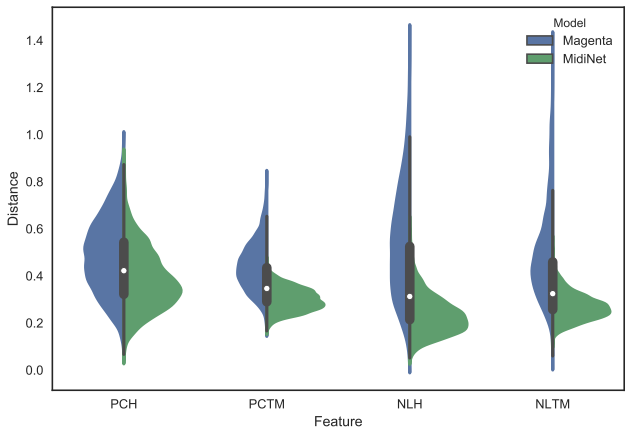


Fig. 4: Visualization of model characteristics through the PDFs of proposed *intra-set* self-distance (Sect. 4.2)

4.3.1 Input datasets

We randomly pick 100 melodies from the training dataset (see Sect. 4.2: 526 MIDI tabs each with 8 bars), and generate 100 samples of melodies with 8 bars each with the two models. To insure an fair comparison, the generation is performed with the same setup as in Sect. 4.2, where we provide one bar for priming and let each model generate 7 continuing bars.

4.3.2 Analysis and discussion

The results of Exp. 3 are shown in Table 4 and Fig. 5. When comparing the generated melodies with the training melodies, the model with active *feature matching*, MidiNet 2, appears to have a larger OA and smaller KLD across almost all features. This indicates that the feature matching is able to deliver the expected improvement. The *intra-set* distance metrics show that both models have —compared to the training dataset— a lower mean and standard deviation in most features. This implies that both systems lose the variety of the training samples. Rather than using the metrics for a quality ranking, we urge the user to use them as index of variability. They could also be used to catch, e.g., an extreme case of losing the variety refered to as *mode collapse* in GANs [47]. In this case, the model is only able to generate very similar samples although the training dataset has significant variability.

Fig. 5 intuitively identifies pitch count (PC), note count (NC), and pitch interval (PI) as the features for

	Training data		MidiNet 1				MidiNet 2			
	Intra-set		Intra-set		Inter-set		Intra-set		Inter-set	
	mean	STD	mean	STD	KLD	OA	mean	STD	KLD	OA
PC	2.527	2.760	2.367	1.805	1.185	0.158	2.214	1.708	0.130	0.541
PC/bar	5.446	2.130	4.551	1.264	0.812	0.042	4.866	1.324	0.572	0.909
NC	12.360	9.954	5.085	3.752	1.058	0.081	6.086	4.596	0.009	0.693
NC/bar	7.804	2.977	5.210	1.520	0.442	0.090	7.511	1.855	0.181	0.916
PCH	0.506	0.169	0.301	0.082	0.012	0.563	0.385	0.127	0.016	0.814
PCH/bar	2.497	0.414	1.546	0.221	0.018	0.319	2.591	0.283	0.025	0.899
PCTM	0.439	0.107	0.263	0.036	0.349	0.277	0.300	0.049	0.172	0.483
PR	4.726	3.935	1.803	1.443	0.502	0.164	3.013	2.631	0.453	0.400
PI	1.062	1.508	0.958	0.767	1.096	0.198	1.105	0.812	0.217	0.443
IOI	0.377	0.403	0.024	0.018	0.141	0.067	0.108	0.095	0.089	0.631
NLH	0.506	0.194	0.174	0.089	0.331	0.187	0.237	0.085	0.024	0.507
NLTM	0.502	0.165	0.208	0.099	0.599	0.201	0.271	0.059	0.112	0.455

Table 4: Experimental result for performance evaluation of generation model (see Sect. 4.3)

which MidiNet 2 outperforms MidiNet 1 (KLD decrease and OA increase drastically). It also points to features such as pitch range (PR) and pitch count across bars (PC/bar) as the dimensions in which both MidiNet models struggle as indicated by a high KLD. Most importantly, the metrics provide the measurement with respect to human interpretable musical features, allowing the user to easily pinpoint the strengths and weaknesses of different system designs.

