Team NoiCE: Final Report

T. Stadtmann, J. Loh, J. Paessens, M. Wabnitz, T. Gemmeke

Chair of Integrated Digital Systems and Circuit Design,
RWTH Aachen University, Germany

In this report, we - team 'NoiCE' (i.e. **Noi**se **C**ancelling **E**xperts) - highlight how we approached the task of neuromorphic audio denoising and our results in the Intel N-DNS Challenge. The overarching paradigm of our work is to incorporate insights from both biology and modern ANN research. To this end, we follow a bottom-up approach - starting from biological insights and tiny networks - and a top-down approach - starting from best-performing ANN methods and embracing neuromorphic principles like sparsity and recurrence. In parallel, we improve the performance of the baseline network. Given the complexity of our planned approaches, straightforward improvements to the baseline must not already outperform them. For instance, if strongly quantizing the baseline provides a well-performing, efficient solution, then more challenging, bio-inspired approaches should not only perform on par but strongly outperform it on one or more metrics.

As the approaches are highly involved, we are still actively working on their design and most optimal implementation - they will be outlined at the end of this report. Within the scope of the Intel N-DNS challenge, we focus on baseline enhancements and results of the extensive ablation study we performed to better understand which hyperparameters, input data, and feature extraction most heavily impact performance. These insights can then steer the optimization of our two upcoming approaches.

Our implementation can be trained and tested using the original interface. Relevant scripts reside in `preprocessing_loss/sdnn_delays`. The training script, `train_sdnn.py`, has been augmented with all parameters of our optimization study and is ready to be used for future exploration.

We performed the optimization in two stages. First, we fixed the network architecture of the baseline and iterated over various parameters: (1) input features (STFT with Hamming, Hann, Bartlett windows; MFCC; MDCT), (2) loss functions (MSE of MFCC coefficients; weighting of MSE to SI-SNR), (3) data augmentation (volume scaling from -20dB to 0dB; amplitude clipping to 80%; lowpass filtering). Our major takeaways are:

1. STFT coefficients result in the best SI-SNR across the board. The window length can be decreased from 512 to around 400 without performance loss.

2. The loss function can use MFCC coefficients instead of STFT to compute MSE to ground truth. This is less computationally heavy and without performance loss.

3. Data augmentation does not increase performance. The network has potentially no more information capacity.

4. Strongly quantized versions of these networks down to binary or ternary retain up to 9dB SI-SNR. While not performant enough, it remains a promising direction for further improvements.

In the second stage, we performed an extensive sweep over network architectures and optimizers on non-spiking networks to speed up the exploration. We opted for (temporal) convolutional layers, assuming that the dense baseline is already optimal. The results of our sweeps are shown in two heatmaps in our repository, `sweep_*.svg`. Network depth only improves performance until around 5 layers. Afterwards, it seems that information propagation is limited and deeper hierarchies do not contribute to better performance. Far more significant is channel width, meaning that more varied feature extraction is crucial to performance. As for optimizers, Adam and AdamW outperform alternatives at low learning rates. In the end, we picked the best performing models, recreated them with spiking neurons in Lava, and performed final adjustments to maximize SI-SNR. Here, interestingly, the networks learned to just recreate the inputs, not denoise them. To solve this, we added a penalty on the similarity between inputs and outputs to the loss function ($|0.5 - MSE(noisy, denoised)|$).

The results of our final model, consisting of 2 convolutional layers with a channel width of 512, are reported in Tab. 1. While its efficiency is impacted by the dense output layer and existing neuromorphic principles have not been augmented, our learnings are extremely valuable for our next steps and future neuromorphic algorithm design in general.

SNR	SNRi	$\text{MOS}_{\text{bak/ovrl/sig}}$	$\text{lat}_{\text{e+d}}$	lat_{all}	power	PDP	params	size
12.96dB	5.60dB	3.64/2.79/3.22	0.05ms	32.05ms	6111	195.87	2100k	8.2kB

Table 1: PDP = PDP-Proxy in M-Ops. Power = Power-Proxy in M-Ops/s. SNRi is the same value for data and encoder+decoder.

With this SI-SNR optimized baseline, and together with insights from other teams in the Intel N-DNS challenge, we will gain a holistic view on neuromorphic audio denoising. This will assist in a faster convergence for our two future approaches:

1. **Efficient Balanced Networks (EBNs)** are SNNs that are based on analytical and experimental evidence for a tight balance between cortical inhibition and excitation. They can learn to mimic complex dynamics with extremely small networks and realistic, sparse spiking activity using local learning rules akin to feedback alignment. For our exploration, we use knowledge distillation - we first train a non-spiking reservoir to solve the denoising task, and then design an EBN to mimic its behavior.

2. **Transformers** have shown impressive feats in the area of Natural Language Processing, amongst others. In our setup, we target a speech resynthesis approach, where noisy input is first transformed into text, from which denoised speech is synthesized.

In the end, leveraging insights from both neuroscience and Machine Learning research will guide future neural networks to ever more efficient and high-performing results.