EVALUATING GENERATIVE AUDIO SYSTEMS AND THEIR METRICS

Ashvala Vinay
Center for Music Technology
Georgia Institute of Technology

Alexander Lerch
Center for Music Technology
Georgia Institute of Technology

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 are generated by computing the distance between embeddings of input and generated data fed to inception networks. The distance is computed using Maximum Mean Discrepancy (MMD), a statistical

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

test that describes the difference between two distributions of data. They were first introduced in a paper by Binkowski et al. [19] where they used MMD to train a critic or discriminator and KID was shown as a metric to evaluate the convergence of the GAN. In order to compute KID, we compute a distribution of embeddings extracted from the Inception classifier for both the reference and generated output and compute the MMD between the distributions of reference and generated embeddings. The score is defined with a lower-bound of zero and an upper bound of infinity. This was used by Nistal et al. to compare input feature representations using GANs [7] and in a separate paper by Nistal et al. to measure the performance of an architecture they built called the VQCPC-GAN [8]. It was also used to evaluate DarkGAN [20].

- **Fréchet Audio Distance (FAD)** [21] is a metric originally developed to evaluate sound enhancement algorithms, but recently it has found use in evaluating NAS systems. The computation of the FAD relies on the VGGish embeddings for both the reference and the generated sounds. The VGGish embeddings are then fitted to multi-variate gaussians. The FAD itself is the Fréchet distance between the two distributions $\mathcal{N}_r$ and $\mathcal{N}_g$ representing the reference and generated gaussian distributions. The mathematical definition can be found in [21].
  This metric was used for evaluating Diffwave [9], Neural Waveshaping Synthesis [22], DarkGAN [20], and CRASH [23].

2.1 Subjective evaluation

Since the quality of the generated outputs is ultimately a perceptual property, listening studies have been previously used to evaluate a neural network's generated audio quality. A lot of listening studies use a popular method called "Mean Opinion Score" (MOS) [24]. Users participating in the listening study are asked to rate the sound they hear on a Likert scale between 1 and 5 across a set of questions. For example, WaveGAN asked participants to rate sounds on their "sound quality, ease of intelligibility, and speaker diversity" [3] where 1 indicates bad and 5 indicates excellent. These questions are asked for a collection of methods that the researcher seeks to evaluate.

The MOS survey strategy has been used for the evaluation of WaveGAN [3], Diffwave [9], and in neural speech generation literature [25, 26]. Kong et al. compared their results to WaveGAN using MOS and found that their network scored higher [9].

It should be noted that surveys that use MOS do not explicitly ask participants to compare the outputs against a reference and instead present participants with a single sound for every question. This means that MOS surveys give you an absolute rating for every sound, not a relative preference. The absence of reference precludes the ability to rank the methods that are being evaluated. For example, a person might like sound A and sound B in isolation and provide high ratings to both, however, it remains unclear whether the person has a relative preference for sound A or B.

A well known alternative to MOS surveys is to use a survey method called MUltiple Stimuli, Hidden Reference and Anchor (MUSHRA) [27] . It is an ITU recommendation for studies designed to evaluate differences in audio quality between audio codecs [28]. While it hasn't seen significant usage in evaluating NAS, it was recently used by Hayes et al. [22] in evaluating the performance of their Neural Waveshaping Synthesis (NeWT) model against the performance of DDSP. Their listening study results showed that their neural network outperformed DDSP on most instruments, with the exception of violins. This was shown to correlate well with the computed FAD scores for DDSP and NeWT.

3. EXPERIMENTAL SETUP

Aiming at our goal of comparing the output quality of popular generative systems and assessing the metrics commonly used for evaluation, we chose three neural networks and re-trained DiffWave and DDSP with the NSynth dataset and used the NSynth network directly. We report both objective metrics and subjective ratings of the outputs of these system.

We break our study into the three phases: (i) comparative analysis using objective metrics, (ii) comparative analysis using listening study results, and (iii) a brief investigation into the perceptual relevance of the objective metrics.