We can also make one counter-intuitive observation: the KLD for the pitch class histogram features slightly increases from MidiNet 1 to MidiNet 2 while the overlapped areas (OA) become larger. This reveals the limitations of KLD as visualized in Fig. 6: the PDFs of the *intra-set* and *inter-set* distances of MidiNet 2 move towards the training data's *intra-set* distances, however, the KLD measure fails to register a performance improvement. Since in discrete probability distributions, the KLD is calculated in an element-wise manner, PDFs with identical shape (as indicated by similar Kurtosis and Skewness) but shifted on the x-axis (distinct in mean value) yield insignificant differences in KLD. As mentioned in Sect. 3.4.3, the calculation of the OA can address these limitations of the KLD. On the other hand, OA can be misleading when the PDFs vary in their Kurtosis but have similar mean values; in this case, the KLD is able to indicate the differences.

5 Conclusion

Evaluation of generative models has been falling behind the system development itself. This is probably due to the challenges of assessing music aesthetic pieces in the aspect of *summative* evaluation [2], where human subjective tests are typically unavoidable. Given the challenges of required resources and listening experiment design, we have proposed to address this issue by

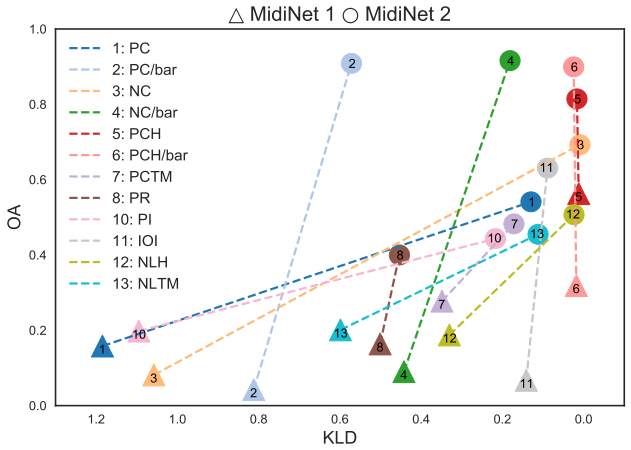


Fig. 5: Visualizing the model performance by the proposed KLD and OA metrics (Sect. 4.3)

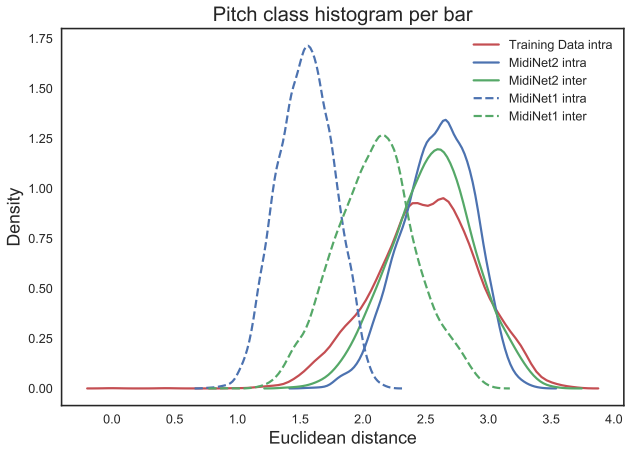


Fig. 6: An example of PDF of the *intra-set* and *inter-set* distances (Sect. 4.3)

using a *formative* objective evaluation for generative music models. This allows for reproducible, reliable, and comparable objective results. It also allows the analysis of large amounts of outputs instead of a small set of hand-picked samples.

