# Siamese SIREN: Audio Compression with Implicit Neural Representations

**Luca A. Lanzendörfer** [1]   **Roger Wattenhofer** [1]

## Abstract

Implicit Neural Representations (INRs) have emerged as a promising method for representing diverse data modalities, including 3D shapes, images, and audio. While recent research has demonstrated successful applications of INRs in image and 3D shape compression, their potential for audio compression remains largely unexplored. Motivated by this, we present a preliminary investigation into the use of INRs for audio compression. Our study introduces Siamese SIREN, a novel approach based on the popular SIREN architecture. Our experimental results indicate that Siamese SIREN achieves superior audio reconstruction fidelity while utilizing fewer network parameters compared to previous INR architectures.

## 1. Introduction

INRs have become known as an alternative representation for 3D shapes (Park et al., 2019; Mescheder et al., 2019), and have since been successfully applied to other data modalities such as radiance fields, images, and audio (Sitzmann et al., 2020; Yu et al., 2020; Chen et al., 2021; Mildenhall et al., 2021; Zuiderveld et al., 2021; Szatkowski et al., 2022).

In this paper, we apply INRs to audio compression, that is we approximate the audio signal function $f : \mathcal{T} \to \mathbb{R}$, where $\mathcal{T}$ is the time input domain and $\mathbb{R}$ is the amplitude output domain, with a small neural network. We take inspiration from the recent work on compression with INRs (Dupont et al., 2021; 2022; Strümpler et al., 2022) and build on previous work in audio INR (Sitzmann et al., 2020; Zuiderveld et al., 2021; Szatkowski et al., 2022).

Even though INRs cannot yet compete with other data compression approaches in the visual and audio domain (Dupont et al., 2022; Strümpler et al., 2022), we believe it still warrants further research. INRs have some interesting prop-
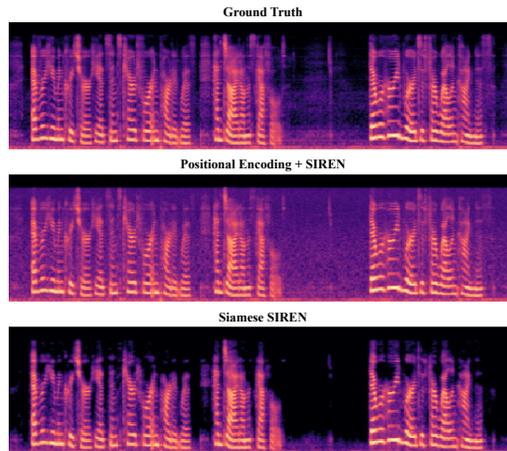


*Figure 1.* Log-mel spectrogram of a random 10-second LibriSpeech sample. We observe background noise produced from SIREN with positional encoding. Siamese SIREN is able to remove the noise by computing the noise estimate.

erties, such as being resolution-invariant to the input data, meaning the storage size does not scale with the input size, as well as having the ability to reconstruct data using any arbitrary resolution during inference.

However, a significant challenge arises when reconstructing audio with INRs. In images, noise may be present, but is often less noticeable. In audio data, however, even relatively small reconstruction errors become clearly perceivable in the form of stationary background noise due to the logarithmic nature of human hearing (Weber–Fechner law, cf. Appendix A). This noise thus becomes more pronounced the further we reduce model size and quantize model weights.

The above trade-off can be phrased in the general case as the following Pareto optimization problem: Let $\boldsymbol{p}$ be the parameters of a candidate INR, let $\mathcal{D}_f = \{(t, f(t)) : t \in \mathcal{T}\}$ be the data of the audio sample $f$, and let $q$ be a quality measure (cf. Section 3). Denote the memory footprint by $|\cdot|$. Solve

$$\max \left( q(\boldsymbol{p}, \mathcal{D}_f), \frac{|\mathcal{D}_f|}{|\boldsymbol{p}|} \right) \tag{1}$$
$$\text{subject to } \mathrm{MSE}\,(\boldsymbol{p}, \mathcal{D}_f) \to 0,$$

where the fraction to be optimized represents the compres-

[1]ETH Zurich, Switzerland. Correspondence to: Luca A. Lanzendörfer <lanzendoerfer@ethz.ch>, Roger Wattenhofer <wattenhofer@ethz.ch>.

sion ratio achieved by the INR, and where constraint convergence is to come from gradient-descent training of $\boldsymbol{p}$ on $\mathcal{D}_f$ until complete convergence (training for overfitting).