3.1 Dataset

The NSynth dataset is a publicly available dataset with approximately 300k sounds of 4 s length spanning acoustic and electronic timbres [2]. The sounds are all sampled at 16 kHz. It is comprised of 11 instrument families. There are ten unique "quality" descriptors that are attached to every sound in the dataset. The descriptors are timbral, for e.g "dark," "bright," "reverb," and "percussive." The NSynth dataset is frequently used in generative audio research and is therefore an obvious choice for this study.

The NSynth dataset's test set was used to generate all the samples that were used in both the listening study and to compute objective metrics.

3.2 Models

Model selection was based on the criteria of age, architecture, and public availability. There are a number of generative audio systems that have been published since 2017, but not all of them were available publicly or were difficult to train due to the lack of computational resources. Models selected were NSynth [2], DDSP [4], and DiffWave [9]. These models are widely known and represent different architectural designs.

NSynth [2] is a deep learning based generative audio system that uses a Variational Autoencoder architecture that learns to generate musical instrument timbres with the ability to be controllable. Its primary novelty is the fact that it can interpolate between multiple sounds and generate new timbres in the process. NSynth is the "oldest" of the

System	NDB/k (↓)	PKID(↓)	IKID(↓)	PIS(↑)	IIS(↑)	MSE(↓)	MAE (↓)	FAD (↓)
Diffwave	0.74	0.0093	0.0021	2.3814	**5.6477**	0.0291	0.1369	7.9488
DDSP	**0.20**	**0.0053**	0.0020	**3.3224**	5.3371	**0.0130**	**0.0666**	**1.1519**
NSynth	0.74	0.0101	0.0024	2.3238	4.6364	0.0329	0.1224	4.0590
Anchor	0.72	0.0123	**0.0006**	2.9356	5.3017	0.0257	0.0857	1.4952

Table 1. Table with objective results for each of the neural networks that we measured. ↓ indicates that a lower score is better and ↑ indicates that a higher score is considered better. Bold indicates best performance.

evaluated systems. The publicly available weights for the NSynth model were used for this study to generate the sounds.

DDSP [4] uses classical synthesis techniques like additive synthesis and noise filtering in the context of deep learning by treating them as differentiable blocks. It uses an encoder-decoder architecture that produces the fundamental frequency, loudness envelope and the necessary variables to control the additive synthesizer. DDSP is also known for it ability to perform "timbre-transfer," where it takes sounds produced from one instrument and outputs it on a different instrument. To train DDSP, we used publicly available code for training with the NSynth dataset and verified it worked by producing metrics similar to the metrics reported in the original paper.

Diffwave [9] is the most recent system that uses a new generative modeling technique called "Diffusion" on audio. Diffusion models start with white noise and through a fixed number of iterations, learn to generate audio. During training, the models use a forward and backward process, where the forward process takes the reference, corrupts it with white noise iteratively until the whole signal is noise. The backwards process learns how to iteratively remove the white noise to recover the reference. Diffwave was primarily evaluated on speech and produced results that seemed to outperform WaveGAN and Wavenet significantly.

DiffWave was trained with an implementation available online [2]. Since this algorithm was not originally trained with NSynth, we had to verify if our model worked properly. We trained the network for 2 million steps and evaluated the network's automatic metrics and found that the metrics aligned with the paper's reported results.

3.3 Objective metrics

To evaluate our neural networks, we used the set of metrics described above. As we discussed in Section 2, some of the objective metrics were designed with specific architectures like GANs in mind. However, since the metrics have found use in the evaluation of other types of architectures we chose to evaluate all of our networks with the same set of metrics.

The pre-trained pitch and instrument inception networks trained on the NSynth dataset from Nistal et al.'s paper on comparing audio representations [7] were used to compute neural network driven metrics such as as PKID and IIS and the "official" implementations of NDB/k [3] and FAD [4] were used. In order to compute NDB/k, we trained the K-means clustering on the NSynth dataset.

The rest of the metrics, like MSE and MAE were computed using SciKit-learn's built-in metrics [29]. We tried to include other metrics but could not select them due to high variance in scoring between implementations (such as PEAQ [17]).

