on the reverse-sorted normalized summarized chroma vector, which results in a right-skewed distribution, of which we took the mean. From the 36-bin harmonic pitch class profiles (HPCP) extracted by essentia, we derived a *tuning complexity* feature, which expresses the average deviation from 440Hz in both directions. [7]

3.2 Beat Correction and Outlier Detection

The extracted features allow us to address two frequently occurring problems. The first problem is that beat tracking is often inconsistent between different versions of a song, due to the large variance in tempo and the high amount of improvisation in the dataset. The second problem is that some song sets may contain incomplete fragments or recordings of other songs, due to mis-annotations in the Live Music Archive, as noted by Page et al [17].

Our *beat correction* method consists of identifying instances of beat features that should rather be double-time, half-time, or two-thirds-time. [8] We start with the original beat pairings $B^i_b = (R^i, b_{R^i})$ of recordings R^i, $i = 1 \dots K$, and their corresponding beat sequences b_{R^i}, which are simply sequences of time points, as extracted by a feature extractor. For each beat sequence b_{R^k} we create three variants, a linearly interpolated *double-time* version d_{R^k}, [9] a *half-time* version h_{R^k} which only contain every other beat, and a *two-thirds-time* version t_{R^k} which contain every third beat of d_{R^k}. We end up with four sets of K beat sequence pairings $B^i_b, B^i_d, B^i_h, B^i_t$, defined analogously to B^i_b. For each of these pairings we extract a set of features $f_1, \dots f_F$ which all depend on the beat sequence in the pairing. For each feature f with a corresponding distance function δ_f, we then create four $K * K$ distance matrices $D^{x,f}_{ij} = \delta_f(B^i_x, B^j_b)$, one for each $x \in \{b, d, h, t\}$. These matrices express how well the four types of beat pairings match with the original pairings b_{R^i}. We normalize these matrices using min-max normalizations and calculate for each pairing B^k_x its average distance from the original pairings over all features

$$\Delta^k_x = \mu_{f=f_1 \dots f_F, j=1 \dots K}(D^{x,f}_{kj})$$

Finally, we choose for each recording R^k its best pairing B^k_x with minimum Δ^k_x.

The set of features that worked well for the dataset used in this paper are chord sequence distributions, onset distributions, and tempo. We generate a chord sequence distribution from a summarized chord sequence, where for every beat interval we take the chord that occupies the longest duration. From this sequence we gather all subsequences of length L at step size 1 and count the occurrences of each possible pattern. Onset distributions are generated by calculating the position of each onset o relative to its neighboring beats $b_<, b_>$, e.g $(o - b_<)/(b_> - b_<)$ and quantizing the resulting relative onsets to a grid of G values by multiplying them by G rounding to the nearest integer. Both kinds of distributions are normalized. Tempo is simply calculated as the average distance between successive beats. For beat correction, we chose the parameters $L = 4, G = 64$.

For distance functions δ_f we used the chi-square distance $\chi^2(x, y) = \frac{1}{2} \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$ for distributions, and the analogous $\frac{(x-y)^2}{x+y}$ for two single values x, y, such as for tempo.

Next, we identify *outlier recordings* using an extension of the feature set above, by adding chord sequence distributions with $L = 2$ as well as overall beat count, i.e. the length of the beat feature vector. Similar to the beat correction method, we calculate a distance matrix $D^f_{ij} = \delta_f(R^i, R^j)$ for each feature f and normalize it. We then calculate a normalized distance $d_{f,k}$ for each feature, consisting of the difference between the average deviation of R^k from the overall average deviation $\mu(D^f)$, normalized by overall standard deviation $\sigma(D^f)$.

$$d_{f,k} = \frac{\mu_{j=1 \dots K}(D^f_{kj}) - \mu(D^f)}{\sigma(D^f)}$$

Finally, we consider recordings as outliers if they deviate by 2.5 standard deviations in at least two features ($d_{f,k} > 2.5$) or by four standard deviations in at least one ($d_{f,k} > 4$). The latter condition particularly ensures that recordings of extreme length, such as incomplete fragments or improperly segmented files that include more than one song, get excluded due to an extreme deviation of overall beat count.

We obtained the best results by applying beat correction and outlier detection *successively and iteratively*, i.e. one after another in a loop, until both the beat positions and the set of included recordings are stable. Especially for songs with less regular structure, the process only terminates after several iterations, due to the reference beat sequences b_{R^j} changing at each step. Furthermore, with iterative application we can catch cases of quadruple time etc.