To address this problem, we propose Siamese SIREN, an INR model built on top of the general-purpose SIREN architecture (Sitzmann et al., 2020). The basic idea of our approach is to add two twin extensions to the standard SIREN model, both extensions trying to approximate the original audio signal $f$. Since both extensions will contain noise, but different noise, their difference can be leveraged to remove the noise from the reconstruction $\hat{f}$.

We demonstrate the viability of our approach via a set of audio metrics, log-mel spectrograms, and audio samples. The code and examples are available at https://github.com/lucala/siamese-siren.

## 2. Background

*INRs* are a class of functions, where one set of function parameters $\boldsymbol{p}$ describes one data sample $\mathcal{D}$. In particular, an INR is a neural network trained on $\mathcal{D}$ that approximates $f$.

*Sinusoidal Representation Networks*, referred to as SIREN, are a particular class of INR models (Sitzmann et al., 2020) that use the multi-layer perceptron (MLP) architecture with sine functions as their activation functions:

$$\phi_i : x_i \rightarrow sin\big(\omega_i \cdot (W_i x_i + b_i)\big), \qquad (2)$$

where $\phi_i$ is the $i^{th}$ layer of the network, $W_i$ and $b_i$ are the weight matrix and bias vector of the $i^{th}$ layer, respectively. The authors found the frequency scaling hyperparameter $\omega_i$ helps SIREN converge faster. They set $\omega_i = 30$, with $\omega_0 = 3000$ in the case of audio.

SIREN INRs have been shown to outperform standard ReLU-activated INRs on images, audio, and 3D geometry (Sitzmann et al., 2020).

*Positional Encoding* (PE) has been shown to help INRs learn high-frequency representations (Tancik et al., 2020; Mildenhall et al., 2021; Benbarka et al., 2021; Strümpler et al., 2022). We observe the same effect and also utilize PE, transforming the input into a high-dimensional embedding:

$$\gamma(t) = \big(t, sin(\sigma^0 \pi t), cos(\sigma^0 \pi t), \ldots,$$
$$sin(\sigma^L \pi t), cos(\sigma^L \pi t)\big), \qquad (3)$$

where $t$ is a normalized point in time, $\sigma$ is a frequency scaling term, and $L$ is the number of frequencies.

We propose a novel extension of SIREN, which we call *Siamese SIREN*. This architecture is motivated by our finding that small reconstruction errors of the waveform produced by SIREN contain audible background noise, even
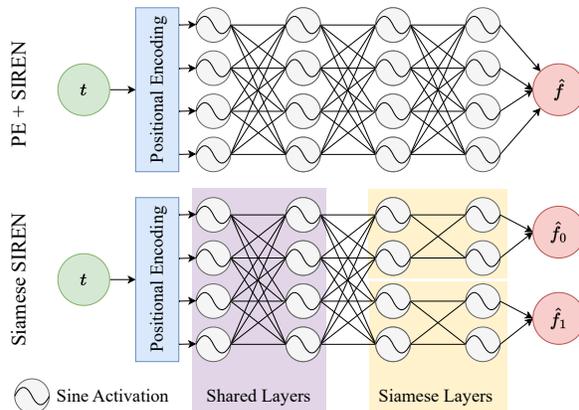


*Figure 2.* Overview of our proposed Siamese SIREN architecture, compared to SIREN with positional encoding (PE+SIREN). The above diagram illustrates a Siamese SIREN with two shared layers and two siamese layers, where each shared layer contains four units and each siamese layer contains two units. Our experiments are centered around two shared layers with 256 units each, and one siamese layer where each siamese head contains 128 units.

when training for tens of thousand of iterations (cf. Appendix C). This error is often magnified after quantizing the network weights. To remove the background noise of the reconstructed signal, we use *Noise Reduce* (Sainburg, 2019), an algorithm which computes a spectrogram of the signal and noise estimate. The signal and noise estimate are used to compute a noise threshold for each frequency band. A noise mask is computed based on the threshold, which is in turn used to remove the noise.