To summarize, the following objective metrics are computed for the network outputs:

1. NDB/k, 2. PKID, 3. IKID, 4. PIS, 5. IIS, 6. MSE, 7. MAE, and 8. FAD.

3.4 Listening study

The listening study uses a variation of the aforementioned MUSHRA methodology. It is popularly used in evaluating "intermediate" differences in low bitrate audio codecs [28]. Given that we are measuring audio reconstruction, we believe that treating the neural network outputs similar to encoded audio is a suitable choice for evaluating audio quality differences. The rating scale used in the study is divided according to the MUSHRA specification. A score between 0 and 20 indicates that the sound is rated *bad*, 20 to 40 indicates that the sound is rated *poor*, 40 to 60 is considered *fair*, 60 to 80 is considered *good*, and 80 to 100 is considered *excellent*.

The 4096 sounds from the NSynth test set are reconstructed using the three generative systems. In addition, one anchor sound is generated for each test sample by low pass filtering the sound at a 1 kHz cutoff frequency and reducing its bit depth to 8 Bits.

Participants were recruited by emails sent to two large academic audio-focused communities. Participants were asked to use a good pair of headphones or speakers in order to participate in the survey. Prior to the listening study, the participants were asked to share their age bracket, experience with audio synthesis, and how much money they have spent on the audio equipment they used. This then led to a training phase where participants were presented with an example sound and necessary introduction to the survey.

When presenting the survey, a sound is randomly selected out of 3237 possible sounds with MIDI note numbers ranging from 22 to 84. [5] Five sliders are presented in random order, with three sliders referring to the generated audio output of the three neural networks to evaluate, and the other two sliders corresponding to the hidden reference and the anchor. The participants were asked to rate audio generated by each neural network on its audio quality

[2] https://github.com/lmntcom/diffwave

[3] https://github.com/eitanrich/gansngmms

[4] https://github.com/googleresearch/googleresearch/

[5] We used the note number range to remove sounds that were either inaudible or could potentially be uncomfortable to listen to.

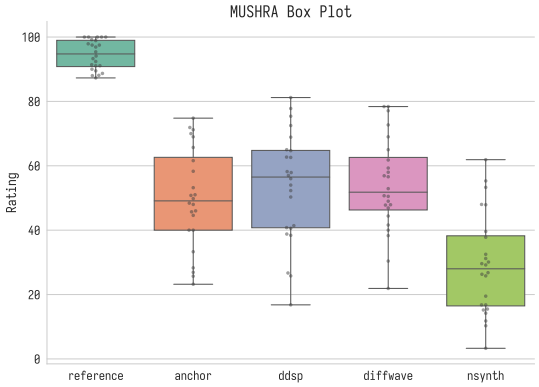

Figure 1. A boxplot of ratings from participants in the listening study. The size of the box represents the interquartile range of ratings from the listeners.

compared to the reference. This presentation is repeated ten times without collision (i.e., no two sounds are ever repeated for a participant).

To clean up the data, responses that are "incomplete" are removed entirely from data that can be used for analysis. MUSHRA analysis generally removes raters who rate the reference below a high rating threshold [30]; thus, participants who rate the reference below an average of 85 will be removed.

Statistical analysis of MUSHRA and MUSHRA-like data relies on running statistical tests such as ANOVA or Wilcoxon tests [30] to measure differences in results between the presented conditions. We will use the Wilcoxon tests for the ratings to measure statistical differences between presented conditions. The test indicates differences between the generative models. The MUSHRA results will also be broken down by demographic information collected above to measure if different demographic groupings rated the models differently.

In order to compare objective and subjective ratings, we will also compute rankings for both ratings. We are interested in knowing how frequently specific networks were considered the best and comparing it to the rankings for each metric. Rank driven correlations were used by Ycart et al. [31] in their paper validating perceptual metrics related to piano transcription.

4. RESULTS

4.1 Objective metric results

The results from the objective metrics as discussed above are shown in Table 1. Overall, these results seem to indicate that DDSP is producing higher quality samples that are more easily classifiable compared to DiffWave and NSynth.

DDSP has a 54% better PKID score over Diffwave and a 62% higher score than NSynth. This trend continues across most metrics, with the exceptions of the IIS, where Diffwave achieves a 5% higher score than DDSP, and the IKID where there is only a 4% difference between DDSP and Diffwave. The kernel distances mentioned here do not measure audio quality.