When applying this method to the dataset, a total of 211 outliers were removed (around 8%), including for a case where we identified a set consisting of recordings of two songs with similar titles, possibly due to wrong annotations. Of the beat features of the remaining versions, 99 were identified as half-time, 8 as two-thirds-time, and 21 as double-time.

3.3 Statistical Analysis

We are now ready to start our diachronic analysis of the feature data, which is structured as follows. For each feature we have one data point per version of every song. Each version is associated with a specific day on which it was played. For feature f we thus have the data points $P^f = \bigcup P^{f,1} \dots P^{f,15}$ where $P^{f,i} = (p^{f,i}_{k_i})_{k_i = 1 \dots K_i}$ is the sequence of data points for song i with K_i versions.

Although our dataset consists of material from only one band, we face a relatively high variation in features, due

[7] These measures are also related to the information-theoretic complexity measures used by Serra et al or Parmer and Ahn [6, 10].

[8] Two-thirds-time has proven to be useful in situations where we have a ternary meter, such as $6/8$, which might be interpreted as either binary or ternary by the beat extractor.

[9] Each successive pair of beats is interspersed with an average of the two, and an additional beat is added at the end, at an interval corresponding to last interpolated one.

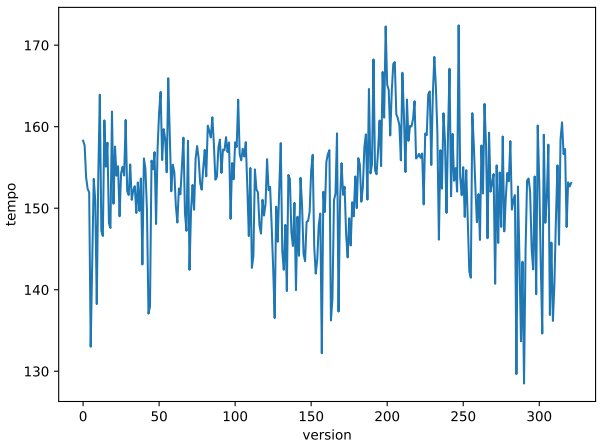

Figure 2: The chronological sequence of tempo data points of all 322 versions of *Sugar Magnolia*.

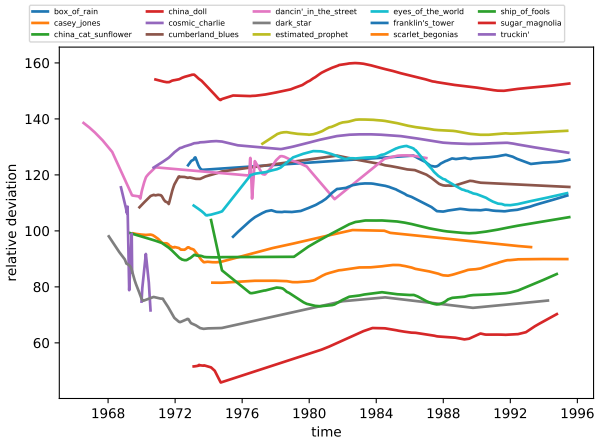

Figure 3: LOWESS plots of tempo for each individual song in the dataset.

to the constant experimentation and improvisation of the band. Figure 2 illustrates how much the tempo of the song *Sugar Magnolia* varies over time, sometimes by as much as 30% from one day to the next. We can identify a global trend in such data by creating a trend line using LOWESS [18], [10] which fits a polynomial regression line at each point, based on a local neighborhood. This neighborhood consists of a fixed proportion of all points, which we set to $f = .2$ or 20%. LOWESS works particularly well for data that is unevenly distributed along the x axis, which is the case here due to the uneven distribution of songs across time (Figure 1).

We can calculate such curves for each song individually, as shown in Figure 3. The top-most curve in this figure corresponds to the line shown in Figure 2 and we can identify two tempo peaks, one around 1973 and one around 1983. It is also apparent that the overall tempo range in the collection varies by more than a factor of 3, from around 50 to almost 160. In order to get a sense of how the tempo changes over all songs, we can create a plot for the whole set of songs, LOWESS(P^f). However, in order to make sure that the curve is not dominated by the values of a single song, we calculate a confidence interval using bootstrapping, leaving out each song once. [11] For each $j = 1 \ldots 15$ we calculate LOWESS($\bigcup_{i \neq j} P^{f,i}$), and determine at each time point the minimum and maximum value among these curves to get the confidence interval.

