Evaluating Generative Audio Synthesizers and their Metrics

Georgia Tech · College of Design
Center for Music Technology

Ashvala Vinay, Alexander Lerch
Music Informatics Group, Georgia Institute of Technology
ashvala@gatech.edu

Introduction and Motivation

1. In recent years, there has been growing interest in Neural Audio Synthesis (NAS). Models like NSynth [1], DiffWave [2] and DDSP [4] are popular examples of NAS systems.
2. Despite the advances, there is a lack of established evaluation methodology. There are several metrics but none of them are applied consistently.
3. This inconsistency makes comparing NAS systems difficult. This makes quantifying the state-of-the-art impossible.
4. Our core contributions are:
   1) A review of currently used metrics and comparitive analysis of 3 NAS systems
   2) A listening study for assessing the perceptual audio quality of these synthesizers
   3) An investigation of the perceptual relevance of objective metrics.

Objective Metrics

1. Objective metrics used today can be categorized into three groups: Reconstruction errors, Sample Diversity Methods, and Distribution Distances.
2. Reconstruction errors are computed as the difference between a given input sound and a reconstructed/generated version of the same sound.
3. Sample Diversity metrics typically focus on the performance of the generator:
   • **NDB/k** [6] is a metric devised to identify mode collapse in GANs.
   • **Inception scores** [7] are a popular method used to measure sample diversity.
      •It was applied to NAS systems by Nistal et al. as Pitch and Instrument Inception Scores [5].
4. Distribution distances measure the similarity in distances between the generated data and input data:
   • **Kernel Inception Distances** (KID) [8]: computed as the Maximum Mean Discrepancy (MMD) between the distribution of embeddings for the input and generated data derived from an Inception Network.
   • **Frechet Audio Distance** (FAD) [3] is a metric originally developed to evaluate sound enhancement algorithms.
      • The computation of FAD relies on using VGGish embeddings, fitting them to gaussians and computing a Fréchet distance between the two distributions.

Subjective metrics

1. A lot of studies use a popular method called Mean Opinion Score surveys (MOS).
   • Users are asked to rate the sound they hear on a Likert scale between 1 and 5 on a variety of questions.
   • MOS surveys do not explicitly ask participants to compare outputs against a reference.
   • MOS gives you an absolute rating for every sound, not a relative rating.
2. A well known alternative is the ITU recommendation called MUltiple Stimuli Hidden Reference and Anchor (MUSHRA).
   • It is commonly used in evaluating audio codecs.
   • Anchor sounds are a worse version of the reference

Experimental Setup

1. To compare 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.
2. We report both objective metrics and subjective ratings of the outputs of these systems.
3. 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.
4. Our subjective ratings were collected using a MUSHRA study. The interface used can be seen in Fig 1.

Results

The results from the objective metrics can be seen in Fig 2. It shows that DDSP is the best performing network of the three.

The listening study results are shown in Fig 3.
• We had a total of 24 valid participants for our survey over a total of 77 total participants.
• DDSP and Diffwave were rated similarly while NSynth was rated the worst.
• When looking at the ranking permutations, we noticed that people selected Diffwave as the best network more frequently than DDSP.

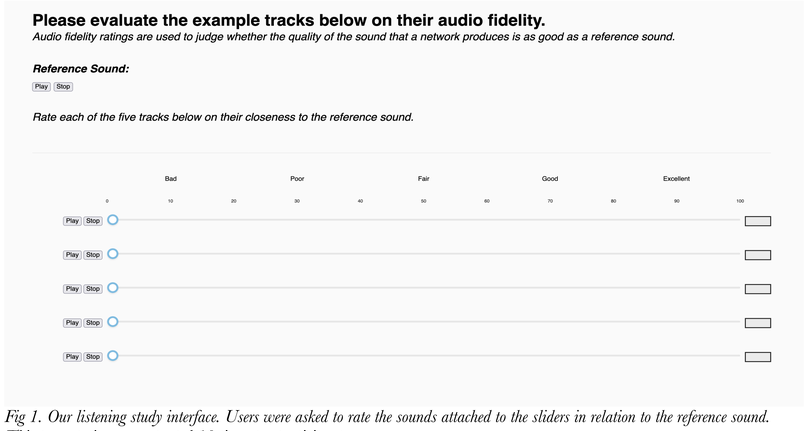


Fig 1. Our listening study interface. Users were asked to rate the sounds attached to the sliders in relation to the reference sound. This presentation was repeated 10 times per participant.