The MSE and MAE results tell us that DDSP is the "closest" to the reference sounds, with an MSE of 0.0130, a 76% and 79% improvement over Diffwave and NSynth, respectively.

DDSP significantly outperforms Diffwave and NSynth on the FrÃ©chet Audio Distance. FAD is also a category where NSynth outperforms Diffwave. This metric indicates that —according to the VGGish representation— DDSP and NSynth are generating samples that are closer to the reference than Diffwave.

Based on these metrics, we can state that DDSP outperforms Diffwave and NSynth and produces a diverse set of easily classifiable samples and produces samples that are much closer to the reference sounds.

4.2 Listening study results

Data was collected from 77 participants from our listening study. After data cleanup, a total of 24 submissions were considered trustworthy. 74% of our participants were between the age of 24 and 50, 22% of our participants between 18 and 24 and 3.7% or 1 participant over the age of 50. 37% of our participants reported that they were very familiar with music production tools and 18.5% reporting that they were extremely familiar. 29% of the participants reported that they were moderately knowledgeable about sound synthesis techniques, with an even distribution of familiarity ranging from "Not very knowledgeable" to "Extremely knowledgeable." 33% of our raters reported that they had spent over $750 on their audio equipment and 25% having spent between $250 and $500.

The overall visualization of the responses can be seen in Fig. 1. Every dot in the chart is an averaged rating from a participant. As expected, the reference scored highly with an average rating of 92. The analysis of the listening study results shows that DDSP and DiffWave scored similarly while NSynth performed considerably worse than the other networks. Both DDSP and Diffwave have average ratings around 53 while sounds generated by NSynth received an average rating of approx. 29.

The inter-rater Krippendorff α score [32] is 0.66 [6], suggesting that the subjects were largely in agreement.

The Wilcoxon test for statistical significance was applied to the data [30]. We found no statistically significant difference between DDSP/Diffwave ($p = 0.629$) and a statistically significant difference between DDSP/NSynth ($p = 8 \times 10^{-6}$) and Diffwave/NSynth ($p = 8 \times 10^{-6}$). There was a significant difference between the Reference and all the systems and the anchor ($p \leq 8 \times 10^{-6}$). There is a significant difference between Anchor/NSynth ($p = 8 \times 10^{-6}$), but no statistically significant difference between Anchor/Diffwave or Anchor/DDSP, with p-values of 0.144 and 0.4385, respectively.

Breaking down the results by the subjects' self-reported familiarity with audio synthesis technologies, we first investigate the inter-rater variation within different "expert levels."

[6] Krippendorff α ranges from -1 to 1, where -1 indicates significant disagreement between raters and 1 suggests maximal agreement

Participants who said they were the least knowledgeable had an Krippendorff α of 0.7, moderately knowledgeable raters had a Krippendorff α of 0.608 while participants who were the most knowledgeable had a Krippendorff α of 0.9. Participants who reported to be either very or extremely knowledgeable tended to rate DDSP and Diffwave lower than the overall scores (DDSP: 47, Diffwave: 43). Less knowledgeable participants tended to rate Diffwave and DDSP higher than the overall scores (DDSP: 56, Diffwave: 59). There was a statistically significant difference in how these two groups rated Diffwave ($p = 0.006$) but no statistically significant difference in how they rated DDSP, NSynth, or the reference and anchor point.

We found no statistically significant differences in ratings between the age categories or in ratings based on money spent on audio equipment.

Instrument results: To investigate whether specific instruments or instrument groups are consistently rated higher or lower than others, the listener ratings were broken down into ratings by instrument family (according to the NSynth dataset). We found no statistically significant differences in ratings between DDSP and Diffwave, but found statistically significant differences between DDSP/NSynth and Diffwave/NSynth. Additionally, there were no statistically significant differences between instrument sources and the ratings from the listeners.

The IKID metric tell us that all three networks are producing samples that are close to the reference and the IIS score tells us that Diffwave should be slightly better than DDSP and much better than NSynth. While we cannot state that this result is accurate for IKID, it is in line with the results from the IIS computation.