With absolute values, this confidence interval turns out to be quite large, leaving us unable to draw conclusions with certainty, as shown by the orange curve in Figure 4. For example, we cannot confidently say that the peak observed in 1984 is higher than the one in 1972, due to the overlapping confidence intervals. However, we can get a much better estimate of how the tempo changes by taking *relative feature values*, which we can simply calculate by

[10] As used by Moss et al in their analysis of the evolution of the distribution of pitch-class content on the line of fifths [13].

[11] An alternative method for this is bootstrap resampling, where we could ensure an equal number of points per song. However, due to limitations found in the dataset (after outlier detection one song only has 22 versions), we chose to use the present method.

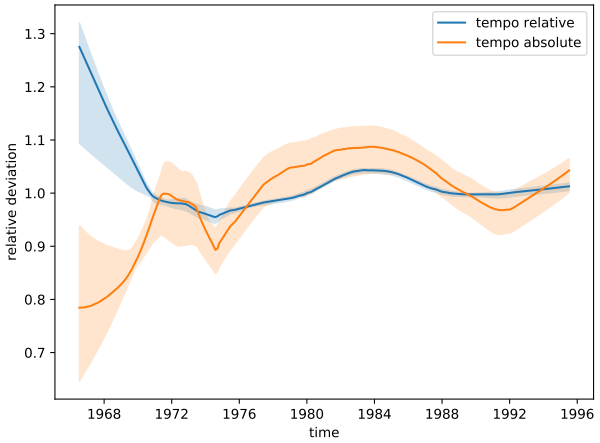

Figure 4: Bootstrapped LOWESS curves on the whole dataset for absolute and relative tempo values.

normalizing the sequences $(p_{k_i}^{f,i})$ as follows:

$$(p_{k_i}^{f,i})' = \frac{(p_{k_i}^{f,i})}{\mu(P^{f,i})}$$

In the case of tempo, each version of every song now has a relative tempo feature, which expresses how its tempo relates to all other performances of the song. The blue curve in Figure 4 is a bootstrapped LOWESS calculated on relative tempo values. We can observe a much narrower confidence interval, now enabling us to confidently claim that in 1984 the tempo was higher than any time in the 70s or 90s. We can also observe that in the beginning of the timeline the curves digress dramatically and with a much larger bootstrapping interval, which is due to the dataset containing few versions of few songs during that time, as observed in Section 2. All subsequent plots are of relative features and calculated as just described.

4. RESULTS

There has been very little musicological work on the Grateful Dead [19, 20], perhaps due to the intimidating size of

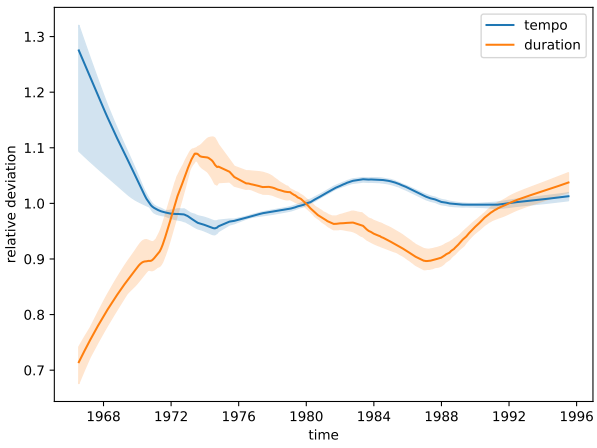

Figure 5: Bootstrapped LOWESS curves for tempo and song duration across all songs.

the band's catalog as well as the intentionally fleeting nature of the performances, as hypothesized by Tift [15]. However, when it comes to in-depth knowledge of the band's music, the best source are undoubtedly the band's fans themselves, the Deadheads. Many of them have spent decades listening to the music and debating it online, and some of them are so knowledgeable that they have an acute sense of how the music changes every single year throughout the band's history. In order to find such general yet detailed descriptions of the evolution of the band's music we searched the Grateful Dead subreddits, [12] the largest forums of their kind with a total of 174k members at the time of writing. We manually read through all pages with discussions on the differences between the band's phases and collected all general descriptions that include the musical characteristics we study, including tempo, timbre, song lengths, etc, with references to specific years. The pages were found using combinations of the search keywords *description, year, decade, 60s, 70s, 80s, 90s, progress, tempo, jam*. Some of the most detailed year by year descriptions found are by Wolfman92097, WesternEstatesHOA, MrCompletely, and an anonymous deleted user [21–24], and some people, such as maximinus-thrax have listened to, rated, and briefly described every of the band's more than 2300 shows [25, 26]. There are also many pages discussing the evolution of individual songs, such as [27]. We will now discuss our results while referring to relevant statements from the community where we can observe a consensus.

