Speech Enhancement with Liquid Neural Networks

Graduate

Yanik Kuster

Introduction: Liquid Time-Constant Neural Networks (LTCs) were implemented on an FPGA-based hardware accelerator in a previous semester project. Currently, due to LTCs' novelty, they lack real-world applications in which they have been tested, despite their claimed superior expressivity. Since LTCs are based on solving systems of ODEs at each iteration, they are well suited for dynamical problems. One such problem is speech enhancement (SE) that tackles rapidly changing noise conditions. Deep neural networks (DNNs) were shown to perform well on SE tasks due to their ability to learn speech patterns.

Approach: This thesis aims to analyze the feasibility of LTCs applied to monaural speech denoising on a wearable device such as a hearing aid. The theoretical SE performance is evaluated by training the CDNN architecture (Fig. 3) using either LSTMs or LTCs. The CDNN-LSTM serves as ground truth to which the CDNN-LTC is compared using PESQ and STOI scores. Both models were trained and tested on the VoiceBank+Demand noisy speech dataset using a training validation split of 80-20% for hyperparameter tuning. The best model is implemented on an embedded platform to show the ability of wearable devices to run SE models in real time. The platform uses ETH's VitalCore as a base, since it implements battery management and BLE. The neural processing unit (NPU) and audio codec are added using the custom-built Neural Extension (Fig. 1). The Keras model was quantized by LiteRT, optimized for the NPU with the Vela compiler and implemented using the TFLite Micro kernel.

Conclusion: The current embedded implementation has an NPU execution time of 18.6 ms, which sums up to 38.6 ms total system latency. The system is considered real-time capable if the latency is lower than 20 ms. Nonetheless, the implementation has still a lot of room for optimization, e.g., by pruning, decreasing the STFT window size and using the Neural Extension's faster NPU. The test results showed that both models were able to increase PESQ and STOI scores of noisy speech (Fig. 2). However, the CDNN-LTC consistently performed worse and was harder to train than the CDNN-LSTM.

Fig. 1: The Neural Extension adds two Ethos-U55 NPUs and an audio codec to ETH's VitalCore base platform.
Own presentment

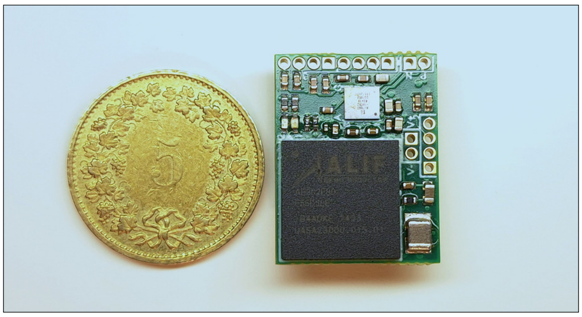
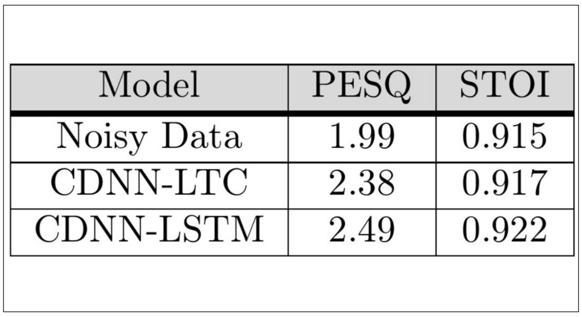
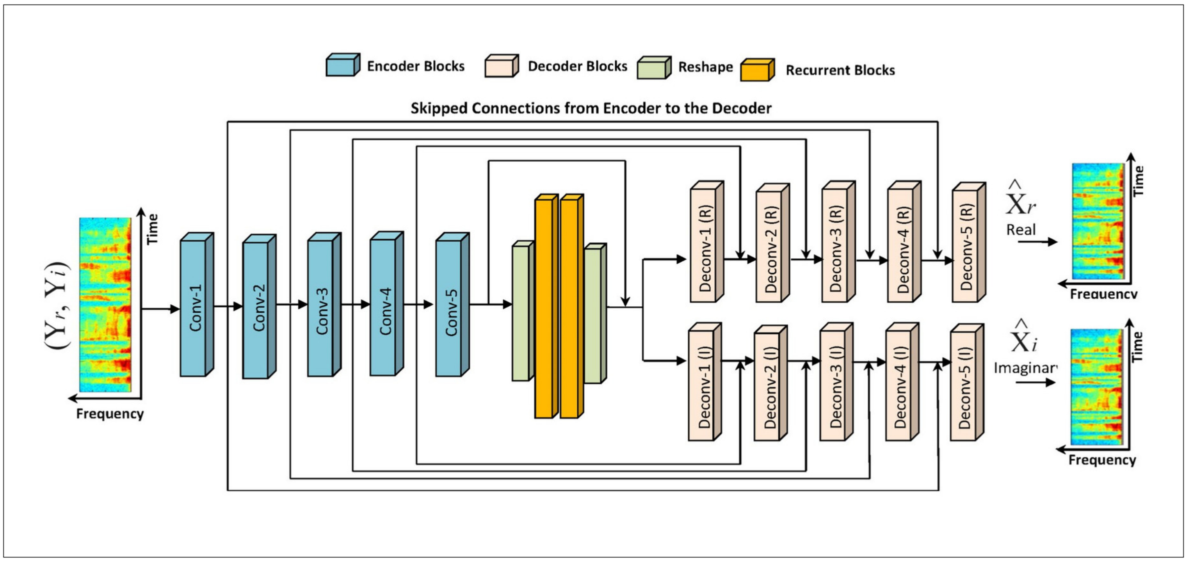


Fig. 2: Measured performance on the test set. The row "Noisy Data" contains the computed metrics if no SE is applied.
Own presentment



Model	PESQ	STOI
Noisy Data	1.99	0.915
CDNN-LTC	2.38	0.917
CDNN-LSTM	2.49	0.922

Fig. 3: The base CDNN architecture was implemented once with LSTMs and once with LTCs as the inner "Recurrent Blocks".
sciencedirect.com/science/article/pii/S0167639323001425



Advisor
Prof. Dr. Andreas Breitenmoser

Co-Examiner
Dr. Sebastian Stenzel, Sonova, Stäfa, Zürich

Subject Area
Electrical Engineering, Data Science