Comparing the listener ratings to objective metrics: In order to identify if our listening study results lined up with our objective metrics, we took the selection of sounds that were presented to participants and computed the correlation between their sample based metrics and our ratings using the Pearson correlation coefficient, which tells us how correlated two samples are from a range of -1 to 1. We have sample based results for MSE and MAE, since all the other metrics in our list rely on computing a difference across a distribution of samples. We also computed the Spearman R score and the R^2 from a linear regression between the ratings and the MSE and MAE. We found that there is no statistically significant correlation between the MSE/MAE errors and the listener ratings.

The listening responses were also ranked based on how frequently a specific network was ranked higher than the others. In the 240 responses, the listeners rated Diffwave higher than DDSP slightly more frequently (123 vs. 99). NSynth was rated the lowest most frequently (163 times). The ranking breakdown showing how frequently each of the 6 permutation of rankings between the 3 networks is shown in 2. We ran a Wilcoxon test on pairings rankings of each network by the listeners and found that the rankings were different with high statistical significance ($p < 0.05$).

When ranking the objective metrics, it is clear that DDSP is ranked the best because it has the best results in seven out of the eight metrics, with Diffwave and NSynth in second and third place. However, the rankings from our listening study tell us that the Diffwave is usually the preferred network. This tells us that the metrics are perhaps not entirely reflective of the perceptual audio quality of the networks.

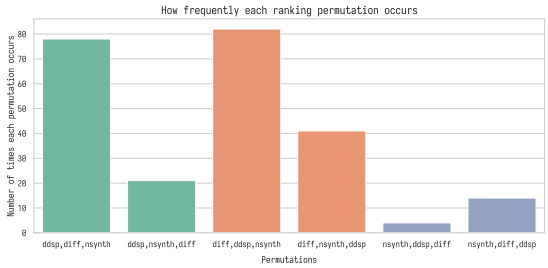

Figure 2. The plot here describes how frequently a unique permutation of rankings occurred. Diffwave is abbreivated to "diff".

5. CONCLUSION

We presented a systematic evaluation of three popular systems for neural audio synthesis, comparing them with a set of previously used objective metrics as well as a listening study. The results clearly show that (i) no previously used objective metric captures the perceptual quality of the synthesized sounds sufficiently well, (ii) any quality ranking based on these objective metrics is questionable, (iii) of the three evaluated audio generators, there is no clear listener preference between DDSP and Diffwave, but NSynth is rated lowly, and (iv) the subjects rate Nsynth worse than the 8 Bit anchor but rate DDSP and Diffwave similar to the anchor indicating that there is still considerable work to do on the quality of neural audio synthesis

These results should give pause to research in the field of neural audio synthesis. How can progress with respect to the audio quality be measured if all available metrics are unable to provide meaningful estimates of audio quality. Many of the objective metrics that we discussed in this paper were designed with the intent of measuring the performance of the sound generating component of the networks, i.e., can the generator produce (i) a diverse set of sounds, (ii) a distribution of samples that are close to the target samples in a relevant dimension, and (iii) accurate reconstructions.

Our results indicate that *measuring generator performance is insufficient to measure audio quality*. We believe that research in this space should not only include subjective results, it should also include greater efforts into critically evaluating the audio quality of network outputs with meaningful objective metrics.

In future work, we will investigate whether other objective metrics might be more meaningful than the ones evaluated here, or will start a research project on developing a more meaningful quality measure for audio synthesis.

6. REFERENCES

[1] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*. ISCA, p. 125. [Online]. Available: http://www.isca-speech.org/archive/SSW_2016/abstracts/ssw9_DS-4_van_den_Oord.html

[2] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, pp. 1068–1077, ISSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v70/engel17a.html

[3] C. Donahue, J. J. McAuley, and M. S. Puckette, "Adversarial audio synthesis," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. [Online]. Available: https://openreview.net/forum?id=ByMVTsR5KQ

[4] J. H. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr

[5] Adam Neely, "Turning BASS into violin (using AI)." [Online]. Available: https://www.youtube.com/watch?v=cIX4y22NWWc

[6] ANDREW HUANG, "Music with artificial intelligence." [Online]. Available: https://www.youtube.com/watch?v=AaALLWQmCdI

[7] J. Nistal, S. Lattner, and G. Richard, "Comparing representations for audio synthesis using generative adversarial networks," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 161–165, ISSN: 2076-1465.