Figure 5 juxtaposes different temporal aspects of the music. We can identify an overall increase in tempo in the 1980s of more than 10%, and a subsequent decrease in the middle of the decade. In terms of duration, the differences are even more dramatic, a sharp increase in the early 70s, followed by a gradual decrease by more than 20% until the late 80s, when duration starts increasing again. Note that the early curve segments in all plots, until about 1969, can be considered biased for the reasons stated in Section 3.3. These observations coincide well with a notion in the community that, compared to the 70s, the early 80s are perceived as more energetic, faster paced, and containing fewer and less extensive jams. Wolfman92097 notes that in 1980 "the tempo [starts] to speed up a little", in 1984 "musically the tempo has really sped up", and in 1985 "the band slows the tempo a little." [22] It is striking that this increase and decrease directly corresponds to the tempo curve in Figure 5. Although a bit more abstractly, leedye similarly highlights that exact time period: "79-84 stands out to me as the disco/cocaine era ... post 84 just seems to be a little more of the slow churned vanilla as opposed to that sweet mint chocolate chip." [28] The same perception is true for individual songs. Discussing a recording of the song Eyes of the World, melwarren says "my 4 year old asked why EOTW was going so fast" and whenthattrain-rollsby replies "They played it too fast for my taste in the 80's." [29] On a different page the forum members observe how the song went back to a refreshingly new slow tempo in 1990 [27]. These observations can also be verified in Figure 3.

In terms of song duration, braney86 notes that the "late 70s was full of monster extended jams, and Jerry was absolutely on fire", while MrCompletely says that "82 through 85 is a long uneven slide down ... 87 is the comeback, shows are very different, most are tightly executed, very light on jams ... 88 starting to stretch back out a little." [23] Wolfman92097 also observes that in late 1986 "the setlists get much longer and way more experimental than they had been all year" and that in 1988 "jazz and extreme psychedelia gets added in." [22] Many deadheads' favorite years are 73/74, "the peak of their spacey, jazzy, psychedelic extended jams" according to devlinontheweb [30]. These observations again directly correspond to the plot in Figure 5. Song duration peaks in 1973, decreases throughout the late 70s and the early 80s with a turnaround point around 87/88, as perceived by MrCompletely.

Figure 6 shows plots for dynamics features. We can observe a long peak in overall loudness of the soundboard recordings, spanning the entire 1980s. With increasing loudness we see a decrease in dynamic range, which may hint at an increased use of compression in the live mix. However, when comparing with Figure 5, we can also see a striking correlation between loudness and tempo, as well as between dynamic complexity and duration. Dynamic complexity may be another indicator of the amount of improvisation in the recordings, similar to song duration. The latter two, however, seem to be somewhat independent nonetheless. We can particularly observe a bulge in dynamic complexity in the late 70s which does not occur in the duration curve, and in the 90s duration increases while dynamic complexity decreases. On the other hand increased loudness may be a direct consequence of higher tempo and energy. EvilLinux admits that "If I am alone in the car, I usually will choose the 80's. There is just more energy, more off the rails." [31] An anonymous deleted user also says that "the Dead did play some incredible

[12] https://www.reddit.com/r/gratefuldead/, https://www.reddit.com/r/grateful_dead/

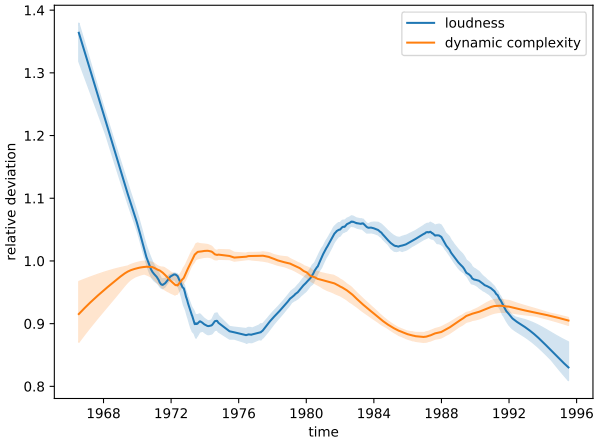

Figure 6: Bootstrapped LOWESS curves for dynamic essentia features (*loudness_ebu128.short_term.median* and *dynamic_complexity*) across all songs.