To construct a noise estimate for *Noise Reduce*, assume that a noisy reconstructed signal $\hat{f}$ can be linearly decomposed into the true signal $f$ and the noise component $\varepsilon$. Since the distribution from which $\varepsilon$ was sampled is not known in general, it has to be estimated. Training two INRs with different random weight initializiations on the same signal $f$ we obtain two approximations $\hat{f}_0, \hat{f}_1$ of $f$. We use the following rule to arrive at an estimate noise signal $\varepsilon_\bullet$

$$\varepsilon_\bullet = \alpha \left( \hat{f}_\bullet - \frac{\hat{f}_0 + \hat{f}_1}{2} \right), \qquad (4)$$

where $\alpha$ is a hyperparameter controlling the amplitude of the noise estimate. We find that tuning $\alpha$ has an impact on results, and we settle at $\alpha = 2$ for all experiments. $\varepsilon_\bullet$ can either be $\varepsilon_0$ or $\varepsilon_1$, and having no preference, we then feed $\varepsilon_0$ as the noise estimate to *Noise Reduce*.

Instead of naively training two INRs to be able to estimate the noise of the signal and thus doubling our parameter count, our proposed Siamese SIREN network merges a subset of layers, reducing the number of required parameters
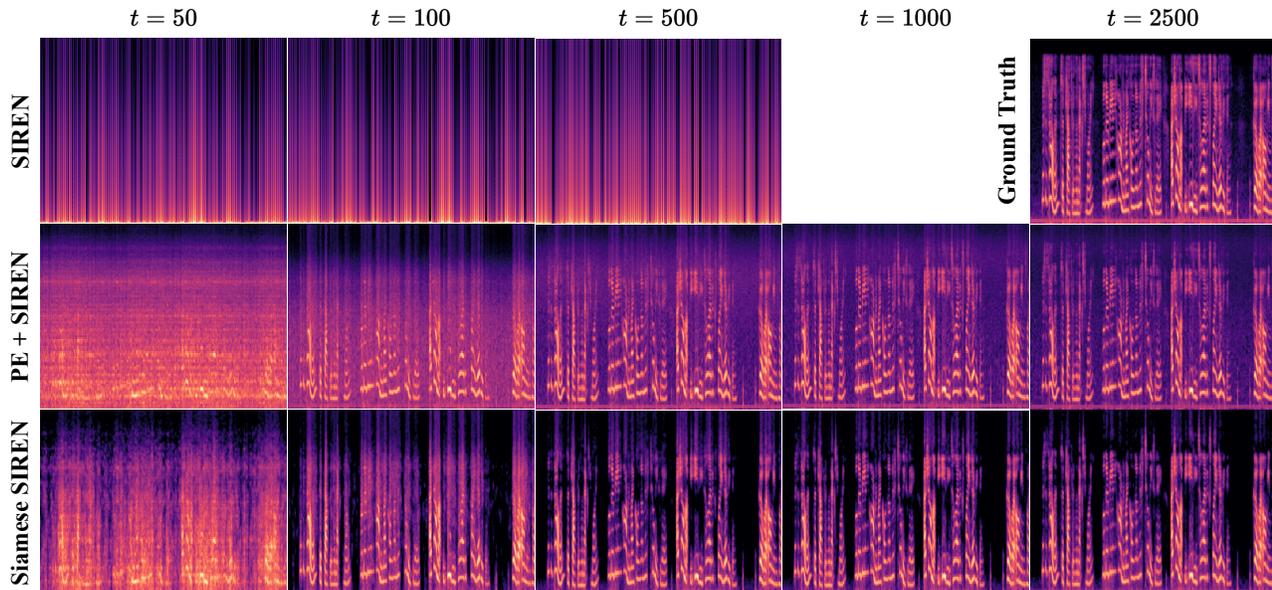
*Figure 3.* Comparison between different quantized models at training iteration $t$ on a 10-second LibriSpeech sample. SIREN without positional encoding (PE) cannot reproduce data after quantization. PE+SIREN is able to reproduce the signal with noise. Siamese SIREN can successfully estimate the background noise and remove it while using less parameters than PE+SIREN.

while still allowing for signals $\hat{f}_0$ and $\hat{f}_1$ to be learned. That way, each siamese twin possesses layers that are *shared* as well as layers that are specific only to it (*siamese layers*). In other words, the shared layers form a common backbone for the INR networks, whereas the siamese layers act as two separate heads, see Figure 2.

