Voice Activity Detector

A Voice Activity Detector (VAD) is a system detecting the presence of voice over the input audio stream (Fig. 1).

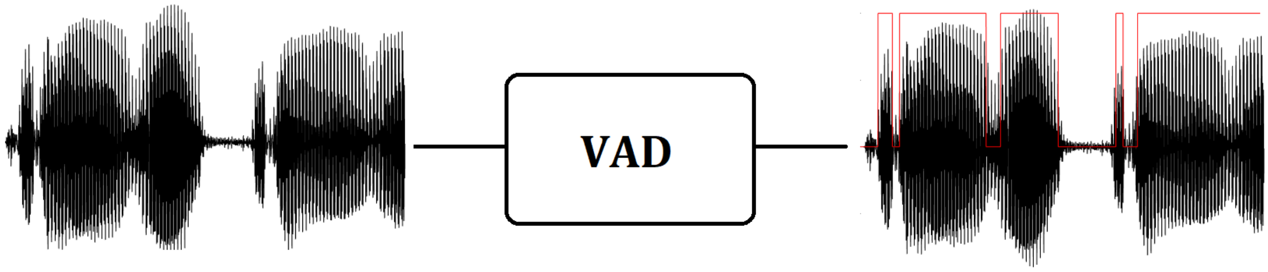

Fig. 1

A simple VAD can be realized by dividing the input stream in audio portions called *frames* of length T_{FRAME}= 16 ms, and comparing the *mean frame Energy* E_{FRAME} with a fixed threshold E_{TH}, where:

$$E_{FRAME} = \frac{1}{N}\sum_{n=1}^{N} x^2[n]$$

where N is the number of samples in a frame and $x[n]$ is the amplitude of the n^{th} frame sample.

Considering an audio sample rate $f_s = 16\ kHz$, realize a VAD that, taking as input a stream of input samples $x[n]$, provides a binary response VAD = 1 (voiced frame) when $E_{FRAME} \geq E_{TH}$ and VAD = 0 (silent frame) otherwise.

The frame start shall be signalized by a pulse of 1 clock cycle on the correspondent FRAME_START input. The VAD output shall be provided at the end of the correspondent frame.

$x[n]$ is included in the range [-1,1) and it is represented by using 16 bits according to the standard C2 representation. E_{TH} = 0.05 and is represented as an <u>unsigned</u> number.

The VAD interface is shown in Fig. 2. The signals clk and rst_n are respectively the clock and the system active low reset.

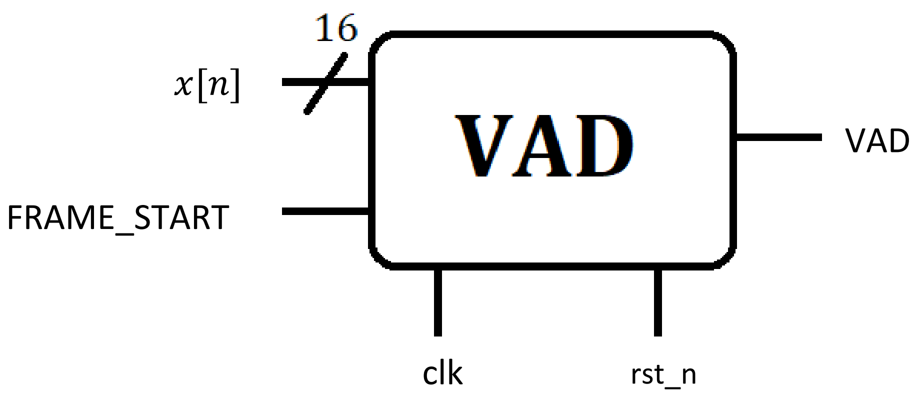

Fig. 2

Report the architecture used, the VDHL code of sources and testbench. Report the test plan.

The design shall be implemented on board the (xc7z010clg400-1) Zynq 7000 FPGA device, reporting the maximum clock frequency of design, the source occupation and the power consumption estimation.

Hint 1: One input $x[n]$ sample is provided to the system every 1 / f_s. Can you exploit this data seriality to calculate the E_{FRAME}?

Hint 2: To calculate $x^2[n]$, it could be convenient performing $|x[n]|$, switching from C2 to unsigned, and then squaring the $|x[n]|$

Hint 3: Do you really need to divide the term $\sum_{n=1}^{N} x^2[n]$ by N? Can you exploit the fact that both E_{TH} and N are constant?