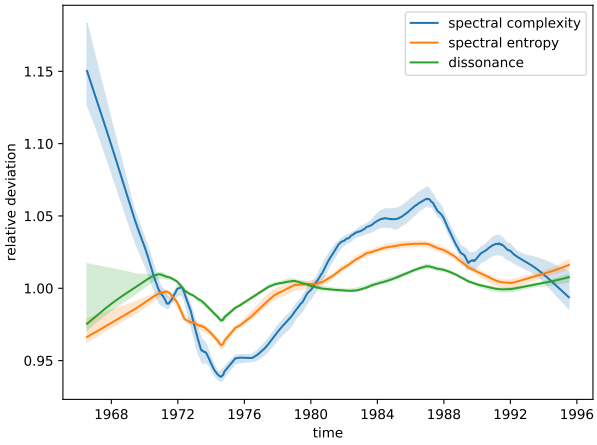

Figure 7: Bootstrapped LOWESS curves for spectral essentia features (*spectral_complexity.median*, *spectral_entropy.median* and *dissonance.median*).

shows in the early to mid-80s. They had a "fatter" sound, especially Jerry" [24].

A similar observation can be made for spectral features (Figure 7). Complexity and entropy increase for the entire duration of the 1980s, while dissonance has a shorter peak in the late 80s. Many listeners agree that the 80s have an entirely different sound, to a large part characterized by the keyboarder Brent Mydland who played with the band from 1979 to 1990 and who used a greater variety of keyboards and many synthesizers. According to BeaverMartin, Brent "really adds a whole different texture to the vocals and keys," [28] and WesternEstatesHOA says that "In 1983 the band truly takes off. The physical change of Brent's new keyboard is enough to change the band's sound alone. It has deep watery effects and adds so much depth to some of the more simple tunes." [22] Wolfman92097: "83 Garcia [guitar] is a little more distorted and Brent is using more fake keyboard sounds Phil [bass] is loud." [21]

As for tonal content, Figure 8 shows a selection of features calculated as described in Section 3.1. Tonal, pitch,

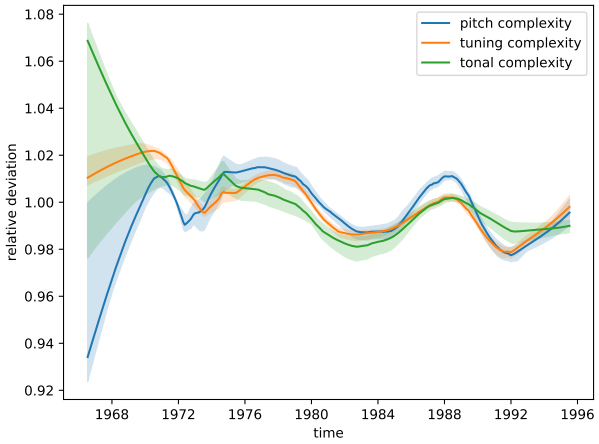

Figure 8: Bootstrapped LOWESS curves for pitch, tuning and tonal complexity features.

and tuning complexities run roughly in parallel and the period of 1980-84 is again demarcated, with a clear dip in all three features. This may again be due to the more streamlined faster performances and the lower degree of improvisation during this period. The late 1980s show parallels with the late 1970s, which may be related to the band starting to improvise more again. However, pitch and tuning complexities also seem highly correlated with dissonance (Figure 7) which may be related to timbre and the typical rich distorted and synthesizer-heavy 1980s Grateful Dead sound.

5. CONCLUSION

We have shown how evolutionary methods can not only be used for the study of general cultural trends in music, but also to investigate how the performances of a band change over time. We have also seen how we can improve the prediction accuracy of musical trends by considering the feature values relative to subsets of the data. It may be possible to apply a similar method in other situations, such as when studying the simultaneous evolution of the music of different composers, considering the music of each of them relative to their own work. We have also discovered limitations with the dataset used and suggest, in future work, to design a larger more systematic one in order to confirm our preliminary discoveries in this paper. Finally, while we have shown the potential reliability of the accounts of experienced listeners in the community, a more systematic collection and processing of online forum data may lead to more detailed results in the future.

6. ACKNOWLEDGMENTS

This work is supported in part by JST PRESTO No. JPMJPR20CB and JSPS KAKENHI Nos. 19H04137, 20K21813, 21K02846, 21K12187, 22H03661.

7. REFERENCES

[1] Y. Liu, Q. Xiang, Y. Wang, and L. Cai, "Cultural style based music classification of audio signals," in *2009*

IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009, pp. 57–60.

[2] D. Moelants, O. Cornelis, and M. Leman, "Exploring african tone scales," in *10th International Society for Music Information Retrieval Conference (ISMIR-2009)*. International Society for music Information Retrieval, 2009, pp. 489–494.

[3] M. Panteli, E. Benetos, and S. Dixon, "A computational study on outliers in world music," *Plos one*, vol. 12, no. 12, p. e0189399, 2017.