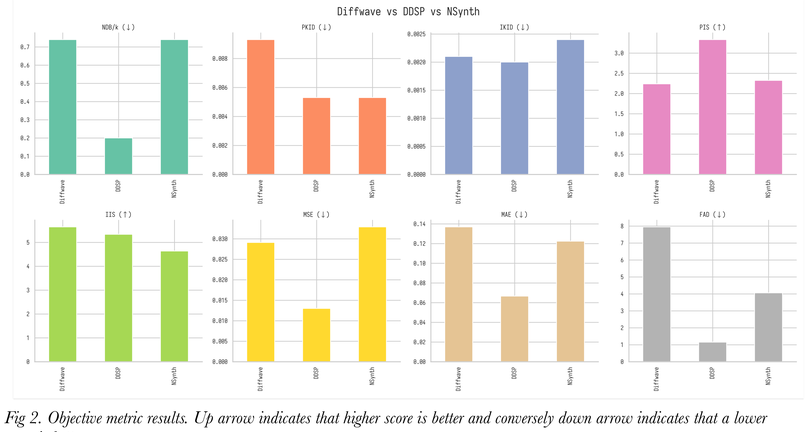
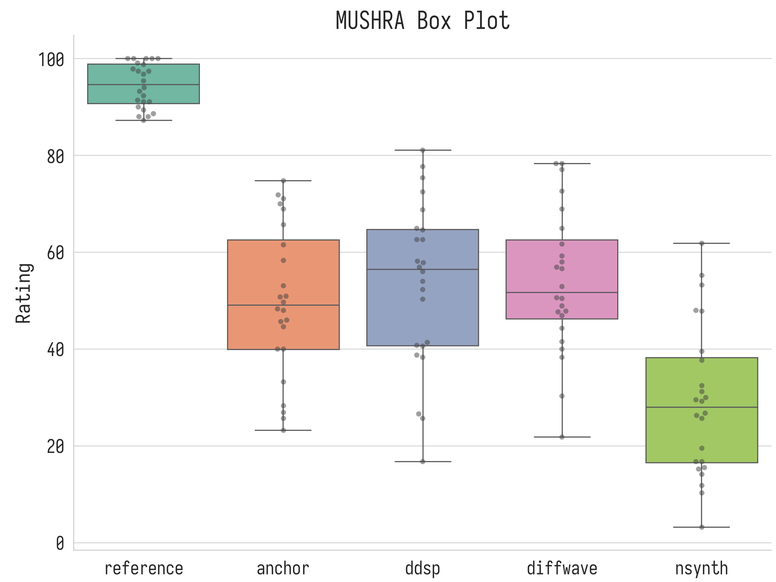


Fig 2. Objective metric results. Up arrow indicates that higher score is better and conversely down arrow indicates that a lower score is better

Discussion

Our results indicate that metrics that measure generator performance may not adequately capture generator quality. The objective metrics we evaluated may not give an accurate, meaningful estimate of the audio quality generated by these networks.