During training, both siamese heads learn to reconstruct the same signal $f$, but due to different random weight initialization of the heads, the reconstructed signals will vary slightly. We leverage this phenomenon to capture the noise estimate that is needed for noise removal, and find it to be effective.

Further on the front of parameter memory footprint reduction, *weight quantization* strategies reduce $|\boldsymbol{p}|$ and increases inference throughput by converting network weights. These are usually often stored with 32-bit floating point precision, but can be often quantized into smaller data types such as 8-bit integers. There exist various quantization schemes – two common approaches are: Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT). QAT tends to achieve lower reconstruction error after quantization. However, QAT needs to either be part of the network while training or fine-tuned after training the unquantized model. PTQ can be applied after training and does not require retraining the network, at the expense of slightly worse reconstruction quality. We use PTQ, as in our early experimentation we found PTQ errors did not significantly affect subjective signal quality.

## 3. Experiments

To evaluate the quality of our models, we use the GTZAN (Tzanetakis et al., 2001) and LibriSpeech datasets (Panayotov et al., 2015). GTZAN contains 1000 music snippets of ten different genres at 30 seconds each. For speech we use the `train.100` split of LibriSpeech which contains 14 second audio snippets of English speakers reading passages of text. We crop each audio snippet on the first 10 seconds at a sampling rate of 22050 Hz.

To evaluate the reconstruction quality we mainly rely on ViSQOL (Chinen et al., 2020), a metric to determine perceived audio quality. ViSQOL is designed to approximate a subjective listening test and produces Mean Opinion Scores between a reference and a test signal. We also employ CDPAM (Manocha et al., 2021), which approximates perceptual audio similarity between two signals. Additionally, we evaluate our models for LibriSpeech with PESQ (Rix et al., 2001) and STOI (Taal et al., 2011), which are designed to measure the perceived quality and intelligibility of speech in a signal. See Appendix B for more details.

## 4. Results

We are interested in the trade-off between compression speed, compression quality $q(\boldsymbol{p}, \mathcal{D}_f)$, and compression ratio $\frac{|\mathcal{D}_f|}{|\boldsymbol{p}|}$. We therefore evaluate SIREN and Siamese SIREN using small MLPs and over a small number of training itera-

*Table 1.* Comparison between different SIREN configurations after quantization, over random LibriSpeech samples. PE refers to Positional Encoding, $\omega_0$ refers to the scaling factor of the first SIREN layer, $\omega$ refers to all other SIREN layer scaling factors. *1x128* Siamese Layer refers to one layer with 128 units for each siamese head. The original SIREN performs worse since it is not able to reconstruct the signal after weight quantization.

| Model Name | Shared Layers | Siamese Layers | PE | $\omega_0$ | $\omega$ | #Params $(10^3)$ | Unquantized File Size (kB) | Quantized File Size (kB) | ViSQOL↑ | CDPAM↓ | PESQ↑ | STOI↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| original SIREN | 3x256 | 0 | 0 | 3000 | 30 | 794 | 1594.6 | 410.9 | 1.01 | 0.84 | 1.05 | -0.01 |
| PE + SIREN | 3x256 | 0 | 16 | 30 | 30 | 843 | 1692.9 | 435.5 | 1.63 | 0.2 | 1.91 | 0.93 |
| optimized SIREN | 3x256 | 0 | 16 | 100 | 100 | 843 | 1692.9 | 435.5 | 1.97 | 0.18 | 2.18 | **0.95** |
| Siamese SIREN | 2x256 | 1x128 | 16 | 100 | 100 | 513 | 1164.9 | 303.7 | **2.12** | **0.16** | **2.58** | 0.93 |

*Table 2.* Comparison between different layer configurations over random LibriSpeech samples.