[4] C. Weiß, S. Balke, J. Abeßer, and M. Müller, "Computational corpus analysis: A case study on jazz solos." in *ISMIR*, 2018, pp. 416–423.

[5] F. Thalmann, K. Yoshii, T. Wilmering, G. A. Wiggins, and M. B. Sandler, "A method for analysis of shared structure in large music collections using techniques from genetic sequencing and graph theory," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 2020.

[6] J. Serrà, Á. Corral, M. Boguñá, M. Haro, and J. L. Arcos, "Measuring the evolution of contemporary western popular music," *Scientific reports*, vol. 2, no. 1, pp. 1–6, 2012.

[7] M. Mauch, R. M. MacCallum, M. Levy, and A. M. Leroi, "The evolution of popular music: Usa 1960–2010," *Royal Society open science*, vol. 2, no. 5, p. 150081, 2015.

[8] E. Deruty and F. Pachet, "The mir perspective on the evolution of dynamics in mainstream music," in *Proceedings of the 16th ISMIR Conference*, vol. 723, 2015.

[9] C. Weiß, M. Mauch, S. Dixon, and M. Müller, "Investigating style evolution of western classical music: A computational approach," *Musicae Scientiae*, vol. 23, no. 4, pp. 486–507, 2019.

[10] T. Parmer and Y.-Y. Ahn, "Evolution of the informational complexity of contemporary western music," *arXiv preprint arXiv:1907.04292*, 2019.

[11] K. Choi and J. Stephen Downie, "A trend analysis on concreteness of popular song lyrics," in *6th International Conference on Digital Libraries for Musicology*, 2019, pp. 43–52.

[12] E. Nakamura and K. Kaneko, "Statistical evolutionary laws in music styles," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[13] F. C. Moss, M. Neuwirth, and M. Rohrmeier, "The line of fifths and the co-evolution of tonal pitch-classes," *Journal of Mathematics and Music*, pp. 1–25, 2022.

[14] M. Benson, *Why the Grateful Dead Matter*. University Press of New England, 2016.

[15] M. C. Tift, "Grateful dead musicking," in *All Graceful Instruments: The Contexts of the Grateful Dead Phenomenon*. Cambridge Scholars Publishing, 2021.

[16] C. Weiß and M. Müller, "Quantifying and visualizing tonal complexity," in *Proceedings of the Conference on Interdisciplinary Musicology (CIM)*, 2014, pp. 184–187.

[17] K. R. Page, S. Bechhofer, G. Fazekas, D. M. Weigl, and T. Wilmering, "Realising a layered digital library: exploration and analysis of the live music archive through linked data," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 2017, pp. 1–10.

[18] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American statistical association*, vol. 83, no. 403, pp. 596–610, 1988.

[19] D. Malvinni, *Grateful Dead and the Art of Rock Improvisation*. Scarecrow Press, 2013.

[20] O. Longcroft-Wheaton, "The stylistic development of the grateful dead: 1965-1973." Ph.D. dissertation, University of Surrey, 2020.

[21] "Can you tell when a song was performed just by hearing it?" https://www.reddit.com/r/gratefuldead/comments/80ifqt/can_you_tell_when_a_song_was_performed_just_by/, accessed 2022-08-30.

[22] "Describing each year of the 80's," https://www.reddit.com/r/gratefuldead/comments/7x5zs9/describing_each_year_of_the_80s/, accessed 2022-08-30.

[23] "1980's Era Dead," https://www.reddit.com/r/gratefuldead/comments/1vf47n/1980s_era_dead/, accessed 2022-08-30.

[24] "Intro to listening to the Dead!" https://www.reddit.com/r/gratefuldead/comments/3crjek/intro_to_listening_to_the_dead/, accessed 2022-08-30.

[25] "Listened to all the shows. Listened to the best again. Here's the top 50 or so," https://www.reddit.com/r/gratefuldead/comments/sny6dx/listened_to_all_the_shows_listened_to_the_best/, accessed 2022-08-30.

[26] "Grateful Dead Reviews," https://raw.githubusercontent.com/maximinus/grateful-dead-reviews/master/dead_reviews.txt, accessed 2022-08-30.

[27] "When did Eyes get slowed down?" https://www.reddit.com/r/gratefuldead/comments/vax00p/when_did_eyes_get_slowed_down/, accessed 2022-08-30.

[28] "Fans of the 80s-90s - what do you like so much about this era?" https://www.reddit.com/r/gratefuldead/comments/sdufh5/fans_of_the_80s90s_what_do_you_like_so_much_about/, accessed 2022-08-30.

