[Public]

DTLN: A Dual-Signal LSTM Transform Speech Enhancement System Based on Zynq

Xuebin Ruan, Jiahao Zhang, Jinrui Wang
Hohai University, Jiangsu Province

OpenHW2023

AMD

On board test by AMD ZYNQ XCZU3EG

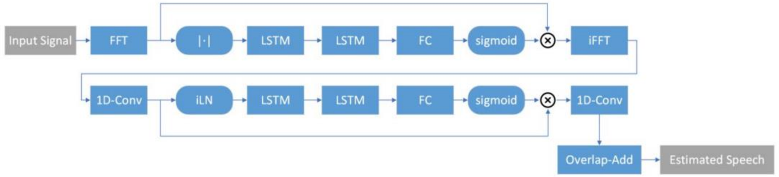
INTRODUCTION

Through the **deep neural network** denoising algorithm DTLN combined with FPGA hardware acceleration, our **ZYNQ** platform has shown good performance in completing **real-time denoising tasks in complex environments**, and can efficiently handle noise in multiple frequency bands. The flexibility of FPGA programming has given us great benefits, and at the same time, we have fully utilized the advantages of heterogeneous acceleration in optimizing the architecture of the ZYNQ platform, achieving high real-time performance.

Using the logarithmic power spectrum of **Fast Fourier Transform (FFT)**, predict the gain or mask applied to noisy FFT, and then reconstruct the speech signal by estimating the amplitude and phase of the noise. The use of DTLN neural network denoising model, with dual signal conversion LSTM as the core, can better capture temporal information in speech signals by learning long-term dependency relationships, thereby effectively reducing noise interference.

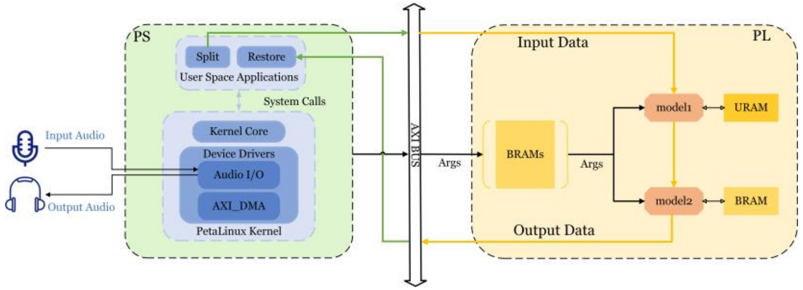
By using the **AXI-DMA IP core** provided by Xilinx, high bandwidth direct memory access is provided between the storage and target peripherals such as AXI4 Stream. Two types of BRAM were used to temporarily store model parameters and intermediate variables. The BRAM that stores parameters only requires one write operation, while the remaining time is read-only. The other type of BRAM requires frequent reading and writing of data.

CREATIVE DESING



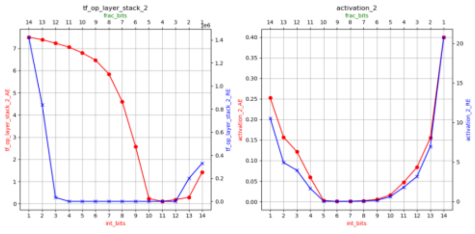
DTLN Network Flow Diagram

The **data flow process** is shown in the above figure. Firstly, FFT is used to separate the audio into amplitude and phase. Then, the mask is calculated through the first separation core, multiplied by the amplitude of the mixed audio, and combined with the input phase, it is converted back to the time domain. The temporal results generated by the first network are then processed through a 1D convolutional layer to form a feature representation, which is then processed through a normalization layer and passed to the second separation core. The final prediction mask of the second core is multiplied by the non normalized feature representation, and the result is converted back to the time domain through 1D convolutional layers.

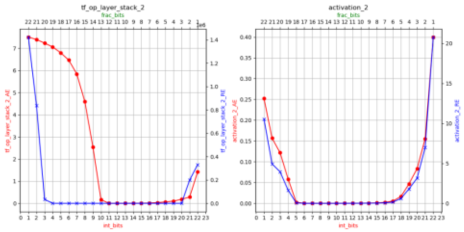


ZYNQ Structural Design Drawing

RESULT



16 bit local error



24 bit local error

From the results, the optimal bit structure in 16bit is 1-8-7, and the optimal bit structure in 24bit is 1-11-12. However, the error of the former is still relatively large, and the background noise of our measured 16bit audio results is also relatively large.Therefore, we chose the quantization structure of 1-11-12 in 24bit.

total_bits	int_bits	frac_bits	MAE	MRE	WEA
16	8	7	0.030429654	10592.23188	3177.690866
	9	6	0.057112397	21341.31752	6402.435235
	10	5	0.112998223	42857.19581	12857.23784
24	10	13	0.001149669	250.7426875	75.22361101
	11	12	0.000641362	124.8317268	37.44996699
	12	11	0.00218341	553.1552291	165.9480971