| Shared | Siamese | #Params $(10^3)$ | ViSQOL↑ | CDPAM↓ | PESQ↑ | STOI↑ |
|---|---|---|---|---|---|---|
| 3x256 | 0 | 843 | 1.5 | 0.23 | 2.26 | **0.94** |
| 2x256 | 1x128 | 513 | **1.92** | **0.2** | **2.62** | 0.9 |
| 2x128 | 1x64 | 142 | 1.28 | 0.31 | 1.58 | 0.68 |
| 2x64 | 1x32 | 42 | 1.34 | 0.34 | 1.18 | 0.43 |

*Table 3.* Evaluating the effect of layer sharing over random LibriSpeech samples.

| Shared | Siamese | #Params $(10^3)$ | ViSQOL↑ | CDPAM↓ | PESQ↑ | STOI↑ |
|---|---|---|---|---|---|---|
| 3x256 | 0 | 843 | 2.06 | 0.34 | 2.21 | **0.95** |
| 2x256 | 1x128 | 513 | **2.22** | **0.14** | **2.73** | 0.93 |
| 1x256 | 2x128 | 151 | 1.94 | 0.24 | 2.00 | 0.87 |
| 0 | 3x128 | 75 | 1.85 | 0.31 | 1.77 | 0.81 |

tions. We train each model for 2500 iterations, which results in a compression time of around 25 seconds per sample on one Titan RTX. We find that the first 2500 iterations have the biggest impact on reconstruction quality. Preliminary experiments conducted by training to 10k iterations had led to slightly better results, but with a clear trend of diminishing returns. Even though the underlying signal can be distinctly heard after a few hundred steps, it is challenging to remove the remaining background noise. Longer training times reduce the presence of the noise, but it is left clearly audible. Our proposed approach solves this by estimating and removing the noise.

To compress the network, we quantize the network weights with PTQ, which reduces the storage size by 4x. We also analyze how the performance scores react to drastic reductions in network size. We find that the reconstruction degrades heavily when the network does not have sufficient parameters to learn to fit the signal, as can be seen in Table 2.

We notice a significant gap in metric performance compared to subjective listening evaluations. This is a well-known problem in audio model evaluation (Cartwright et al., 2016; Kilgour et al., 2019; Vinay & Lerch, 2022). We find that audio evaluation metrics tend to hold up better in speech signal analysis when compared to music signals.

To analyze the effect of estimating the noise distribution using our Siamese SIREN approach, we conduct an ablation study as shown in Figure 4. We observe a more pronounced cut-off when no noise estimate is provided, especially for music signals. We also noticed that the largest discrepancy in signal fidelity comes from reconstruction – the signal is subjectively only slightly more degraded after quantization.

Table 1 shows the results of our ablation study between the original SIREN and our Siamese SIREN. We observe the lowest score for the original SIREN, as the model cannot reconstruct the output after weight quantization. This can be seen in Figure 3.
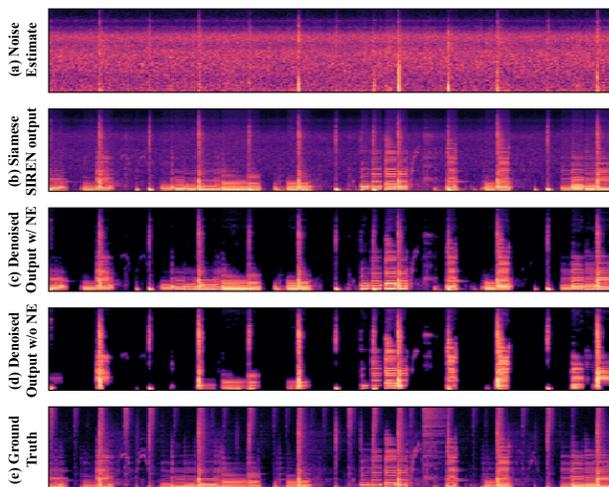


*Figure 4.* Comparison of noise removal. We visualize noise estimate $\varepsilon_0$ (a) and signal $\hat{f}_0$ (b). We demonstrate denoising results with noise estimate (c) and without noise estimate (d), we observe better results when using a noise estimate.

We also measure the impact of increasing the proportion of shared layers (cf. Table 3). Unsurprisingly, we find that the parameter count has a large influence on reconstruction quality of the signal, indicating that there is a trade-off between reducing network size and maintaining reconstruction quality. Our experiments further show that keeping large parts of the network shared and only splitting the last layer into siamese heads achieves the best quality-size trade-off.