[29] "What are the fastest versions of Dead songs?" https://www.reddit.com/r/grateful_dead/comments/57btul/what_are_the_fastest_versions_of_dead_songs/, accessed 2022-08-30.

[30] "Best decade and year?" https://www.reddit.com/r/gratefuldead/comments/pdey6g/best_decade_and_year/, accessed 2022-08-30.

[31] "Can we just cut the shit and agree that the best Dead is from the 70s? Who's with me?" https://www.reddit.com/r/gratefuldead/comments/rblwiy/can_we_just_cut_the_shit_and_agree_that_the_best/, accessed 2022-08-30.

EVALUATING GENERATIVE AUDIO SYSTEMS AND THEIR METRICS

Ashvala Vinay, Alexander Lerch
Center for Music Technology
Georgia Institute of Technology
ashvala@gatech.edu

ABSTRACT

Recent years have seen considerable advances in audio synthesis with deep generative models. However, the state-of-the-art is very difficult to quantify; different studies often use different evaluation methodologies and different metrics when reporting results, making a direct comparison to other systems difficult if not impossible. Furthermore, the perceptual relevance and meaning of the reported metrics in most cases unknown, prohibiting any conclusive insights with respect to practical usability and audio quality. This paper presents a study that investigates state-of-the-art approaches side-by-side with (i) a set of previously proposed objective metrics for audio reconstruction, and with (ii) a listening study. The results indicate that currently used objective metrics are insufficient to describe the perceptual quality of current systems.

1. INTRODUCTION

There has been growing research interest in building deep learning models that are capable of generating audio. Models such as WaveNet [1], NSynth [2], WaveGAN [3] and, most recently, DDSP [4] paved the way for data-driven Neural Audio Synthesis (NAS). Models like DDSP and NSynth have been advertised in YouTube videos prominently [5, 6], where artists make music using neural network generated audio, indicating that there is real world applicability to generative audio models.

Despite the many advances in generative modeling of sounds in the past couple of years, the evaluation of these systems lacks established methodology. Most importantly, systems are not evaluated with a consistent set of metrics. For instance, researchers have suggested the following to evaluate the output of such generative systems: the classification accuracies of neural networks trained to classify the sounds into predefined categories such as pitch and sound qualities [2,3,7,8], statistical methods [9,10], and subjective evaluation through listening studies [3, 9, 10].

This inconsistency in metrics and methodology makes a direct comparison of systems difficult at best, making it hard to understand the current state of the art and to measure the impact of new innovations in the field.

In this study, we measure and compare the quality of the output of neural networks capable of producing short time-invariant samples. The goal is to evaluate state-of-the art systems comparatively with previously used evaluation metrics and to investigate the perceptual relevance of these metrics for measuring audio quality.

To pursue these goals, we trained three widely known neural networks, DDSP [4], NSynth [2], and Diffwave [9]. The evaluation results published with the introduction of these methods all imply that these models synthesize sounds at high quality. However, since different metrics are used for each system, no comparison is possible. To enable such a comparative analysis, we survey and implement a set of metrics and apply them to these systems. Furthermore, we conduct a listening study to measure the perceptual sound quality of the system outputs.

The core contributions of this paper are: (i) a review of currently used metrics for the evaluation of synthesis quality and a comparative analysis of 3 popular neural audio synthesizers, (ii) a listening study for assessing the perceptual audio quality of these synthesizers, and (iii) an investigation on the perceptual relevance of the objective metrics.

2. EVALUATION OF NAS SYSTEMS TODAY

Generative systems are notoriously difficult to evaluate [11] and new metrics are proposed frequently to understand whether generative networks are able to capture desirable characteristics required for a given task. In audio and music, the task of evaluation is difficult because (i) the ground truth is not well defined, as various outputs might be considered "correct," [12] (ii) the previously used set of objective metrics for NAS most likely misses or insufficiently models perceptual qualities of a sound [7, 13], (iii) and aesthetic preferences are, by definition, subjective. [14]

Some contemporary metrics for NAS derive from literature on evaluating Generative Adversarial Networks (GANs), where diversity of samples and modeling the distribution of data play a significant role [15]. Objective evaluation metrics typically rely on either mathematical formulations of a success measure, or —in the case of contemporary GAN literature— using a separate neural network to identify if the model is working appropriately. Accordingly, we categorize the metrics into the four groups (i) reconstruction metrics, (ii) sample diversity measures,

(iii) distribution distance measures, (iv) and measures derived from subjective evaluation methods.