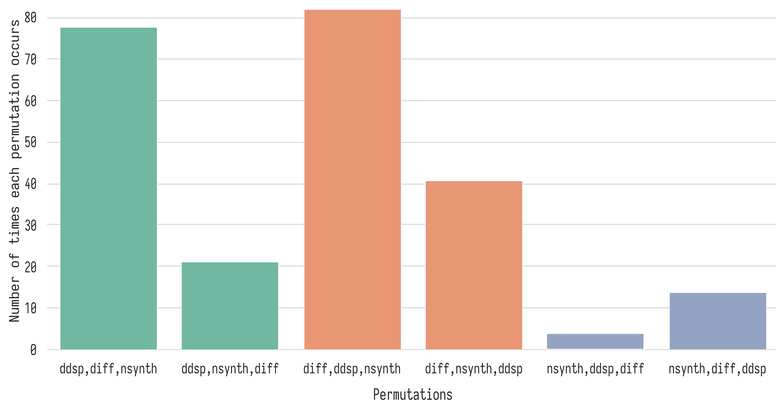
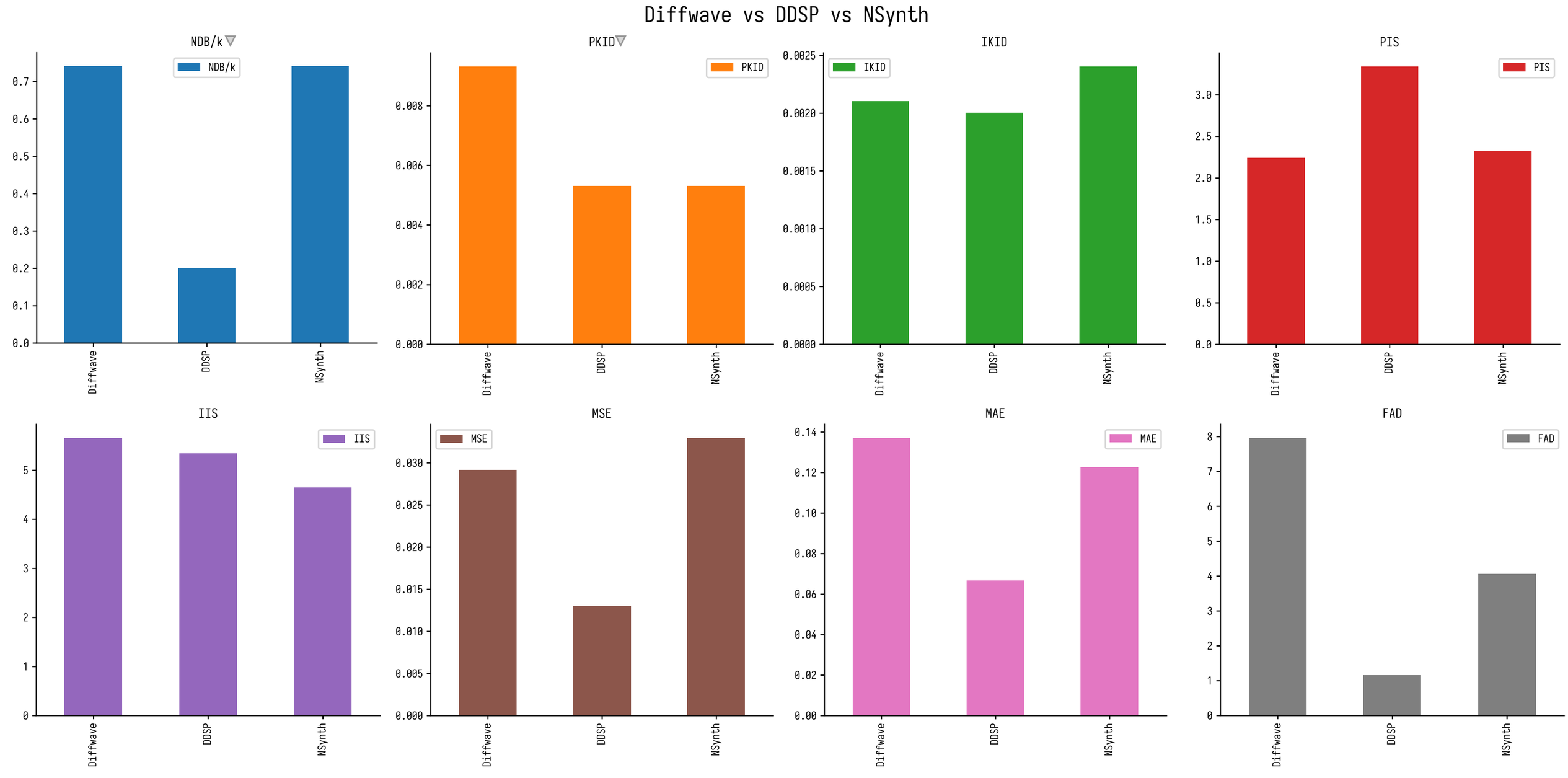


Fig 3. These are the results from our listening study. The plot above shows the results from our MUSHRA study and the plot below shows the various permutations of rankings that were selected by the participants.

References

[1] Engel, Jesse, et al. "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders." Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, pp. 1068–77, https://proceedings.mlr.press/v70/engel17a.html.
[2] Engel, Jesse H., et al. "DDSP: Differentiable Digital Signal Processing." 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020, https://openreview.net/forum?id=B1x1ma4tDr.
[3] Kilgour, Kevin, et al. "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms." Interspeech 2019, ISCA, 2019, pp. 2350–54, https://doi.org/10.21437/Interspeech.2019-2219.
[4] Kong, Zhifeng, et al. "DiffWave: A Versatile Diffusion Model for Audio Synthesis." 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021, https://openreview.net/forum?id=a-xFK8Ymz5J.
[5] Nistal, Javier, et al. "Comparing Representations for Audio Synthesis Using Generative Adversarial Networks." 2020 28th European Signal Processing Conference (EUSIPCO), 2021, pp. 161–65, https://doi.org/10.23919/Eusipco47968.2020.9287799.
[6] Richardson, Eitan, and Yair Weiss. "On GANs and GMMs." Advances in Neural Information Processing Systems, edited by S. Bengio et al., vol. 31, Curran Associates, Inc., 2018, https://proceedings.neurips.cc/paper/2018/file/0172d289da48c48de8c5ebf3de9f7ee1-Paper.pdf.
[7] Salimans, Tim, et al. "Improved techniques for training gans." Advances in neural information processing systems 29 (2016).
[8] Binkowski, Mikolaj, et al. "Demystifying MMD GANs." 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018, https://openreview.net/forum?id=r1lUOzWCW.


Diffwave vs DDSP vs NSynth
NDB/k
NDB/k
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Diffwave
DDSP
NSynth
PKID
PKID
0.000
0.002
0.004
0.006
0.008
Diffwave
DDSP
NSynth
IKID
IKID
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
Diffwave
DDSP
NSynth
PIS
PIS
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Diffwave
DDSP
NSynth
IIS
IIS
0
1
2
3
4
5
Diffwave
DDSP
NSynth
MSE
MSE
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Diffwave
DDSP
NSynth
MAE
MAE
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Diffwave
DDSP
NSynth
FAD
FAD
0
1
2
3
4
5
6
7
8
Diffwave
DDSP
NSynth