In summary, we present a first approach to audio compression using INRs. We introduce Siamese SIREN – an extension to SIREN designed for audio compression and denoising tasks – and find it to be a viable candidate for INR-driven compression of audio. We hope our work will help to facilitate future research on INRs for sound and speech.

# References

Benbarka, N., Höfer, T., ul Moqeet Riaz, H., and Zell, A. Seeing implicit neural representations as fourier series. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2283–2292, 2021.

Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M. Fast and easy crowdsourced perceptual audio evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 619–623, 2016. doi: 10.1109/ICASSP.2016.7471749.

Chen, Y., Liu, S., and Wang, X. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8628–8638, 2021.

Chinen, M., Lim, F. S. C., Skoglund, J., Gureev, N., O'Gorman, F., and Hines, A. Visqol v3: An open source production ready objective speech and audio metric, 2020.

Dupont, E., Goliński, A., Alizadeh, M., Teh, Y. W., and Doucet, A. Coin: Compression with implicit neural representations, 2021.

Dupont, E., Loya, H., Alizadeh, M., Goliński, A., Teh, Y. W., and Doucet, A. Coin++: Data agnostic neural compression. *arXiv preprint arXiv:2201.12904*, 2022.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Manocha, P., Jin, Z., Zhang, R., and Finkelstein, A. Cdpam: Contrastive learning for perceptual audio similarity, 2021.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5206–5210. IEEE, 2015.

Park, J. J., Florence, P. R., Straub, J., Newcombe, R. A., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 165–174, 2019.

Rix, A. W., Beerends, J. G., Hollier, M., and Hekstra, A. P. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, 2:749–752 vol.2, 2001.

Sainburg, T. timsainb/noisereduce: v1.0, June 2019. URL https://doi.org/10.5281/zenodo.3243139.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

Strümpler, Y., Postels, J., Yang, R., van Gool, L., and Tombari, F. Implicit neural representations for image compression, 2022.

Szatkowski, F., Piczak, K. J., Spurek, P., Tabor, J., and Trzciński, T. Hypersound: Generating implicit neural representations of audio signals with hypernetworks, 2022.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. R. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19: 2125–2136, 2011.

Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains, 2020.

Tzanetakis, G., Essl, G., and Cook, P. Automatic musical genre classification of audio signals, 2001. URL http://ismir2001.ismir.net/pdf/tzanetakis.pdf.

Vinay, A. and Lerch, A. Evaluating generative audio systems and their metrics. In *International Society for Music Information Retrieval Conference*, 2022.

Yu, A., Ye, V., Tancik, M., and Kanazawa, A. pixelnerf: Neural radiance fields from one or few images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4576–4585, 2020.

Zuiderveld, J., Federici, M., and Bekkers, E. J. Towards lightweight controllable audio synthesis with conditional implicit neural representations, 2021.

## A. Noise Perception

We demonstrate the effect of logarithmic hearing perception by adding noise $\varepsilon \sim \mathcal{N}(0, 10^{-3}\mathbf{I})$ to a LibriSpeech sample (cr. Figure 5). We visualize this effect using log-mel spectrograms. The log-mel spectrogram is a perceptually-relevant amplitude and frequency representation of an audio sample. We observe a strong distinction between the ground truth log-mel spectrogram and the log-mel spectrogram of the noisy sample. However, this difference is not clearly noticeable when examining the amplitude waveform. Since we train SIREN and its extensions to learn on the amplitude waveform, it is unsurprising that there exists an inherent difficulty in removing noise.