The method can be applied to two main tasks, the analysis of characteristics or the objective evaluation with interpretable metrics. Given a pair of datasets, features rooted in musical domain-knowledge are extracted, providing absolute measures to the user quantifying the characteristics of a dataset in various dimensions. When used as evaluation metric, a relative measurement allows to look into *intra-set* and *inter-set* distances with respect to the training and the output data. The statistic analysis with respect to both the absolute measure and the similarity measure serves as a tool for the analysis of quantifiable dataset characteristics. This analysis allows the researcher to draw conclusions about the system's ability to model a certain musical feature of the training dataset, as well as to estimate the variability and the stability of different model designs.

We have released the evaluation framework as an open source toolbox which implements the demonstrated evaluation and analysis methods along with visualization tools. Our future work will include the extension of the current toolbox with additional dimensions (e.g., dynamics) and to expand it towards polyphonic music. This toolbox is available in an online repository [3].

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Agarwala, N., Inoue, Y., Sly, A.: Music composition using recurrent neural networks (2017)
2. Ariza, C.: The interrogator as critic: The turing test and the evaluation of generative music systems. Computer Music Journal **33**(2), 48–70 (2009)
3. Asmus, E.P.: Music assessment concepts: A discussion of assessment concepts and models for student assessment introduces this special focus issue. Music educators journal **86**(2), 19–24 (1999)
4. Babbitt, M.: Twelve-tone invariants as compositional determinants. The Musical Quarterly **46**(2), 246–259 (1960)
5. Balaban, M., Ebcioğlu, K., Laske, O. (eds.): Understanding Music with AI: Perspectives on Music Cognition. MIT Press, Cambridge, MA, USA (1992)
6. Bech, S., Zacharov, N.: Perceptual audio evaluation-Theory, method and application. John Wiley & Sons (2007)
7. Boot, P., Volk, A., de Haas, W.B.: Evaluating the role of repeated patterns in folk song classification and compression. Journal of New Music Research **45**(3), 223–238 (2016)
8. Bretan, M., Weinberg, G., Heck, L.: A unit selection methodology for music generation using deep neural networks. In: International Conference on Computational Creativity (ICCC). Atlanta, Georgia, USA (2017)
9. Briot, J.P., Hadjeres, G., Pachet, F.: Deep Learning Techniques for Music Generation-A Survey. Springer International Publishing (2019)
10. Chordia, P., Rae, A.: Raag recognition using pitch-class and pitch-class dyad distributions. In: International Society of Music Information Retrieval (ISMIR), pp. 431–436. Vienna, Austria (2007)
11. Chu, H., Urtasun, R., Fidler, S.: Song from pi: A musically plausible network for pop music generation. In: International Conference on Learning Representations (ICLR). San Juan, Puerto Rico (2016)
12. Chuan, C.H., Herremans, D.: Modeling temporal tonal relations in polyphonic music through deep networks with a novel image-based representation. In: Association for the Advancement of Artificial Intelligence (AAAI). New Orleans, Louisiana, USA (2018)
13. Colton, S., Pease, A., Ritchie, G.: The effect of input knowledge on creativity. In: Technical Reports of the Navy Center for Applied Research in Artificial Intelligence. Washington, DC, USA (2001)
14. Dong, H.W., Hsiao, W.Y., Yang, L.C., Yang, Y.H.: Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In: Association for the Advancement of Artificial Intelligence (AAAI). New Orleans, Louisiana, USA (2018)
15. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. In: The Annual Meeting of the Vision Sciences Society. St. Pete Beach, Florida, USA (2016)
16. Geisser, S.: Predictive inference, vol. 55. CRC press (1993)
17. Geman, D., Geman, S., Hallonquist, N., Younes, L.: Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences **112**(12), 3618–3623 (2015)
18. Gero, J.S., Kannengiesser, U.: The situated function–behaviour–structure framework. Design studies **25**(4), 373–391 (2004)
19. Gurumurthy, S., Sarvadevabhatla, R.K., Radhakrishnan, V.B.: Deligan: Generative adversarial networks for diverse and limited data. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA (2017)
20. Guyot, W.M.: Summative and formative evaluation. The Journal of Business Education **54**(3), 127–129 (1978). DOI 10.1080/00219444.1978.10534702. URL `https://www.tandfonline.com/doi/abs/10.1080/00219444.1978.10534702`
21. Hadjeres, G., Pachet, F.: Deepbach: a steerable model for bach chorales generation. In: International Conference on Machine Learning (ICML). New York City, NY, USA (2016)
22. Hale, C.L., Green, S.K.: Six key principles for music assessment. Music Educators Journal **95**(4), 27–31 (2009). DOI 10.1177/0027432109334772. URL `https://doi.org/10.1177/0027432109334772`
23. Henrik norbeck's abc tunes (Last accessed: March, 2018). URL `http://www.norbeck.nu/abc/`
24. Huang, C.Z.A., Cooijmans, T., Roberts, A., Courville, A., Eck, D.: Counterpoint by convolution. In: International Society of Music Information Retrieval (ISMIR). Suzhou, China (2017)
25. Huang, K.C., Jung, Q., Lu, J.: Algorithmic music composition using recurrent neural networking (2017)