Reconstruction errors are computed as the difference between a given input sound S_i and a generated sound S_g. The error is typically defined as either an ℓ_2 norm (squared error) or a ℓ_1 norm (absolute error). These differences can be computed on either the time-domain signal or a spectrogram. Typically, the MSE/MAE is computed on spectrograms, given their ubiquity as the input and generated representation for neural networks. MSE and MAE have a range between 0 and infinity, with a MSE/MAE of 0 indicating perfect reconstruction.

To give examples of practical use, Engel et al. use the multi-scale spectrogram loss, which compares spectrograms across a set of FFT sizes as a measure of reconstruction error and as a loss term for use in training [4]. This was recently used as a metric by Shan et al. to compare DDSP with their proposed Differential WaveTable Synthesis [16].

With respect to reconstruction metrics, there appears to be an ambiguity about the purpose of these metrics, as they seem to be used as both the training loss to be minimized and the evaluation metric itself. Also note that these errors are known to not have a clear perceptual meaning, as a high MSE between two sounds does not necessarily mean that such two sounds will be perceived as dissimilar (or vice versa). There is plenty of evidence in the field of (perceptual) audio coding verifying that measuring the power of the coding error is insufficient to capture the perceptual quality of the sound [17].

Sample diversity metrics: In GAN literature, a lot of emphasis is placed on the performance of the "generator" and to ensure that it is able to produce classifiable samples that capture the diversity of classes from the training dataset. The two metrics we will discuss in this section use machine learning and deep learning driven approaches to measure sample diversity.

- **Number of statistically Different Bins** (NDB/k) is a metric devised to identify mode collapse in GANs, a phenomena where a network produces a lot of outputs that look like alike, therefore lacking sample *diversity* [18]. NDB/k is computed on a Voronoi decomposition from the k-means centroids of the training samples. The clusters are computed directly in the sample space. [1] The k clusters are referred to as the "bins." To compute a score, test samples are assigned to the k clusters/bins using an $L2$ distance measure between the samples and the centroids of the clusters. A two-sample t-test on each bin identifies the statistically different bins. The final NDB score is given by counting the number of statistically different bins and dividing by the number of clusters.
  NDB/k scores are between the range of 0 and 1. According to the interpretation of the score provided by Richardson and Weiss [18], a score of 0 indicates that the network is producing a large diversity of samples that captures the training distribution well, whereas a score approaching 1 suggests that the generator has collapsed.
  This was used by Diffwave [9] and GANSynth [10] as part of their evaluation metrics.
  Richardson and Weiss [18] state in their definition of the metric that computing the distance for each pixel in an image using an $L2$ distance is perhaps not meaningful. As we have stated above, distances like $L2$ are not good at capturing perceptual qualities of sound, making this of particular concern when evaluating audio and the outputs of generative audio systems.
- **Inception scores (IS)** were proposed by Salimans et al. as a way to evaluate image generating GANs by using a classifier [15]. This classifier is used to automatically evaluate whether the output of a GAN was of reasonable quality and captured the *diversity* of samples in the dataset.
  The Inception classifier produces class label probabilities for a given input. Ideally, each input produces a high probability for one class label and our generative system is able to produce many of them. At the same time, the generator should be able to produce many such images, that can be classified uniquely into a large set of labels. If the difference between the probability distribution of predicted labels for the generated images and the marginal distribution of the labels from the generated data is small, it implies that the the generator is unable to produce a a diverse number of easily classifiable images. The mathematical formulation of IS can be found in [15]. The score itself has a lower bound of zero and an upper bound of infinity. The higher the score, the better.
  Within the context of audio, IS was used in evaluating WaveGAN [3], where an Inception network was trained on the SC09 dataset with spectrogram inputs. More recently, two inception scores have been proposed for NAS: the Pitch Inception Score (PIS) and the Instrument Inception Score (IIS) [8]. These measures tell us if generator has captured the true distribution of discrete MIDI pitch classes and the instrument classes in the NSynth dataset [2], and are thus focused on inherent sound properties that are not directly related to audio quality.
  Salimans et al. [15] stated that the Inception Score for an image generator was found to correlate well with human judgment of image quality [15]. However, no such work has been done in evaluating the perceptual meaningfulness of IS with audio.

Distribution distance metrics: Another important facet of evaluating GANs is identifying whether the distribution of data produced by the generator is close to the distribution of real data. Like the Inception score mentioned above, the metrics discussed here use neural networks. The difference here is that these metrics rely on using embeddings from a neural network.

As noted by Ananthabhotla et al. [13], matching distributions cannot guarantee a perceptually closer result.

- **Kernel Inception Distances (KID)** are scores that

[1] in the case of images, the pixel distance