## B. Model Training Parameters

We train all models for 2500 iterations using Adam optimizer (Kingma & Ba, 2017) with the $\beta_1$ and $\beta_2$ parameters as proposed in the paper, learning rate of $1e^{-4}$, a weight decay of $1e^{-5}$, and mean squared error as the loss function. We use frequency scaling with $\omega = 100$, which we found to give better results compared to the widely used $\omega = 30$. We use $L = 16$ positional frequency encoding with scaling $\sigma = 2$, resulting in a 33-dimensional input embedding which is passed into the MLP. Furthermore, we normalize time inputs into the range of $\mathcal{T} = [-1, 1]$. In our early experiments we tested other loss functions, scaling and positional frequencies, optimizers, learning rate and weight decay values, and learning rate schedules, but found them to have little effect on reconstruction quality. Furthermore, the *noise reduce* algorithm we used allows to only remove a percentage of detected noise. We did not use this feature, as we found keeping partial noise did not increase reconstruction quality.
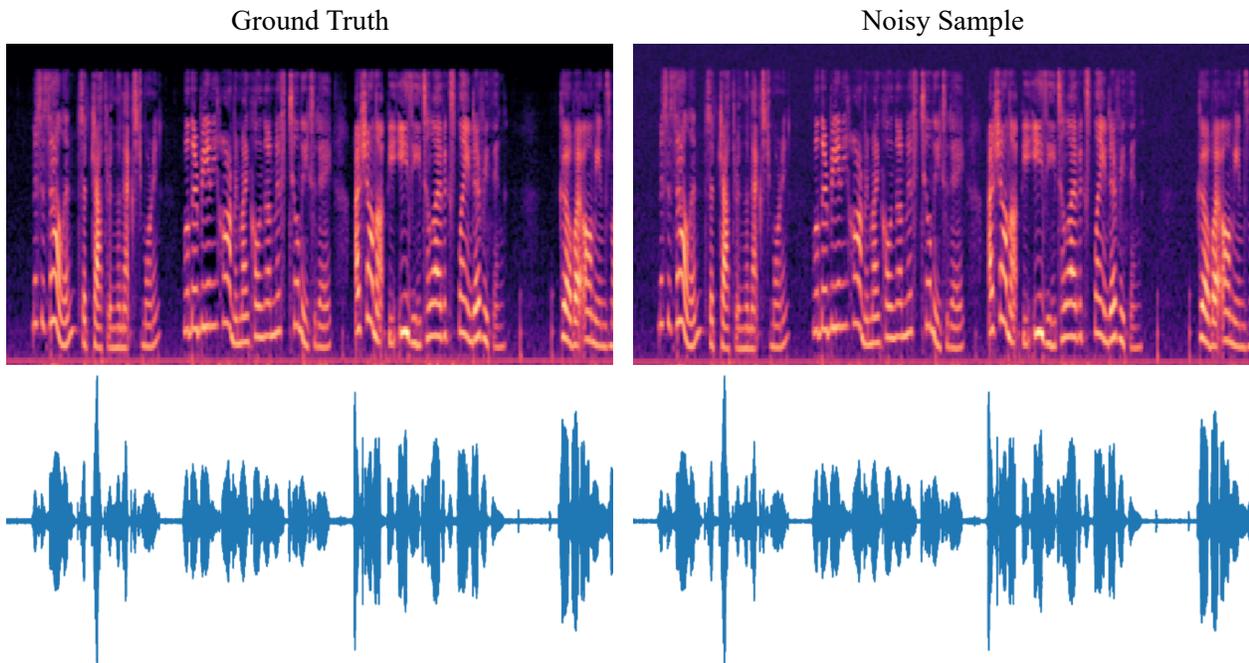


| Ground Truth | Noisy Sample |

*Figure 5.* Log-mel spectrogram of a LibriSpeech sample and the same sample with added $\varepsilon \sim \mathcal{N}(0, 10^{-3}\mathbf{I})$ noise. We can clearly see the difference in the spectrogram, but not in the waveform. This discrepancy is at the root of the challenge to remove noise. Even if the waveform is closely approximated, small errors are magnified and lead to distinctly audible noise.