[8] J. Nistal, C. Aouameur, S. Lattner, and G. Richard, "VQCPC-GAN: Variable-length adversarial audio synthesis using vector-quantized contrastive predictive coding," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 116–120, ISSN: 1947-1629.

[9] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. [Online]. Available: https://openreview.net/forum?id=a-xFK8Ymz5J

[10] J. H. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial neural audio synthesis," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. [Online]. Available: https://openreview.net/forum?id=H1xQVn09FX

[11] L. Theis, A. v. d. Oord, and M. Bethge, "A note on the evaluation of generative models," number: arXiv:1511.01844. [Online]. Available: http://arxiv.org/abs/1511.01844

[12] M. Caetano and N. Osaka, "A formal evaluation framework for sound morphing," in *ICMC*.

[13] I. Ananthabhotla, S. Ewert, and J. A. Paradiso, "Towards a perceptual loss: Using a neural network codec approximation as a loss for generative audio models," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, pp. 1518–1525. [Online]. Available: https://dl.acm.org/doi/10.1145/3343031.3351148

[14] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," vol. 95, no. 4, pp. 489–508. [Online]. Available: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/0007126042369811

[15] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training GANs," p. 9, QID: Q46993880.

[16] S. Shan, L. Hantrakul, J. Chen, M. Avent, and D. Trevelyan, "Differentiable wavetable synthesis," number: arXiv:2111.10003. [Online]. Available: http://arxiv.org/abs/2111.10003

[17] T. Thiede, "PEAQ–the ITU standardfor ObjectiveMeasuremenot f perceived audio quality," vol. 48, no. 1, p. 27.

[18] E. Richardson and Y. Weiss, "On GANs and GMMs," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/0172d289da48c48de8c5ebf3de9f7ee1-Paper.pdf

[19] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. [Online]. Available: https://openreview.net/forum?id=r1lUOzWCW

[20] J. Nistal, S. Lattner, and G. Richard, "DarkGAN: Exploiting knowledge distillation for comprehensible audio synthesis with GANs," in *Proceedings of the*

22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. v. Kranenburg, and A. Srinivasamurthy, Eds., pp. 484–492. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000060.pdf

[21] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "FrÃ©chet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Interspeech 2019*. ISCA, pp. 2350–2354. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2019/kilgour19_interspeech.html

[22] B. Hayes, C. Saitis, and G. Fazekas, "Neural waveshaping synthesis," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. v. Kranenburg, and A. Srinivasamurthy, Eds., pp. 254–261. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000031.pdf

[23] S. Rouard and G. Hadjeres, "CRASH: Raw audio score-based generative modeling for controllable high-resolution drum sound synthesis," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. v. Kranenburg, and A. Srinivasamurthy, Eds., pp. 579–585. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000072.pdf

[24] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2416–2419. [Online]. Available: http://ieeexplore.ieee.org/document/5946971/

[25] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621, ISSN: 2379-190X.

[26] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017*. ISCA, pp. 4006–4010. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2017/wang17n_interspeech.html

[27] International Telecommunications Union (last). BS.1534Â :Â method for the subjective assessment of intermediate quality level of audio systems. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534/en

[28] S. Zielinski, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential biases in MUSHRA listening tests." Audio Engineering Society. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=14237

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others, "Scikit-learn: Machine learning in python," vol. 12, pp. 2825–2830, QID: Q28365500.

[30] C. MendonÃ§a and S. Delikaris-Manias, "Statistical tests with MUSHRA data," in *Audio Engineering Society Convention 144*. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19402

[31] A. Ycart, L. Liu, E. Benetos, and M. T. Pearce, "Investigating the perceptual validity of evaluation metrics for automatic piano music transcription," vol. 3, no. 1, pp. 68–81. [Online]. Available: http://transactions.ismir.net/articles/10.5334/tismir.57/

[32] K. Krippendorff, "Computing krippendorff's alpha-reliability." [Online]. Available: https://repository.upenn.edu/asc_papers/43