[3] https://github.com/RichardYang40148/mgeval

26. Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., Belongie, S.: Stacked generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, Nevada, USA (2016)
27. Johnson, D.D.: Generating polyphonic music using tied parallel networks. In: International Conference on Evolutionary and Biologically Inspired Music and Art, pp. 128–143. Amsterdam, The Netherlands (2017)
28. Jordanous, A.: A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. Cognitive Computation **4**(3), 246–279 (2012)
29. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (ICLR). Toulon, France (2017)
30. Krumhansl, C., Toiviainen, P., et al.: Dynamics of tonality induction: A new method and a new model. In: International Conference on Music Perception and Cognition (ICMPC). Keele, UK (2000)
31. Lee, K.: Automatic chord recognition from audio using enhanced pitch class profile. In: International Computer Music Conference (ICMC). New Orleans, Louisiana, USA (2006)
32. Liang, F., Gotham, M., Johnson, M., Shotton, J.: Automatic stylistic composition of bach chorales with deep lstm. In: International Society of Music Information Retrieval (ISMIR). Suzhou, China (2017)
33. Likert, R.: A technique for the measurement of attitudes. Archives of psychology **22**(140), 5–55 (1932)
34. Marsden, A.: Music, intelligence and artificiality. In: Readings in music and artificial intelligence, pp. 25–38. Routledge (2013)
35. Meredith, D.: Computational music analysis. Springer (2016)
36. Meyer, L.B.: Emotion and meaning in music. University of chicago Press (2008)
37. Mogren, O.: C-rnn-gan: Continuous recurrent neural networks with adversarial training. In: Advances in Neural Information Processing Systems, Constructive Machine Learning Workshop (NIPS CML). Barcelona, Spain (2016)
38. Moog, R.A.: Midi: musical instrument digital interface. Journal of the Audio Engineering Society **34**(5), 394–404 (1986)
39. Mroueh, Y., Sercu, T.: Fisher gan. In: Advances in Neural Information Processing Systems (NIPS). Long Beach, CA, USA (2017)
40. O'Brien, C., Lerch, A.: Genre-specific key profiles. In: International Computer Music Conference (ICMC). Denton, Texas, USA (2015)
41. Pati, K.A., Gururani, S., Lerch, A.: Assessment of Student Music Performances Using Deep Neural Networks. Applied Sciences **8**(4), 507 (2018). DOI 10.3390/app8040507. URL `http://www.mdpi.com/2076-3417/8/4/507`
42. Pearce, M., Meredith, D., Wiggins, G.: Motivations and methodologies for automation of the compositional process. Musicae Scientiae **6**(2), 119–147 (2002)
43. Pearce, M.T., Wiggins, G.A.: Evaluating cognitive models of musical composition. In: International joint workshop on computational creativity, pp. 73–80. London, UK (2007)
44. Pease, A., Colton, S.: On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In: Proceedings of the AISB symposium on AI and Philosophy, p. 39. York, United Kingdom (2011)
45. Pease, T., Mattingly, R.: Jazz composition: theory and practice. Berklee Press (2003)
46. Ritchie, G.: Some empirical criteria for attributing creativity to a computer program. Minds and Machines **17**(1), 67–99 (2007)
47. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain (2016)
48. Scott, D.W.: Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons (2015)
49. Shin, A., Crestel, L., Kato, H., Saito, K., Ohnishi, K., Yamaguchi, M., Nakawaki, M., Ushiku, Y., Harada, T.: Melody generation for pop music via word representation of musical properties (2017). `arXivpreprintarXiv:1710.11549`
50. Silverman, B.W.: Density estimation for statistics and data analysis, vol. 26. CRC press (1986)
51. Simon, I., Morris, D., Basu, S.: Mysong: automatic accompaniment generation for vocal melodies. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 725–734. Florence, Italy (2008)
52. Sturm, B.L., Ben-Tal, O.: Taking the models back to music practice: evaluating generative transcription models built using deep learning. Journal of Creative Music Systems **2**(1) (2017)
53. Temperley, D., Marvin, E.W.: Pitch-class distribution and the identification of key. Music Perception: An Interdisciplinary Journal **25**(3), 193–212 (2008)
54. Theis, L., van den Oord, A., Bethge, M.: A note on the evaluation of generative models. In: International Conference on Learning Representations (ICLR). Caribe Hilton, San Juan, Puerto Rico (2016). URL `http://arxiv.org/abs/1511.01844`
55. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236), 433–460 (1950)
56. Turlach, B.A., et al.: Bandwidth selection in kernel density estimation: A review. Université catholique de Louvain Louvain-la-Neuve (1993)
57. Verbeurgt, K., Dinolfo, M., Fayer, M.: Extracting patterns in music for composition via markov chains. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 1123–1132. Springer, Ottawa, ON, Canada (2004)
58. Waite, E., Eck, D., Roberts, A., Abolafia, D.: Project Magenta: Generating long-term structure in songs and stories (2016). `https://magenta.tensorflow.org/blog/2016/07/15/lookback-rnn-attention-rnn/`
59. Wu, C.W., Gururani, S., Laguna, C., Pati, A., Vidwans, A., Lerch, A.: Towards the objective assessment of music performances. In: International Conference on Music Perception and Cognition (ICMPC). Hyderabad, AP, India (2016)
60. Yang, L.C., Chou, S.Y., Yang, Y.H.: Midinet: A convolutional generative adversarial network for symbolic-domain music generation. In: International Society of Music Information Retrieval (ISMIR). Suzhou, China (2017)
61. Zbikowski, L.M.: Conceptualizing music: Cognitive structure, theory, and analysis. Oxford University Press on Demand (2002)
62. Zhang, W., Wang, J.: Design theory and methodology for enterprise systems. Enterprise Information Systems **10**(3), 245–248 (2016). DOI 10.1080/17517575.2015.1080860
63. Zhang, W.J., Yang, G., Lin, Y., Ji, C., Gupta, M.M.: On definition of deep learning. In: World Automation Congress (WAC). Stevenson, Washington, USA (2018)

64. Zhou, Z., Cai, H., Rong, S., Song, Y., Ren, K., Zhang, W., Wang, J., Yu, Y.: Activation maximization generative adversarial nets. In: International Conference on Learning Representations (ICLR). Vancouver, Canada (2018)