## C. SIREN and Stationary Background Noise

We demonstrate the challenge of removing noise from an audio reconstruction. We train the original SIREN setup for audio, as described by (Sitzmann et al., 2020), and SIREN with positional encoding, over 100k iterations on a LibriSpeech sample (cf. Figure 6). For this experiment we do not quantize model weights. Comparing the spectrograms, we notice that without positional encoding the original SIREN struggles to reproduce high-frequency bands while simultaneously containing substantial stationary background noise in the mid-frequency and low-frequency bands. SIREN with positional encoding is capable of learning high-frequency content, however, it also produces substantial noise in these high-frequency bands. Furthermore, we observe only minimal improvement when training for more than 5000 iterations.
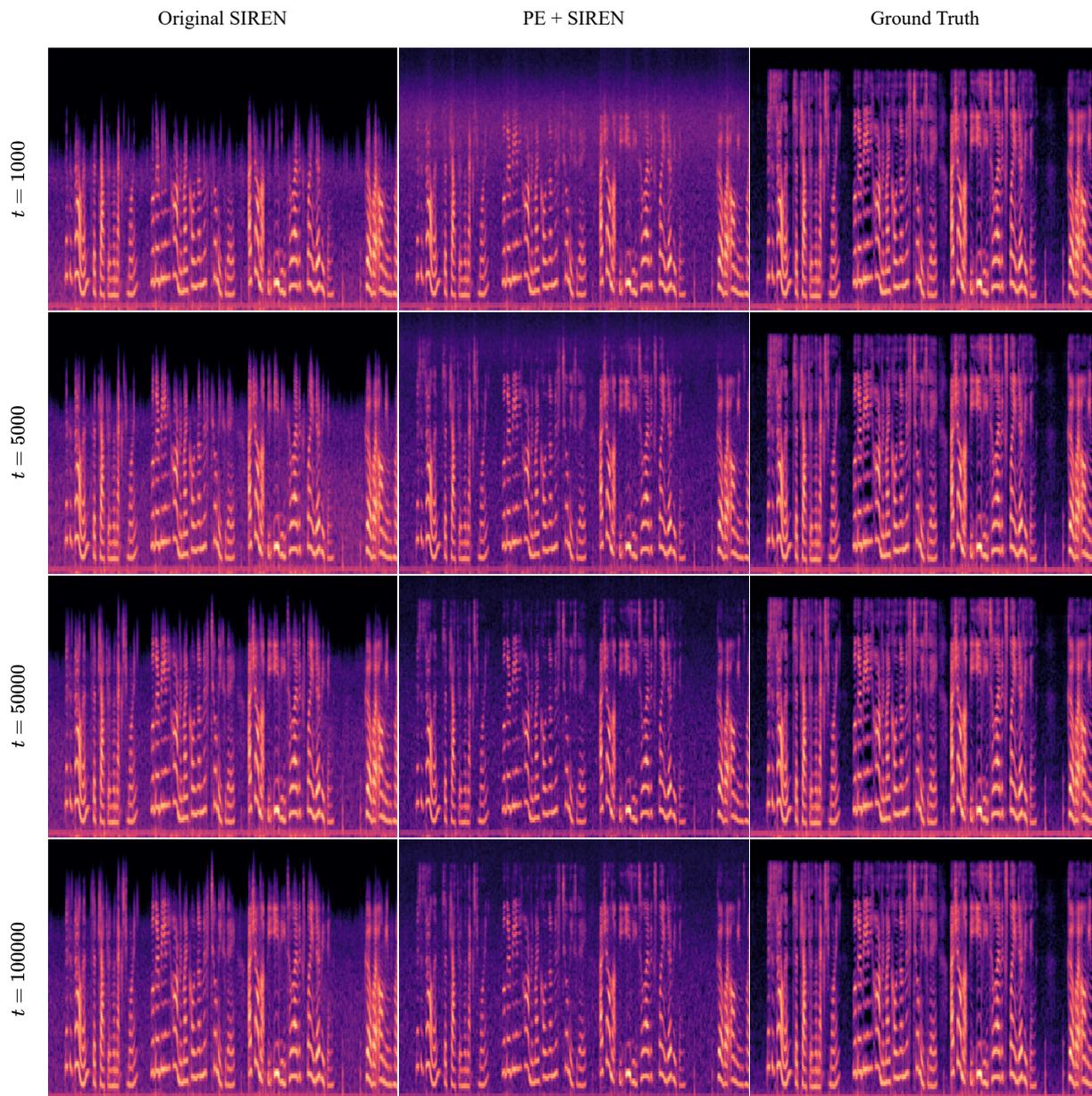


*Figure 6.* Log-mel spectrogram of unquantized original SIREN and unquantized SIREN with positional encoding over 100k iterations. We observe significant background noise in both reconstructions.